

SUPPORTING MULTILINGUAL INTERNET SEARCHING AND
BROWSING

by

Yilu Zhou

Copyright © Yilu Zhou 2006

A Dissertation Submitted to the Faculty of the
COMMITTEE ON BUSINESS ADMINISTRATION

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Yilu Zhou entitled Supporting Multilingual Internet Searching and Browsing and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: July 24, 2006
Hsinchun Chen

_____ Date: July 24, 2006
Jay F. Nunamaker, Jr.

_____ Date: July 24, 2006
J. Leon Zhao

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: July 24, 2006
Dissertation Director: Hsinchun Chen

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Yilu Zhou

ACKNOWLEDGEMENTS

First of all, I would like to thank my dissertation advisor, Professor Hsinchun Chen, for his guidance and encouragement throughout my five years at the University of Arizona. It has been an invaluable opportunity for me to work in the Artificial Intelligence Lab under his direction. Many thanks go to my committee members, Dr. Jay F. Nunamaker, Jr. and Dr. J. Leon Zhao for their guidance and encouragement. I also thank all the faculty members in the MIS Department for their support.

My dissertation has been partly supported by grants from the National Science Foundation. Many thanks to my colleagues in the AI Lab and in the department for their tremendous help, advice and support through the past five years. Thanks also to Ms. Barbara Sears and Ms. Sarah Marshall for editing my papers.

I especially thank Jennifer Xu, Yiwen Zhang, Gang Wang, Ming Lin, Jason Li, Harry Wang and many others for their encouragement and emotional support during my stressful time. I am most grateful for Jialun Qin who has been with me and shared all my happiness and distress. Most of all I appreciate the constant support from my parents and my family.

DEDICATION

This dissertation is dedicated to my grandpa, Jianhua Zhou, who passed away February 9,
1990.

TABLE OF CONTENTS

LIST OF TABLES.....	9
LIST OF FIGURES.....	11
ABSTRACT.....	12
CHAPTER 1 INTRODUCTION.....	14
1.1 Background.....	14
1.2 Web Search in Non-English Languages.....	15
1.3 Multi-lingual Web Retrieval.....	17
1.4 Translation of Proper Names.....	19
1.5 Research Framework.....	21
CHAPTER 2 SUPPORTING MULTI-REGIONAL INFORMATION SEEKING: A STUDY IN THE CHINESE MEDICAL DOMAIN.....	24
2.1 Introduction.....	24
2.2 Related Work.....	26
2.2.1 Information Seeking Behaviors on the Internet: Searching and Browsing.....	26
2.2.2 Techniques Facilitating Information Seeking.....	28
2.2.3 Information Seeking in a Multilingual World: Research Gaps.....	33
2.3 Research Questions.....	38
2.4 A Research Testbed in the Chinese Medical Domain.....	39
2.5 Proposed Approach.....	41
2.5.1 An Integrated Knowledge Portal Approach.....	41
2.5.2 A Research Prototype in the Chinese Medical Domain: CMedPort.....	43
2.6 Evaluation Methodology.....	56
2.6.1 Search and Browse Tasks.....	56
2.6.2 Benchmarks.....	58
2.6.3 Hypotheses.....	59
2.6.4 Experimental Design.....	60
2.7 Experimental Results.....	62
2.7.1 Results from Search Tasks.....	63
2.7.2 Results from Browse Tasks.....	64
2.7.3 Results from Usability Questionnaire.....	67
2.7.4 Subjective Feedback.....	68
2.8 Conclusions and Future Directions.....	71
CHAPTER 3 FACILITATING CROSS-LINGUAL WEB RETRIEVAL: AN EXPERIMENT IN ENGLISH- CHINESE BUSINESS INTELLIGENCE.....	73
3.1 Introduction.....	73
3.2 Literature Review.....	75
3.2.1 Query Translation Approaches.....	76

TABLE OF CONTENTS - *CONTINUED*

3.2.2 Reducing Translation Ambiguities and Errors	79
3.2.3 CLIR for Web applications	82
3.2.4 Summary	83
3.3 Proposed Approach to Multilingual Web Retrieval.....	85
3.3.1 Web Spider and Indexer.....	86
3.3.2 Pre-translation Query Expansion	88
3.3.3 Query Translation	88
3.3.4 Post-translation Query Expansion.....	90
3.3.5 Document Retrieval	91
3.4 ECBizPort: An English-Chinese Web Portal for Business Intelligence.....	91
3.4.1 Domain Selection.....	93
3.4.2 Web Spider and Indexer.....	93
3.4.3 Query Translation	95
3.4.4 Document Retrieval	99
3.5 An Example of Query Translation and Expansion	100
3.6 System Evaluation	104
3.6.1 CLIR Evaluation Methodologies	104
3.6.2 Experiment Design and Measure	105
3.7 Experimental Results and Discussions	107
3.7.1 Precision.....	108
3.7.2 Efficiency.....	110
3.8 Conclusions and Future Directions.....	111
 CHAPTER 4 DEVELOPING A MULTILINGUAL WEB RETRIEVAL SYSTEM: EXPERIMENTS ACROSS WESTERN AND EASTERN LANGUAGES.....	
4.1 Introduction.....	113
4.2 Literature Review: MLIR on the Web.....	114
4.2.1 Resources to Support MLIR	115
4.2.2 MLIR for Web Applications.....	117
4.3 Research Questions.....	118
4.4 A Proposed Multilingual Web Retrieval (MWR) Approach	120
4.4.1 Multilingual Collection Building.....	122
4.4.2 Query Translation	125
4.4.3 Document Retrieval	129
4.5 A Multilingual Web Portal for Business Intelligence: An Experiment	130
4.5.1 Multilingual Collection Building.....	130
4.5.2 Query Translation	134
4.5.3 Document Retrieval	135
4.5.4 Sample User Sessions	135
4.6 System Evaluation	138

TABLE OF CONTENTS - *CONTINUED*

4.6.1 MLIR Evaluation Methodology.....	138
4.6.2 Experiment Design and Measures	139
4.6.3 Hypotheses.....	141
4.7 Experiment Results and Discussions	143
4.7.1 Overall Comparison Between Multilingual and Monolingual.....	145
4.7.2 Phrasal and Co-occurrence Translation Comparison.....	146
4.7.3 Efficiency.....	151
4.8 Conclusions and Future Directions.....	153
CHAPTER 5 NAME transliteration BY COMBINING THE	
PROBABILITY MODEL AND THE WEB MINING	
MODEL	
5.1 Introduction.....	156
5.2 Related Works.....	157
5.2.1 Transliteration Problem	157
5.2.2 Transliteration Models Overview	160
5.2.3 A Taxonomy of Transliteration Research.....	165
5.3. Research Questions.....	167
5.4 Proposed Framework: <i>Arizona NameTran</i>	167
5.4.1 Training Statistical Model.....	170
5.4.2 Transliteration Process.....	174
5.5 Experiment Design and Hypotheses	177
5.5.1 Hypotheses.....	178
5.5.2 Experiment Measure	179
5.5.3 Language Pairs and Dataset.....	181
5.5.4 Experiment Methodology	182
5.6 Experiment Results	183
5.6.1 English-Arabic back transliteration	183
5.6.2 English-Chinese forward transliteration	187
5.6. Conclusions and Discussions.....	193
CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS.....	
6.1 Conclusions.....	194
6.2 Future Directions	197
6.3 Relevance to Business and MIS Research.....	197
REFERENCES	199

LIST OF TABLES

Table 2.1: Major Chinese search engines in the three regions	49
Table 2.2: Computer system usability questionnaire (Lewis, 1995)	62
Table 2.3.1: Searching performance of CMedPort and benchmark systems by regions	63
Table 2.3.2: Hypotheses testing for search tasks by regions	64
Table 2.3.3: Searching performance of CMedPort and benchmark systems with combined regions	64
Table 2.4.1: Browsing performance of CMedPort and benchmark systems by regions	66
Table 2.4.2: Hypotheses testing for browse tasks of CMedPort and benchmark systems	66
Table 2.4.3: Hypotheses testing for browse tasks of CMedPort summarizer and categorizer	67
Table 2.5.1: User satisfaction rating of CMedPort and benchmark systems	68
Table 2.5.2: Hypotheses testing for user satisfaction	68
Table 3.1: Precision and time.....	108
Table 3.2: Paired <i>t</i> -test results.....	109
Table 4.1: Spider program settings and number of pages collected for each language.....	131
Table 4.2: Bilingual dictionaries used in query translation	134
Table 4.3: Summary of system effectiveness performance	144
Table 4.4: Summary of system effectiveness performance	145
Table 4.5: Comparison between monolingual and best multilingual performance	146
Table 4.6: Efficiency of Multilingual Business Intelligence Portal.....	152
Table 5.1: Transliteration problems studied in previous research	159
Table 5.2.1: Taxonomy of Transliteration Research	166
Table 5.2.2: Taxonomy of Transliteration Research using Statistical Approach	166
Table 5.3: Summary of system performance (accuracy) with different models and their improvement over a Simple Statistical model (English-Arabic).....	184
Table 5.4: Summary of average accuracy achieved and <i>t</i> -test results (English-Arabic).....	185
Table 5.5: Summary of average accuracy achieved and <i>t</i> -test results using combined Probability and Web mining model (English-Arabic).....	186
Table 5.6: Summary of system performance (accuracy) with different models and their improvement over a Simple Statistical model (English-Chinese).....	188
Table 5.7: Summary of average accuracy achieved and <i>t</i> -test results (English to Pinyin)	189
Table 5.8: Summary of average accuracy achieved and <i>t</i> -test results (Pinyin to Chinese).....	189

LIST OF TABLES - *CONTINUED*

Table 5.9: Summary of average accuracy achieved and <i>t</i> -test results (English to Chinese).....	190
Table 5.10: Summary of average accuracy achieved and <i>t</i> -test results using combined Probability and Web mining model (English-Chinese).....	192

LIST OF FIGURES

Figure 2.1: The CMedPort system architecture	44
Figure 2.2: The CMedPort user interface	55
Figure 3.1: CLIR system architecture.....	86
Figure 3.2: Sample screenshots of ECBizPort.....	92
Figure 4.1: Proposed architecture for a multilingual Web retrieval system (the MWR system).....	121
Figure 4.2 Query translation steps with an example in English-Chinese translation	126
Figure 4.3: User interface of Multilingual Business Intelligence Portal	137
Figure 5.1: Training Statistical Model.....	169
Figure 5.2: Transliteration Process	169
Figure 5.3: Pseudo-codes for word alignment process	172
Figure 5.3: Performance comparison of probability models (accuracy) (English- Arabic).....	185
Figure 5.4: Performance comparison of combined probability and Web mining models (accuracy) (English-Arabic)	187
Figure 5.5: Performance comparison of probability models (accuracy) (English- Pinyin)	189
Figure 5.6: Performance comparison of probability models (accuracy) (Pinyin- Chinese).....	189
Figure 5.7: Performance comparison of probability models (accuracy) (English- Chinese).....	191
Figure 5.8: Performance comparison of combined probability and Web mining models (accuracy) (English-Chinese)	192

ABSTRACT

The amount of non-English information has proliferated rapidly in recent years. The broad diversity of the multilingual content presents a substantial research challenge in the field of knowledge discovery and information retrieval. Therefore there is an increased interest in the development of multilingual systems to support information sharing across languages. The goal of this dissertation is to study how different techniques and algorithms could help in multilingual Internet searching and browsing through a series of case studies.

A system development research process was adopted as the methodology in this dissertation. In the first part of the dissertation, I discuss the development of CMedPort, a Chinese medical portal to serve the information seeking needs of Chinese users. A systematic evaluation has been conducted to study the effectiveness and efficiency of CMedPort in assisting human analysis. My experimental results show that CMedPort achieved significant improvement in searching and browsing performance compared to three benchmark regional search engines.

The second and third case studies aim to investigate effective and efficient techniques and algorithms that facilitate multilingual Web retrieval. An English-Chinese multilingual Web retrieval system in the business IT domain was developed and

evaluated. It was then extended into five languages: English, Chinese, Japanese, German and Spanish. A dictionary-based approach was adopted in query translation. Corpus-based co-occurrence analysis, relevance feedback, and phrasal translation algorithms were used for disambiguation purposes. Evaluation results showed that the system's phrasal translation and co-occurrence disambiguation led to great improvement in performance. The last part of this dissertation studies proper name translation problem. Proper names are often out-of-vocabulary terms and are critical to multilingual Web retrieval. This study proposes a combined Hidden Markov Model and Web mining model to automatically generate proper name translations. The approach was evaluated on two language pairs: English-Arabic and English Chinese. My results are encouraging and show promise for using transliteration techniques to improve multilingual Web retrieval.

This dissertation has two main contributions. Firstly, it demonstrated how information retrieval, Web mining and artificial intelligence techniques can be used in a multilingual Web-based context. Secondly, it provided a set of tools that can facilitate users in their multilingual Web searching and browsing activities.

CHAPTER 1

INTRODUCTION

1.1 Background

Rapid growth of the World Wide Web has produced a wealth of information in almost every major language, making the Internet a truly multilingual world. A report published by Internet World Stats (<http://www.internetworldstats.com/stats7.htm>) in March of 2006 showed that the majority of total Internet users are from non-English speaking areas (69.4%). Furthermore, the population of non-English-speaking Internet users is growing much faster than that of English-speaking users. Asia, Africa, the Middle East and Latin America are the areas with the fastest growing online population. As a result, study of Internet searching and browsing support has become an interesting and challenging research problem in a multilingual world.

Internet users often face an information overload problem (Blair and Maron, 1985). A search employing a general-purpose search engine such as Google can result in thousands of hits. This problem is particularly serious for non-English-speaking users because most Internet searching and browsing techniques are developed for speakers of English. The wealth of information available on the Internet has led to much research directed toward

developing techniques to support non-English information retrieval and knowledge discovery.

On the other hand, it is often difficult for an English speaker to access non-English content on the Web. In general, it is difficult for a user to retrieve documents written in a language that is not spoken by that user. As a result, retrieval from the Web of documents in different languages presents a very interesting and challenging research problem.

1.2 Web Search in Non-English Languages

Because the number of non-English resources available on the Web is increasing rapidly, developing information retrieval techniques for non-English languages is becoming an urgent and challenging issue.

Major search engines have provided search support in both English and non-English languages. Several Web search engines such as Google, Yahoo!, and AltaVista can handle multiple languages in addition to English and can specify the target language of the documents to be retrieved. Google currently supports searching in more than 100 languages. Yahoo has local sites in 25 different languages or regions. AltaVista provides searching in 25 languages. Google and AltaVista also provide machine translation services for certain languages. In addition to providing easy access to more than 3 billion

Web pages, Google has many special features such as cached links, site search, link search, Web page translation, stock quotes, and more. Yahoo, the first human-compiled directory-based search engine, offers search results with directory-category links that have been reviewed by human experts. Microsoft's MSN Search provides a blend of human-powered directory information and crawler coverage different from any of the other top choices listed. It uses a Looksmart-powered directory, with secondary results from Inktomi. Other popular search engines in English include AltaVista (<http://www.altavista.com>), AOL (<http://search.aol.com/>), Ask Jeeves (<http://www.askjeeves.com>), HotBot (<http://www.hotbot.com>), etc. A number of these, such as Google, AltaVista, and Yahoo, have gradually expanded their services to non-English speakers.

However, there still exists a technology gap between systems in English and those in other languages which contributes to an information gap in various areas. Many Internet searching and browsing support techniques have been shown to be effective for English search engines, including meta-search (Chen et al., 2001; Meng et al., 2001; Selberg and Etzioni, 1995), document categorization (Chen et al., 2001; Hearst and Pedersen, 1996; Zamir and Etzioni, 1999), and summarization returned (McLellan et al., 2001; McDonald & Chen, 2002). Similar technologies for non-English languages are not as well developed. This is mainly due to the fact that non-English content brings many challenges that existing English-based information retrieval techniques do not address. For example, there are no explicit word boundaries in Chinese, making existing English indexing and

searching techniques not directly applicable in Chinese information retrieval systems. There is a desire to study how to generalize these techniques to non-English languages to support human information seeking.

1.3 Multi-lingual Web Retrieval

Although some online search engines have moved toward offering multilingual support, from the user's perspective these search engines are essentially a collection of monolingual search engines. There are a wide variety of circumstances in which a reader needs to search for documents in totally unfamiliar languages, for example, companies seeking international business opportunities, researchers seeking references and information on a particular topic, intelligence agencies researching global intelligence, etc.

Multilingual information retrieval (MLIR) appears to be a promising approach to addressing that problem. Multilingual Information Retrieval is the study of responding to a query by searching for documents in other languages (Hull and Grefenstette, 1996). MLIR is an extension of Bilingual Information Retrieval (BLIR), where target documents are in a single language that is different from the query (Chen et al., 2002). Both multilingual and bilingual retrieval are sometimes referred to in a broad sense as Cross-lingual Information Retrieval (CLIR). While bilingual information retrieval has been studied more extensively in different-language pairs such as English-Spanish, English-

Chinese, and English-Arabic, multilingual information retrieval has not been widely considered.

Most reported approaches translate queries into the document language, and then perform monolingual retrieval. There are three main approaches in MLIR: using machine translation, a parallel corpus, or a bilingual dictionary. The machine translation-based (MT-based) approach uses existing machine translation techniques to provide automatic translation of queries. The MT-based approach is simple to apply, but the output quality of MT is still not very satisfying, especially for western and oriental language pairs (Sakai, 2000; Aljlal et al. 2002). A corpus-based approach analyzes large document collections (parallel or comparable corpora) to construct a statistical translation model (Xu & Weischedel, 2000; Nie et al., 1999). Although the approach is promising, the performance relied largely on the quality of the corpus. Also, parallel corpora are very difficult to obtain, especially for western and oriental language pairs. In a dictionary-based approach, queries are translated by looking up terms in a bilingual dictionary and using some or all of the translated terms (Ballesteros & Croft, 1996; 1997; Oard and Wang, 2001). This is the most popular approach because of its simplicity and the wide availability of machine-readable dictionaries. Much research work has been reported and evaluation results have, in general, been satisfactory.

Most MLIR research has used standard Text REtrieval Conference (TREC) collections, predominately news articles, as their test set, but little research has investigated Web-

based MLIR systems. Because, as several researchers (Kando, 2002; Oard, 2002) have suggested, operational applications will be the next step in MLIR research, a need to study how to integrate MLIR techniques into a multilingual Web retrieval system has arisen. While traditional MLIR techniques are promising, they cannot be employed directly in Web applications. Several factors make multilingual Web retrieval different from traditional MLIR research. First, Web pages are more unstructured and are very diverse in terms of document content and document format (such as HTML or ASP). As a result, Web pages are typically much “noisier” than such standard collections as news articles and therefore need extensive work in document pre-processing. Second, traditional MLIR usually focuses on effectiveness, measured in recall and precision, whereas efficiency also is important to end users in Web retrieval scenarios. If a single search were to last for more than a few minutes, a user would probably lose interest.

1.4 Translation of Proper Names

In multilingual retrieval, queries often involves unknown words that cannot be found in dictionaries, known as out-of-vocabulary (OOV) terms (Chen and Lee, 1998). Most of these OOV phrases are proper names and are some of the most difficult phrases to translate (Al-Onaizan and Knight, 2001). Proper names, such as organizations, company names, product names, and person names, play an important role in search queries (Bian & Chen 2000). It was reported that 67.8%, 83.4%, and 38.8% of queries to Wall Street Journal, Los Angeles Times, and Washington Post respectively involved name searching

(Thompson and Dozier 1997). During translation between language pairs which employ the same alphabets (e.g., English/Spanish), proper names stay the same. For language pairs employ different alphabets (e.g., English/Arabic), proper names are translated phonetically, referred to as *transliteration*. For example, President “George Bush” is transliterated into Chinese as “乔治 布什” and the company name “SONY” is transliterated into Arabic as “سوني.” Being able to identify correct transliterations of proper names as well as identifying the origin of transliterated words would largely affect the precision of multilingual Web retrieval and would be beneficial in machine translation systems or Question Answering systems as well. While the identification of proper names has received significant attention, transliteration of proper names has not (Al-Onaizan and Knight, 2002).

Previous transliteration models can be categorized into four approaches: a rule-based approach (Darwish et al., 2001; Kawtrakul et al., 1998; Wan and Verspoor, 1998), a machine learning approach (Arbabi et al., 1994), a statistical approach (Knight and Graehl, 1997; Meng et al., 2001; Al-Onaizan and Knight, 2002; AbdulJaleel and Larkey, 2002) and a Web mining approach (Goto et al., 2001; Lu et al., 2004). A statistical approach is the most promising approach. It does not rely on rules or heuristics; training data can be obtained fairly easily; and it has achieved reasonably good accuracy in previous research.

Several research gaps have been identified in transliteration. First, little transliteration research has considered context information in the transliteration model. Second, most previous research studied only one pair of languages. However, alphabet-based languages and character-based languages have many different features that need to be considered in transliteration process. It is a challenge to develop a generic approach for name transliteration to support knowledge discovery in a multilingual content.

1.5 Research Framework

Traditional MLIR techniques have not been widely used and evaluated in Web applications. To address these issues, three research questions have been posed:

1. How can I develop a generic approach to facilitating Internet searching and browsing by integrating selected search engine and post-retrieval analysis techniques for non-English content?
2. How can I develop a generic approach for a multilingual Web retrieval system that incorporates both European and Asian languages?
3. How can I develop a translation model that addresses out-of-vocabulary proper nouns in multilingual Web retrieval?

The system development research process described in Nunamaker et al. (1991) has been adopted to study these research questions. Following this methodology, several prototype systems were developed and evaluated. Each system investigates the research questions

by studying how users interact with them and the techniques incorporated. Chapters 2, 3, 4, and 5 will present several system prototypes that have been designed, implemented and evaluated to address those research questions.

In Chapter 2, I present my work in CMedPort, a Chinese Web portal in the medical domain that not only allows users to search for Web pages from local collections and meta-search engines but also provides encoding conversion between simplified and traditional Chinese to support cross-regional search and document summarization and categorization. User studies were conducted to compare the effectiveness and efficiency of CMedPort with those of three major Chinese search engines.

Chapter 3 presents my research in developing and evaluating a multilingual English-Chinese Web portal that incorporates various CLIR techniques for use in the business domain. A dictionary-based approach was adopted and combines phrasal translation, co-occurrence analysis, and pre- and post-translation query expansion. The portal was evaluated by domain experts, using a set of queries in both English and Chinese.

Chapter 4 extends techniques and algorithms used in Chapter 3. I present the design and evaluation of a multilingual Web portal in the business domain in English, Chinese, Japanese, German and Spanish. Web pages relevant to the domain were collected. Search queries were translated using bilingual dictionaries while phrasal translation and co-occurrence analysis were used again for query translation disambiguation. Western and

Eastern language features and influence of linguistic resources on performance are discussed.

In Chapter 5, I describe a generic proper name transliteration framework which incorporates an enhanced Hidden Markov Model (HMM) and a Web mining model. I improved the traditional statistical-based transliteration in three areas: 1) incorporated a simple phonetic transliteration knowledge base; 2) incorporated a bigram and a trigram Hidden Markov Model (HMM); 3) incorporated a Web mining model that uses word frequency of occurrence information from the Web. The framework was evaluated on two different language pairs, English-Arabic and English-Chinese.

Chapter 6 summarizes contributions of this dissertation and suggests future directions.

CHAPTER 2

SUPPORTING MULTI-REGIONAL INFORMATION SEEKING: A STUDY IN THE CHINESE MEDICAL DOMAIN

2.1 Introduction

As discussed in Chapter 1, the Web is growing exponentially and information on it has become increasingly diverse and comprehensive. Online information in languages other than English is growing even faster. A recent report shows that the non-English online population has exceeded the English online population [Global Internet Statistics. <http://www.gltreach.com/globstats/>]. Globalization has been a major trend of the Internet. However, most research in information retrieval (IR) has involved only English language programs. As non-English speakers wish to access information in their native languages, there is a need to study how to facilitate information seeking in a multilingual world. As the second most popular language online, Chinese occupies 10.8% of Internet languages (China Internet Network Information Center, 2003). It would be desirable to take Chinese as an example to study how techniques used in English IR could facilitate IR in other languages, because Chinese differentiates from English in many aspects.

While the Web provides convenient information searching, users often face an information overload problem (Blair and Maron, 1985); a search employing a general-

purpose search engine such as Google can result in thousands of hits. This problem is even more serious for non-English-speaking users because most Internet searching and browsing techniques are developed for speakers of English. A generic approach to facilitating Internet searching and browsing in any language is sorely needed.

Other than information overload, there are also some additional problems specifically faced by non- English speakers. As the second most often used language on the Web, Chinese provides a good example. Chinese users in mainland China, Hong Kong, and Taiwan make up 12.2% of the total world online population. Although Internet users in these 3 regions share enormous common information needs, Chinese search engine developers usually find it difficult to provide information from all regions because of problems such as the diversity of data sources and encoding differences. This greatly hinders information sharing among users from different regions. A generic Internet searching and browsing approach is needed to address these problems.

In this research, I propose an integrated approach to facilitating Internet searching and browsing in non-English languages. I developed an experimental Chinese medical Web portal, called *CMedPort* (<http://ai30.eller.arizona.edu:8080/gbmed>), based on my proposed approach. I also conducted user studies to evaluate the performance of the proposed approach in assisting human information seeking behavior.

The remainder of the chapter is structured as follows. Section 2.2 reviews related research, including studies on information seeking behavior, technologies that support searching and browsing, and search engines and medical portals in English and Chinese. In Section 2.3, I present my research questions. Section 2.4 describes my research testbed. In Section 2.5, I propose a generic framework to support information seeking and present the architectural design and major components of a prototype system: CMedPort. Section 2.6 describes my experimental design. In Section 2.7 I discuss experimental results and lessons learned. Finally, in Section 2.8, I conclude my study and suggest some future directions.

2.2 Related Work

In this section, I review various issues related to Internet searching and browsing in a multilingual world. These include different information-seeking behaviors on the Internet, different Internet searching techniques, and some special issues in supporting Chinese Internet searching and browsing.

2.2.1 Information Seeking Behaviors on the Internet: Searching and Browsing

Information seeking can be viewed as a “process in which humans purposefully engage in order to change their state of knowledge” (Marchionini, 1995). A significant amount

of research contributes to more recent information-seeking models on the World Wide Web. Ellis (1989) proposed a general model of information seeking behaviors with six categories: starting, chaining, browsing, differentiating, monitoring, and extracting. Kuhlthau (1991) provided a model that focuses on the information search process from the user's perspective. Belkin's model (Belkin et al., 1993) addresses the idea of cognitive and situational aspects as the reason for information seeking. More recently, Meho and Tibbo (2003) further expanded Ellis' model and identified four additional information seeking behaviors: extracting, verifying, networking, and information management. These studies show that users' particular patterns of information seeking enter into users' mental models of search. Among all the information seeking behaviors identified, two typical types of Internet behavior have received the most attention from researchers: searching and browsing (Marchionini and Shneiderman, 1988; Carmel et al., 1992; Chen et al., 1998).

Internet searching is a process in which an information seeker uses a query to describe a request for information and the system must locate information that matches or satisfies that request (Chen et al., 1998). Through searching, individuals seek to retrieve either specific information on a given topic or a specific piece of information. In other words, he or she has a specific object or target in mind. Internet browsing has been defined by Marchionini & Shneiderman as "an exploratory, information seeking strategy that depends upon serendipity" and is "especially appropriate for ill-defined problems and for exploring new task domains" (Marchionini and Shneiderman, 1988). Through browsing,

individuals explore an information space to gain familiarity with it or to locate something of interest to them (Chen et al., 1998). Browsing reflects a mental model in which the target is comparatively unfocused.

2.2.2 Techniques Facilitating Information Seeking

Understanding of the information seeking process helps researchers develop tools and techniques to augment the process. The wealth of information available on the Internet has led to much research directed toward developing techniques to support Internet searching and browsing. In this section, I briefly review some of their strengths and weaknesses.

2.2.2.1 Web Search Engines

General-Purpose Search Engines. General-purpose search engines are the most popular tools to help users locate information on the Web. A search engine usually consists of the following components: (1) Spiders (a.k.a. Crawlers) to retrieve Web pages by recursively following URL links, (2) an Indexer to tokenize Web pages into words or phrases, (3) a Query and Ranking Engine to retrieve search results, and (4) a User Interface (Chau and Chen, 2003).

Currently, many general-purpose search engines are available on the Web, each having its own characteristics and preferred algorithm for indexing, ranking, and presenting Web

documents. For example, AltaVista and Google allow users to submit queries and retrieve Web pages in a ranked order, while Yahoo groups Web sites into categories, creating a hierarchical directory of a subset of the Internet. Most prevailing search engines, such as Google, are keyword-based (Arasu et al., 2001). Although their search speeds are fast, their results are often overwhelmingly numerous and imprecise. It is often difficult to obtain specialized, domain-specific information from them.

Domain-Specific Search Engines. Many domain-specific search engines (or vertical search engines) seek to support more effective searching by providing precise search results in particular domains and extra functionalities that are not possible with general search engines (McCallum et al., 1999). For example, LawCrawler (<http://www.lawcrawler.com>) allows users to search for legal information. CampSearch (<http://www.campsearch.com>) searches for summer camps for children and adults and Excite NewsTracker (<http://nt.excite.com>) searches for news articles.

A good domain-specific search engine should contain as many relevant, high-quality pages and as few irrelevant, low-quality pages as possible. To address this need, domain-specific search engines collect Web pages by using intelligent spiders that can predict whether a URL is likely to point to relevant material and thus should be fetched first. Some examples of algorithms that have been developed to guide spiders to locate Web pages relevant to desired domains include HITS (Charkrabarti, 1999), PageRank (Cho et

al., 1998), Hopfield Net (Chau & Chen, 2003), and Reinforcement Learning (McCallun et al., 1999).

Meta-Search Engines. Selberg and Etzioni (1995) suggested that by relying on a single search engine, users could miss over 77% of the references they would find most relevant. A study by NEC Research Institute drew some similar conclusions, revealing that Internet search engines cannot keep up with the Internet's growth and that each search engine studied covered only about 16% of total available Web sites (Lawrence & Giles, 1999). However, meta-search engines leverage the capabilities of multiple Web search engines and other types of information sources, providing a simple, uniform user interface and relieving the user of having to deal with different search engines and information overload (Chen et al., 2001; Meng et al., 2001; Selberg and Etzioni, 1995). For instance, SavvySearch (<http://www.savvysearch.com>) supports up to 100 engines and allows users to customize a selection of search engines. MetaCrawler (<http://www.metacrawler.com>) searches the Internet's top search engines, including Google, Yahoo, AltaVista, and Ask Jeeves. Copernic Agent (<http://www.copernic.com>) collates results from more than 1000 search engines and provides filtering, summarization, and analysis on search results. MedTextus (Leroy and Chen, 2002) and HelpfulMed (Chen et al., 2003a) are two medical search engines that meta-search Web pages as well as online medical databases and journals.

2.2.2.2 Post-retrieval Analysis

In most current search engine systems, returned results are presented as lists of ranked URLs. Post-retrieval analysis can be performed on a result list to help users quickly locate the information needed. Document categorization and summarization are the two major techniques used in post-retrieval analysis.

Summarization—Document Preview. Summarization is a post-retrieval analysis technique that provides previews of documents (Greene et al., 2000). It can reduce the size and complexity of Web document lists by offering concise representations of the documents returned (McLellan et al., 2001; McDonald & Chen, 2002). In a browsing scenario, summarization provides an efficient way to allow users to judge the relevance of a document and let the users decide whether or not the full text is worth viewing (Rush et al., 1964).

Two major approaches to text summarization are text abstraction and text extraction. Text abstraction, which generates grammatical sentences that summarize a document, involves a great degree of document processing and computation. Text extraction utilizes sentences from an original document to form a summary. It usually involves assigning importance scores to sentences, based on term frequency and other characteristics of the document, and top-scoring sentences are selected as a summary. Recent research in text summarization has focused on the text extraction approach (Hovy and Lin, 1999; McDonald and Chen, 2002).

Categorization--Document Overview. Users are often frustrated by long-list results returned by search engines. In a browsing scenario, it is highly desirable for a search system to provide an overview of the retrieved document set so that a user can explore a specific topic and gain a general view of a particular area of interest. Categorization has been shown to be a powerful post-retrieval document processing tool that can cluster similar documents and present the resulting clusters to the user in an intuitive and sensible way (Chen et al., 2001). Hearst and Pedersen (1996) and Zamir and Etzioni (1999) demonstrated that document categorization has the potential to improve performance in document retrieval.

Document categorization is based on the Cluster Hypothesis: “closely associated documents tend to be relevant to the same requests” (Rijsbergen, 1979). There are two approaches to categorization (Zamir and Etzioni, 1999). It can be based on individual document attributes, such as query term frequency, size, source, topic, or author for each document. NorthernLight (<http://www.northernlight.com>) is an example of this approach. Categorization can also be based on inter-document similarities. This approach usually includes some machine learning techniques. For example, the self-organizing map (SOM), which uses a neural network algorithm to cluster documents, has been incorporated in several information retrieval systems (Chen et al., 1998). In both approaches, documents need to be systematically segmented and indexed so that key

phrases can be identified. Examples of key phrase extraction techniques include AZ Noun Phraser (Tolle & Chen, 2000) and Mutual Information (Ong & Chen, 1999).

2.2.2.3 Chinese Text Processing

Textual information collected from the Web must first be processed by document indexing techniques. Indexing techniques have been widely studied for English documents, but these techniques do not apply to languages such as Chinese in which there are no explicit separators to indicate word boundaries (Chien & Pu, 1996). Kwok (1999) investigated three indexing techniques for Chinese that can be applied to virtually any languages without explicit boundaries: character-based (or 1-gram), bi-gram, and lexicon-based indexing. His results showed that character-based indexing is good and efficient while bi-gram and simple word-based indexing achieved higher precision. However, bi-gram indexing led to a large indexing term space and was not efficient. Lexicon-based indexing usually matched text to an existing word lexicon, but many valuable words could be missed if they were not included in the matching lexicon.

2.2.3 Information Seeking in a Multilingual World: Research Gaps

The Web provides convenient information searching and browsing. As more and more users on the Internet are from non-English-speaking countries, major search engines have attempted to provide search support in non-English languages. However, search engines originally designed for English speakers usually cover a very limited amount of non-

English content and cannot serve the information need of the fast growing non-English-speaking online population. Although some English online search engines have moved toward offering multilingual support, there still exists a technology gap between systems in English and those in other languages which contributes to an information gap in various areas. This is mainly due to the fact that non-English contents bring many challenges that existing English-based information retrieval techniques do not address. For example, there are no explicit word boundaries in Chinese, making existing English indexing and searching techniques not directly applicable in Chinese information retrieval systems.

2.2.3.1 Comparison between General-purpose Search Engines

A report from Nielsen/NetRatings in January 2003 rated Google (<http://www.google.com>), Yahoo (<http://www.yahoo.com>), and MSN (<http://www.msn.com>) as the most popular search engines in the United States (Sullivan, 2003). In addition to providing easy access to more than 3 billion Web pages, Google has many special features such as cached links, site search, link search, Web page translation, stock quotes, and more. Yahoo, the first human-compiled directory-based search engine, offers search results with directory-category links that have been reviewed by human experts. Microsoft's MSN Search provides a blend of human-powered directory information and crawler coverage different from any of the other top choices listed. It uses a Looksmart-powered directory, with secondary results from Inktomi.

Other popular search engines in English include AltaVista (<http://www.altavista.com>), AOL (<http://search.aol.com/>), Ask Jeeves (<http://www.askjeeves.com>), HotBot (<http://www.hotbot.com>), etc. A number of these, such as Google, AltaVista, and Yahoo, have gradually expanded their services to non-English speakers.

Chinese is spoken by most people in mainland China, Hong Kong, and Taiwan, and most Chinese search engines have been developed to serve one of these regions. The most popular are Baidu (<http://www.baidu.com>) and Sina (<http://www.sina.com>) in mainland China, Yahoo Hong Kong (<http://hk.yahoo.com>) in Hong Kong, and Yam (<http://www.yam.com.tw>) and Openfind (<http://www.openfind.com.tw>) in Taiwan. Baidu has indexed 200 million Chinese Web pages and is the backend engine for over half of the major Chinese portals. In addition to basic Boolean search, it provides cached links and Chinese character encoding conversion. Sina, the most popular Web Portal in mainland China, offers directory-based searching with over 10,000 subcategories and recently adopted Google's search technology. However, advanced features such as encoding conversion were not provided in Sina. Yahoo Hong Kong, another directory-based search engine, returns results in both simplified and traditional Chinese. Openfind provides cached links and term suggestion functions, while Yam integrates encoding conversion to support cross-regional search. Compared with English search engines, Chinese search engines face more challenging issues, have fewer indexed pages, and provide fewer features. Contents focus on their own regions and contain much local information related to people's daily life.

2.2.3.2 Comparison between Medical Domain Search Engines

I further compared vertical search engines in the medical domain in two languages, because medical information is among the most popular resources on the Web. Supporting medical information searching on the Web also has attracted much attention from researchers and search engine builders, so a comparison of tools and search engines developed to support English and Chinese medical information seeking on the Web can help further illustrate the research gaps between them.

Numerous sites have been built to provide access to medical information over the Internet in English. For example, the National Library of Medicine's Gateway (<http://gateway.nlm.nih.gov/gw/Cmd>) and CliniWeb (<http://www.ohsu.edu/clinweb/>) provide access to Web pages from reputable organizations and institutions. In Gateway, Web pages are indexed according to the UMLS Metathesaurus, and in CliniWeb, the MeSH tree hierarchy. MDConsult (<http://www.medconsult.com>) and Medscape (<http://www.medscape.com>) aggregate journals, books, news, clinical symposia, and continuing medical education resources. MedTextus (<http://ai.bpa.arizona.edu/go/medical/MedTextus.html>) (Leroy and Chen, 2001) and HelpfulMed (<http://ai.bpa.arizona.edu/helpfulmed>) (Chen et al., 2003a) are two systems that search Web pages as well as medical databases and provide automatic thesaurus and document clustering functions to lower the requirement for human intervention. .

For Chinese, the major Chinese medical portals include 999 (<http://www.999.com.cn>), MedCyber (<http://www.medcyber.com>), and WSJK (<http://www.wsjk.com.cn>) from mainland China, and TrustMed (<http://www.trustmed.com.tw>) from Taiwan. These portals have quite diverse content, ranging from general health to drugs, industry, research conferences, etc. However, surprisingly few of them incorporate a search function. Most medical portals serve as a medical content provider and their contents are manually updated. Only 999 provides a basic search function for Chinese medical information on the Internet. MedCyber and TrustMed provide a search function only within their own sites, while WSJK has no search ability. Most of these portals maintain a small collection of fewer than 10,000 pages and provide only the Chinese character version for their own region. Although there is a lot of medical information available on the Internet, few medical domain search engines have been developed. Most researchers need to rely on medical search engines in English which contain little Chinese information or they have to use general-purpose Chinese search engines.

My comparison shows that in the medical domain, the gap between English and Chinese search engines is even greater than that of general-purpose search engines. English medical domain search engines have incorporated plenty of advanced features such as meta-search, medical thesaurus, and document clustering, while there exist few medical domain search engines in Chinese.

2.3 Research Questions

Many Internet searching and browsing support techniques have been shown to be effective for English search engines, including meta-search, document categorization, and summarization. However, technologies for non-English languages are not as well developed. There is a desire to study how to generalize these techniques to non-English languages to support human information seeking.

Based on my review, I believe developing a generic framework with various searching and browsing support techniques promises to narrow the information gap between the English and non-English languages. In this study, I aimed to address the following research questions:

1. How can I develop a generic approach to facilitating Internet searching and browsing by integrating selected search engine and post-retrieval analysis techniques for non-English content?
2. Would this integrated approach be more effective and efficient in facilitating information searching and browsing than other existing Internet systems for non-English content?
3. What is the users' level of satisfaction toward this integrated approach in comparison with existing systems for non-English content?

The remainder of the chapter presents my work in studying these three questions.

2.4 A Research Testbed in the Chinese Medical Domain

Because of the importance of the Chinese language on the Web and the popularity of medical information, I selected the Chinese medical domain as my research testbed to investigate various issues in supporting Internet searching for non-English content.

As the largest non-English-speaking Internet population, the Chinese-speaking users in Mainland China, Hong Kong, and Taiwan make up 12.2% of the world online population (Global Internet Statistics. <http://www.greach.com/globstats/>). A recent report from China Internet Network Information Center (CNNIC, 2003) shows that Internet population in mainland China grew at a rate of 68% half-yearly, from 45.8 million in late 2002 to 68 million in early 2003, while Hong Kong and Taiwan are among the few regions in the world that have the highest Internet penetration rates. With the rapid growth of the Chinese online population, the need for information service is increasing dramatically.

Medical Web sites are among the most frequently visited Web sites on the Internet (Shortliffe, 1998). A tremendous number of Chinese medical information resources have been created on the Web, ranging from scientific papers and journals to general health topics and clinical symposia. These medical resources are of widely varied quality. However, Internet users are often frustrated when they try to look for Chinese health information online. The sheer volume of results returned by general Chinese search engines often overwhelms the users and there are few medical domain-specific search

engines built for Chinese users. Compared to the wide availability of English medical information services such as MEDLINE, CANCERLIT, and HelpfulMed (Chen et al., 2003a), Chinese medical information services are under-developed to meet the growing medical information needs of the Chinese users.

Various factors contribute to the difficulties of supporting Chinese information-seeking in the medical area. One important problem is the regional differences between mainland China, Hong Kong, and Taiwan. Although the populations of all three regions speak Chinese, they use different Chinese characters. People from mainland China, where simplified Chinese is used, usually find it difficult to read traditional Chinese that is used in Hong Kong and Taiwan, while people from the latter two areas also have similar problems in reading simplified Chinese. Moreover, simplified Chinese and traditional Chinese are encoded differently in computer systems. Simplified Chinese is usually encoded using the GB2312 system and traditional Chinese is encoded using the Big5 system. When searching in a system encoded one way, users usually cannot get information encoded in the other.

Furthermore, Chinese medical search engines in mainland China, Hong Kong, and Taiwan usually keep only information from their own regions, while it is desirable for users to track medical information in all regions. For example, during the SARS epidemic, users who wanted to find detailed information about SARS outbreaks and control in all regions had to use different systems. These factors greatly hinder the

medical information sharing among mainland China, Hong Kong, and Taiwan and result in information gaps between these regions.

2.5 Proposed Approach

2.5.1 An Integrated Knowledge Portal Approach

I propose to use an “integrated knowledge portal” approach to supporting Internet searching and browsing in a multilingual world. My portal approach adopts the common architecture of most search engines that collects Web documents, indexes them, and makes them searchable to information seekers. In addition, I add three key extensions to this basic structure: the generic language process ability, the integration of multiple information resources, and post-retrieval analysis ability.

2.5.1.1 Generic Language Process Ability

Based on my review in Section 2.2.3, I adopted the character-based indexing technique in my research. In addition, the positional information on the words or characters within a document was captured and stored such that when the query was a phrase, documents containing the exact phrase could be retrieved and given higher ranking than pages with separated words.

In addition to the basic character-based indexing, the ability to extract meaningful phrases from documents is also desired because such phrases are often useful for other analyses. For this purpose, I adopted a mutual information approach. The mutual information approach is a statistical method that identifies significant patterns as meaningful phrases from a large amount of text in any language (Church & Hanks, 1989; Chien, 1997; Ong & Chen, 1999). The approach is an iterative process of identifying significant lexical patterns by examining the frequencies of word co-occurrences in a large amount of text. I experimented with this approach in processing Chinese, Spanish, and Arabic text collections and got satisfactory results.

2.5.1.2 Integration of Multiple Information Resources and Post-Retrieval Analysis

As reviewed in a previous section, relying on a single document collection could result in low information coverage. In the proposed approach, my own collections were complemented by information collated from different regions and resources using meta-searching. Such an integration of multiple information resources has been shown to offer precise and diverse information and facilitate efficient information seeking (Chen et al., 2001).

Post-retrieval analysis is another important feature in my design that provides added value to searching and browsing. I adopted text summarization technique to provide a

preview of the search results and document categorization techniques to provide an overview of the search results. These techniques have been successfully applied in previous research (Greene et al., 2000; Chen et al., 2003a).

2.5.2 A Research Prototype in the Chinese Medical Domain: CMedPort

Based on my proposed approach, I developed CMedPort as a research prototype to investigate whether integrated techniques can help improve Internet searching and browsing in languages other than English. It uses a three-tier architecture (as shown in Figure 1). The main components are: (1) Content Creation; (2) Meta-search Engines; (3) Encoding Converter; (4) Chinese Summarizer; (5) Categorizer; and (6) User Interface. In this section, I discuss each major component in depth.

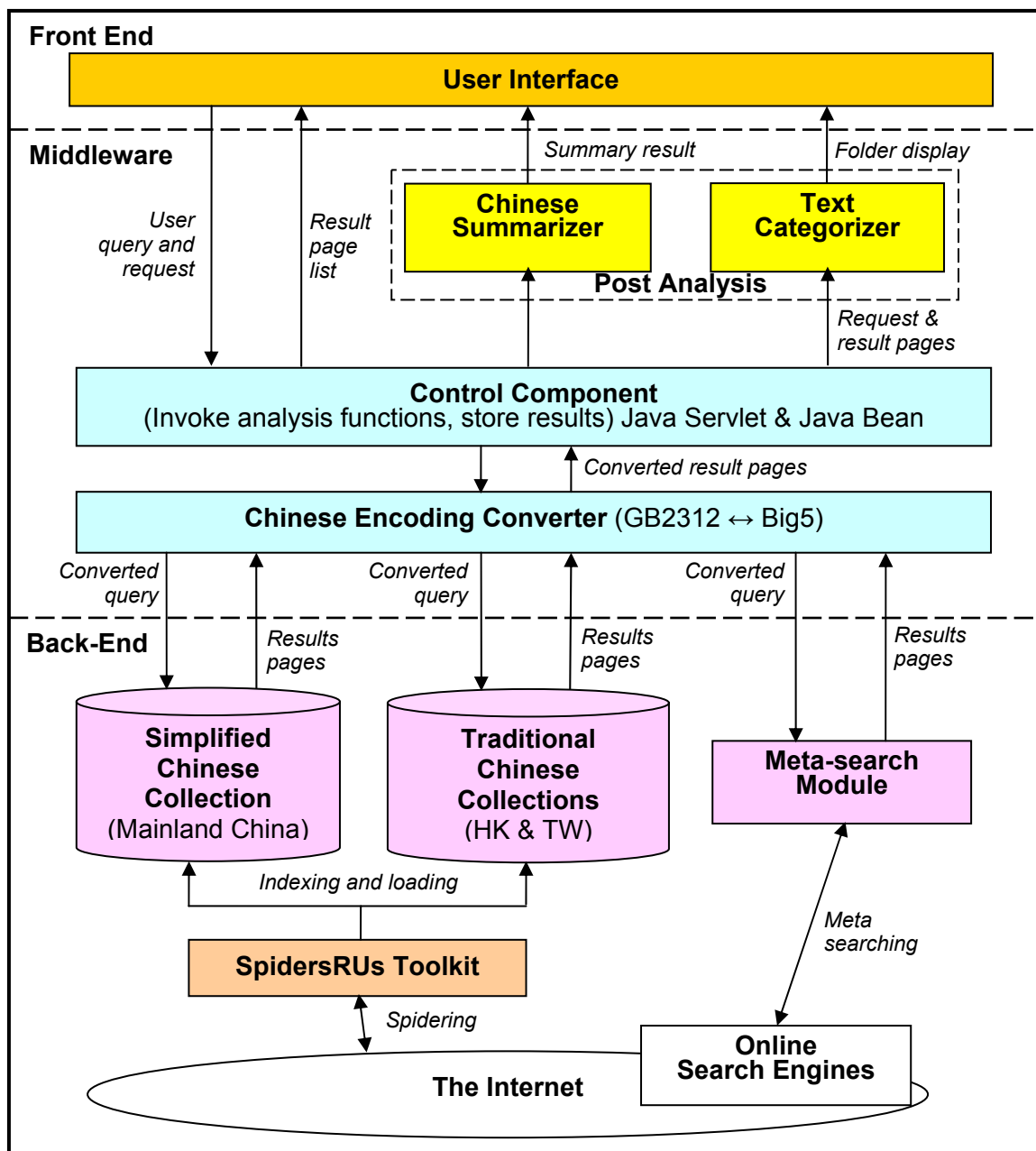


Figure 2.1: The CMedPort System Architecture

2.5.2.1 Content Creation

As a cross-regional search engine that covers medical information from mainland China, Hong Kong, and Taiwan, CMedPort needs to be able to collect information from all three regions. The AI Lab's SpidersRUs toolkit (<http://ai.bpa.arizona.edu/spidersrus/>), a digital library development tool developed by our research group, has been used to build collections for the Web portal. The toolkit contains components that support document fetching, document indexing, collection repository management, and document retrieval. It is able to deal with different encodings of Chinese (GB2312, Big5, and UTF8) and index different document formats, including HTML, SHTML, text, PDF, and MS Word. SpidersRUs also supports other languages, including English, Spanish, Arabic, etc.

Spidering

Based on suggestions from medical domain experts in the regions, 210 starting URLs were manually selected, including 87 from mainland China, 58 from Hong Kong, and 65 from Taiwan. They cover a large variety of medicine-related topics, from public clinics to professional journals, and from drug information to hospital information. Beginning with these medically related URLs, the SpidersRUs toolkit searched the Internet using a breadth-first search algorithm. It is assumed that medical pages included in the list will be likely to point to sites that they consider useful (Chen et al., 2003a). The three regional Web page collections created contain more than 300,000 Web pages in total. Web pages from mainland China, Hong Kong, and Taiwan were collected separately in

order to differentiate among sources and identify encoding schemes. During the spidering process I found more medical Web sites in mainland China and Taiwan than in Hong Kong. This observation is reasonable because mainland China and Taiwan have much larger online populations than Hong Kong, and Hong Kong residents very often use English medical Web sites.

Indexing

In CMedPort I used character-based indexing with positional information for document retrieval. Character-based indexing is known to be efficient and achieve high recall. In my approach, the positional information about words or characters within a Web page was captured and stored such that, when the query was a phrase, Web pages containing the exact phrase could be retrieved and given higher ranking than pages with separated words. This also ensures the precision of Chinese document retrieval and could be useful in working with languages lacking explicit word boundaries.

To perform advanced information retrieval techniques, such as document categorization and summarization, I need to extract meaningful Chinese phrases from textual information. In order to capture up-to-date phrases in my collection, I adopted the mutual information approach, a statistical method to identify significant lexical patterns examining the frequencies of word co-occurrences in a large amount of text (Ong & Chen, 1999). This approach computes how frequently a pattern appears in the corpus, relative to its sub-patterns. Based on the algorithm, the MI of a pattern c (MI_c) can be found by

$$MI_c = \frac{f_c}{f_{left} + f_{right} - f_c}$$

where f stands for the frequency of a set of words. Intuitively, MI_c represents the probability of co-occurrence of pattern c , relative to its left sub-pattern and right sub-pattern. Phrases with high MI are likely to be extracted and used in automatic indexing. For example, if the Chinese phrase “乙肝病毒” (hepatitis B virus, HBV) appears in the corpus 100 times, the left sub-pattern (乙肝病) appears 110 times, and the right sub-pattern (肝病毒) appears 105 times, then the mutual information (MI) for the pattern “乙肝病毒” is $100 / (110 + 105 - 100) = 0.87$. Phrases of all lengths were examined in the MI program. Furthermore, Stop words like “的” (of), “了” (function word, no meaning), and “及” (and) are removed. The included word list, which has priority over the stop-word list, allows users to have the flexibility to retain words that appear in the stop-word list. For example, the Chinese phrase “目的” (aim) can be listed in the included words although the word “的” (of) appears in the stop-word list. Using this approach I created simplified and traditional Chinese lexicons. Indexing against MI lexicon were then saved into a separate index file and later used by the CMedPort Categorizer.

The indexed files were loaded into a MS SQL Server database in which the data were separated into the three regions such that, during retrieval, the system could tell which region a Web page came from. Pages from each region were ranked by $tf*idf$ during

retrieval. Tf*idf combines the frequency of occurrence of every word in a document as well as the word's total occurrences in the collection, which indicated correlation between documents and a particular keyword.

2.5.2.2 Meta-search Engines

In addition to the regional collections, CMedPort also integrates information from different sources by meta-search engines. As discussed, few reputable medical-domain Chinese search engines are available for incorporation as information sources. Although some medical databases have been developed, online versions such as China Academy CBM (Chinese Biomedical Database) (<http://www.imicams.ac.cn/cbmdisc/cbmdisc.html>) are not stable. Based on suggestions from domain experts, six key Chinese search engines were chosen for meta-searching, two from each region (as shown in Table 2.1). These search engines contain a large portion of medically related information, usually from different parts of the Internet. Access to these systems provides a richer representation and a fresh coverage of information to supplement mylocal collection. More meta-search engines could be incorporated into CMedPort as they become available on the Internet.

Table 2.1: Major Chinese Search Engines in the Three Regions

Region	Information Source	Description
Mainland China	Baidu (www.baidu.com)	The biggest Internet search service provider in mainland China and has indexed millions of medical Web pages.
	Sina China (www.sina.com.cn)	The biggest Web portal in mainland China, containing more than 100,000 manually classified medical and health related Web sites.
Hong Kong	Yahoo! Hong Kong (hk.yahoo.com)	The most popular directory-based search engine in Hong Kong. Its “Health and Medicine” directory contains both public health and professional medical information.
	Hong Kong Government Information Center (search2.info.gov.hk)	A high quality search engine provided by the Hong Kong government, providing information including Hong Kong Health Department news, policies, etc.
Taiwan	Yam (www.yam.com)	The biggest Chinese search engine in Taiwan with a Health directory of more than 100,000 manually classified Web sites.
	Sina Taiwan (www.sina.com.tw)	One of the biggest Chinese Web portals in Taiwan containing information about hospitals, traditional Chinese medicine, etc.

When searching for information via CMedPort, a user is able to choose which meta-search engines are of interest and indicate how many results from each search engine he or she would like the system to retrieve. The default is 20 results from each source, but the user can choose up to 100. This gives the user some control over the retrieval process and lessens the effects of information overload. By sending queries to the search engines listed in the table, users could get results from all three regions, thus alleviating the problem of regional variations. While most meta-search engines merge the search results from the various data sources into a single list and re-rank the results using their own

heuristics, I decided to leave the results in separate categories (still in a single HTML page shown to the users) without merging them into a re-ranked list. The reason is that the meta-search results come from a diverse set of data sources as well as different regions. Merging these together may make it more difficult for users to locate the information they need.

2.5.2.3 Encoding Converter

In order to share medical information in different forms of written Chinese (simplified Chinese and traditional Chinese) among all three regions, an encoding conversion program is employed in CMedPort. The encoding converter uses a dictionary of 6,737 entries that map between simplified and traditional Chinese characters. Since some simplified characters map to multiple traditional equivalents, the conversion from simplified characters to traditional ones is sometimes ambiguous. I have picked the candidate character that is most frequently selected as equivalent to the original one.

In the simplified Chinese version of CMedPort, when a user enters a query in simplified Chinese the query is sent to all mainland China information sources using simplified Chinese. At the same time, the query is converted into traditional Chinese and sent to all information sources from Hong Kong and Taiwan that use traditional Chinese. When displaying results, the encoding conversion program is invoked again to convert results from traditional Chinese into simplified Chinese. The whole process is transparent to the

user. The encoding conversion program enables cross-regional search and addresses the problem of dissimilar Chinese character forms.

2.5.2.4 Chinese Summarizer

Automatic summarization has been applied as a document preview tool in many English retrieval systems. In CMedPort, a Chinese Summarizer was developed based on the sentence extraction approach. The Chinese Summarizer is a modified version of the AI Lab TXTRACTOR, a summarizer for English documents developed in my previous research (McDonald and Chen, 2002). Its major components include: 1) sentence evaluation, 2) segmentation or topic boundary identification, and 3) segment ranking and extraction. First, a sentence evaluation component parses the original Web page and extracts all sentences. These sentences are evaluated based on linguistic heuristics including presence of cue phrases (e.g. “总而言之” (in summary), “所以” (therefore)), $tf*idf$ score normalized for the sentence length, sentence position, and sentence length. Second, the TextTiling algorithm (Hearst, 1994) is used to analyze the Web page and determine where the topic boundaries are located. The Web page is thus segmented into its main topics. Third, the Summarizer ranks the document segments based on the scores given to the sentences and extracts high-ranking sentences from different segments as summary sentences.

The Chinese Summarizer was embedded in the CMedPort system and can be invoked at real time. Users can start the summarizer by choosing the number of sentences to be summarized under each returned result. The Chinese Summarizer dynamically retrieves the Web page on the Internet, processes the content, and presents a summary in a pop-up summarizer window. On the summarizer page, summary sentences are displayed on the left-hand side and the original Web page is displayed on the right-hand side with summary sentences highlighted. Users can click on any summary sentence on the left-hand side and go to the location of that sentence in the original page on the right-hand side. This feature is especially useful for browsing long documents where the Summarizer can help a user quickly determine whether or not a Web page is of interest.

2.5.2.5 Categorizer

Another component of CMedPort is the categorizer, which organizes returned results into various folders labeled by key phrases. When the categorizer is invoked all returned results are processed, and key phrases that have appeared in their titles and summaries are extracted by matching to the Chinese lexicons obtained from the Mutual Information program. As described in Section 2.5.2.1, the lexicons were constructed from fresh Web page collections and are more up-to-date and highly related to the medical domain than previous Chinese lexicons. Key phrases with high occurrences in returned results are extracted as folder topics. Web pages that contain a folder topic are included in that folder. One Web page may appear in multiple folders if it contains multiple folder topics.

I are using only title and summary to extract keywords because it is practical and permits dynamic categorization. Previous research has shown that clustering based on snippets is almost as effective as clustering based on a whole document (Zamir and Etzioni, 1999).

2.5.2.6 User Interface

CMedPort has two versions of User Interface to accommodate users from different regions: the traditional Chinese version and the simplified Chinese version. They look the same and provide the same functionalities, except that they use different encoding schemes and Chinese characters (simplified vs. traditional).

On the search page (see Figure 2.2.a), users can begin searching by typing keywords in the search box and indicating which local database and meta-search engines are to be searched. Multiple keywords can be entered into the search box at the same time, one keyword per line. Available information sources are organized into three columns by the region to which they belong and can be chosen by selecting the checkbox in front of a name.

On the result page, the top 20 results from each information source are displayed as ranked lists. Results from the CMedPort local collection are displayed in sorted order of relevancy to the query as measured by $tf*idf$ score. For each result in the lists, the title and a short summary are displayed (see Figure 2.2.b). All results of different encodings

are converted into the same encoding as the interface and displayed together (see Figure 2.2.c). By clicking on the name of a particular information source in the navigation bar at the top right-hand side of the page, users can go to the first result from that information source.

There is a drop-down box beneath each result in the list that users can use to select a sentence length and let the system automatically generate a 1-to-5-sentence summary of a Web page (see Figure 2.2.d). Users can also click on the ‘Analyze Results’ button to go to the analyzer page where all the results are categorized into folders with extracted topics. Clicking on the folders of interest produces a list of URL titles that is displayed under the relevant folder for him/her to browse (see Figure 2.2.e).



Figure 2.2: The CMedPort User Interface

2.6 Evaluation Methodology

In order to evaluate my system's performance in assisting Internet searching and browsing, a user experiment was designed and conducted. My study mainly addressed the following questions: (1) whether the integrated approach in CMedPort can facilitate searching and browsing of Chinese medical information more effectively and more efficiently than other existing Chinese search engines; (2) whether the summarizer and categorizer in CMedPort are effective and efficient support tools for browsing; and (3) whether CMedPort generates higher user satisfaction than other Chinese search engines. In this section, I discuss the experimental design and the results of my study.

2.6.1 Search and Browse Tasks

Consistent with the Interactive track in TREC (Text Retrieval Conference) evaluations (Voorhees & Harman, 1997), I gave users a list of questions. They are required to find answers using the given systems with their own queries. During the experiment, they were allowed to change their query words and do multiple retrievals. Since CMedPort has been designed to facilitate both searching and browsing, two types of tasks were designed: search tasks and browse tasks.

Search tasks in my user study were short questions that required specific answers. This type of task was designed to have just one correct answer. An example of a searching task would be "Which disease was Cocktail Therapy FIRST used on?" Since there is

always just one correct answer to the question, accuracy was used as the primary measure of effectiveness in searching tasks as follows:

$$\text{Accuracy} = \frac{\text{Number of correct answers given by the subject}}{\text{Total number of questions asked}}$$

In browse tasks, subjects were given a topic that defined an information need accompanied by a short description regarding the task. Subjects were expected to summarize the findings of their Web browsing session as a number of themes. A theme was defined as “a short phrase, which describes a certain topic” (Chen et al., 2001). Different from search tasks, there are always multiple themes that a user needs to address in browse tasks. An example of a browse task would be:

Topic: Liu Wei Di Huang Wan (A traditional Chinese medicine)

Description: Please summarize the effect of Liu Wei Di Huang Wan and diseases that it can be used on.

Theme identification has been used to evaluate performance of browse tasks (Chen et al., 2001; Chen et al., 2003b). Theme precision and theme recall were used as the primary measures of effectiveness in browsing tasks. They are defined as follows:

$$\text{Theme precision} = \frac{\text{Number of correct themes identified by the subject}}{\text{Number of all themes identified by the subject}}$$

$$\text{Theme recall} = \frac{\text{Number of correct themes identified by the subject}}{\text{Number of correct themes identified by expert judges}}$$

A theme is considered correct if it matches any of the themes identified by experts. By examining the themes that subjects came up with using different searching tools, I was able to evaluate how effectively each system helped a user locate relevant information from the Web.

Efficiency in both tasks was directly measured by the time that subjects spent on the tasks using different systems.

2.6.2 Benchmarks

To compare the performance of CMedPort with existing Chinese Search Engines, I selected one search engine from each region as a benchmark system. Existing Chinese medical portals were not considered suitable for benchmarks because they do not have adequate search functionalities and they usually search only inside their own Web sites. Thus, CMedPort was compared with three major commercial Chinese search engines from the three regions: Sina, Yahoo! Hong Kong, and Openfind. Although their content is not focused on the medical domain, their indexes of Chinese Web pages cover a large variety of medical information. Among the existing Chinese search engines they are the ones having the most functions comparable with those of CMedPort and are the most popular Chinese search engines in their respective regions.

2.6.3 Hypotheses

I tested three groups of hypotheses in my Chinese medical-domain search engine, CMedPort.

In hypothesis 1, I hypothesized that CMedPort would be more effective and efficient than benchmark search engines in search tasks where specific information is needed.

H1a: CMedPort is more effective than the existing benchmark Chinese search engines in searching.

H1b: CMedPort is more efficient than the existing benchmark Chinese search engines in searching.

In hypothesis 2, I hypothesized that CMedPort would be more effective and efficient than benchmark search engines in browse tasks where users need to become familiar with a topic. CMedPort integrates information sources from three regions and provides broad coverage of results. At the same time, I believed Chinese summarization and categorization would be post-retrieval analysis tools that would further improve effectiveness and efficiency in CMedPort.

H2.1a: CMedPort is more effective than the existing benchmark Chinese search engines in browsing.

H2.1b: CMedPort is more efficient than the existing benchmark Chinese search engines in browsing.

H2.2a: CMedPort's Chinese summarization further improves effectiveness in browsing.

H2.2b: CMedPort's Chinese summarization further improves efficiency in browsing.

H2.3a: CMedPort's Chinese categorization further improves effectiveness in browsing.

H2.3b: CMedPort's Chinese categorization further improves efficiency in browsing.

In hypothesis 3, I hypothesized that CMedPort would achieve higher user satisfaction in terms of usability. CMedPort provides a user-friendly interface with clearly organized information sources. The ability to do cross-regional search, summarization, and categorization would ease the process of information seeking.

H3: CMedPort achieves higher user satisfaction than existing benchmark Chinese search engines.

2.6.4 Experimental Design

Forty-five subjects, 15 from each region, were recruited for the experiment. Each subject was required to perform four search tasks and eight browse tasks. I specified more browse tasks because I also were interested in the performance of post-retrieval analysis tools in browsing. A time limit of 10 minutes was given for each task. Among the search tasks, two were performed using CMedPort and another two using the benchmark search engine from the subject's own region. Among the browse tasks, two were performed using CMedPort with general search function without summarizer or categorizer, two

used CMedPort with summarizer, two used CMedPort with categorizer, and two used the benchmark search engine. The order of the questions used in the tasks and the systems used in the experiment were rotated to avoid any potential bias. During the experiment, the tasks performed by subjects were timed and their answers were judged by domain experts. In order to test the CMedPort system as a whole, I did not specify the queries. The users chose their own query words. They could modify their query words anytime during the experiment.

At the end of each experiment, subjects were required to fill out a user satisfaction questionnaire. The questionnaire included 1) 19 Computer System Usability Questionnaire (CSUQ) items from Lewis (1995), specifically designed to assess aspects of usability, (Table 2.2); 2) questionnaire on individual components of CMedPort to evaluate the user perspective on cross-regional search, summarizer, and categorizer; and 3) open-ended positive or negative comments on CMedPort.

Three Chinese graduate students from the medical school at the University of Arizona, one from each region, were recruited as domain experts. They helped design the tasks used in my experiment and provided the standard answers for search and browse tasks. I did not control the type of websites experts got the answers from. They can use any search engines that they are comfortable with to provide the correct answers. The final version of standard answers was an agreed-upon version from the experts.

Table 2.2: Computer System Usability Questionnaire (Lewis, 1995)

Overall
Overall, I am satisfied with this system
Easiness of Use
Overall, I am satisfied with how easy it is to use this system
It was simple to use this system
I feel comfortable using this system
It was easy to learn to use this system
The information provided for the system is easy to understand
The information (such as online help, on-screen messages, and other documentation) provided with this system is clear
It is easy to find the information I needed
Effectiveness and Efficiency
I can effectively complete my work using this system
I am able to complete my work quickly using this system
I am able to efficiently complete my work using this system
I believe I became productive quickly using this system
The information is effective in helping me complete the tasks and scenarios
Error Recovery
The system gives error messages that clearly tell me how to fix problems
Whenever I make a mistake using the system, I recover easily and quickly
Interface
The organization of information on the system screens is clear
The interface of this system is pleasant
I like using the interface of this system
Functionality
This system has all the functions and capabilities I expect it to have

2.7 Experimental Results

In this section, I report the evaluation results and observations based on the user study.

2.7.1 Results from Search Tasks

Table 2.3.1 summarizes the systems' search-task performance by regions. In order to determine whether there were significant differences among the performances of the systems, paired t-tests were performed for each pair of methods. The statistical results (p-values) are shown in Table 2.3.2. Table 2.3.3 shows average improvement achieved in CMedPort compared to benchmark systems.

The results of testing hypotheses H1a and H1b showed that in terms of effectiveness CMedPort performed significantly (at $\alpha=0.05$) better than Sina and comparably to Yahoo HK and Openfind in search tasks. In terms of efficiency, CMedPort was significantly (at $\alpha=0.05$) better than Sina and Openfind and comparable to Yahoo HK. On average CMedPort achieved 20% higher accuracy and was 30% more efficient than the benchmark systems.

Table 2.3.1: Searching Performance of CMedPort and Benchmark Systems by Regions

Region	System Used	Accuracy	Time Spent (Sec)
Mainland China	Sina China	0.625	149.039
	<i>CmedPort</i>	0.917	97.962
Hong Kong	Yahoo HK	0.857	117.967
	<i>CMedPort</i>	0.929	95.033
Taiwan	Openfind	0.846	114.767
	<i>CMedPort</i>	0.962	72.433

Table 2.3.2: Hypotheses Testing for Search Tasks by Regions

H1: Search Task	Hypotheses	p-value	Result
H1a: Effectiveness	CMed > Sina	0.008*	Confirmed
	CMed > Yahoo	0.163	Not Confirmed
	CMed > Openfind	0.092	Not Confirmed
H1b: Efficiency	CMed > Sina	0.040*	Confirmed
	CMed > Yahoo	0.194	Not Confirmed
	CMed > Openfind	0.045*	Confirmed

Table 2.3.3: Searching Performance of CMedPort and Benchmark Systems with Combined Regions

	Accuracy	Improvement	Time Spent	Improvement
Benchmark Systems	0.776	-	127.3	-
CMedPort	0.935	20.5%	88.5	30.5%

2.7.2 Results from Browse Tasks

Table 2.4.1 summarizes the system performance in browse tasks. Hypotheses test results are shown in Tables 2.4.2 and 2.4.3.

The results of testing hypotheses H2.1a and H2.1b (Table 3.4.2) showed that CMedPort achieved significantly (at $\alpha=0.05$) higher theme precision than Yahoo HK and Openfind, and significantly (at $\alpha=0.05$) higher theme recall than all three benchmark systems for browse tasks. Meanwhile, users spent significantly (at $\alpha=0.05$) less time in CMedPort than in Sina and Openfind. On average I found 22.2% improvement in theme precision, 113% improvement in theme recall, and 19.5% improvement in efficiency comparisons between CMedPort and benchmark Chinese search engines. The results from browse

tasks are very encouraging. The high theme precision probably resulted from my focused collection building technique, including my character-based indexing approach for Chinese. The theme recall achieved is greater than I expected. This probably is because CMedPort has the ability to search information in all three regions and integrates results from different search engines by meta-searching.

The results of testing hypotheses H2.2a and H2.2b (Table 2.4.3) showed that there were no significant (at $\alpha=0.05$) differences in theme precision and recall when using or not using CMedPort's summarizer. No significant (at $\alpha=0.05$) differences in efficiency were found. Surprisingly, the Chinese summarizer did not further improve performance. I observed that the Chinese summarizer is not always fast. Before processing the page, the summarizer needs to fetch Web pages from remote servers and processing time is largely affected by the length of the Web page. Users usually lost patience after 10 seconds of waiting and would try other results. These factors affected the performance of the Chinese summarizer as a browsing support tool.

The results of testing hypotheses H2.3a and H2.3b (Table 3.4.3) showed that there were no significant (at $\alpha=0.05$) differences in theme precision and recall when using or not using CMedPort's categorizer. However, using CMedPort's categorizer was significantly (at $\alpha=0.05$) more efficient than not using the categorizer. Results suggested that, as a document overview tool, the categorizer could help users identify topics of interest more quickly when browsing.

Table 2.4.1: Browsing Performance of CMedPort and Benchmark Systems by Regions

Region	System Used	Theme Precision	Theme Recall	Time Spent (Sec)
Mainland China	Sina China	0.675	0.250	412.231
	<i>CMedPort Basic</i>	0.819	0.472	312.961
	<i>CMedPort Summarizer</i>	0.849	0.385	270.231
	<i>CMedPort Categorizer</i>	0.859	0.510	265.039
Hong Kong	Yahoo HK	0.651	0.228	376.700
	<i>CMedPort Basic</i>	0.790	0.524	360.400
	<i>CMedPort Summarizer</i>	0.878	0.517	454.800
	<i>CMedPort Categorizer</i>	0.825	0.506	286.133
Taiwan	Openfind	0.636	0.215	318.267
	<i>CMedPort Basic</i>	0.789	0.480	218.100
	<i>CMedPort Summarizer</i>	0.739	0.450	255.933
	<i>CMedPort Categorizer</i>	0.845	0.514	230.733

Table 2.4.2: Hypotheses Testing for Browse Tasks of CMedPort and Benchmark Systems

H2: Browse Tasks		Hypotheses	p-value	Result
H2.1a: Effectiveness	Theme Precision	CMed > Sina	0.071	Not Confirmed
		CMed > Yahoo	0.050*	Confirmed
		CMed > Openfind	0.031*	Confirmed
	Theme Recall	CMed > Sina	<0.001*	Confirmed
		CMed > Yahoo	<0.001*	Confirmed
		CMed > Openfind	<0.001*	Confirmed
H2.1b: Efficiency		CMed > Sina	0.003*	Confirmed
		CMed > Yahoo	0.290	Not Confirmed
		CMed > Openfind	<0.001*	Confirmed

Table 2.4.3: Hypotheses Testing for Browse Tasks of CMedPort Summarizer and Categorizer

H2: Browse Tasks		Hypotheses	p-value	Result
H2.2a: Effectiveness	Theme Precision	CMed Summ> CMed Basic	0.214	Not Confirmed
	Theme Recall	CMed Summ> CMed Basic	-0.201	Not Confirmed
H2.2b: Efficiency		CMed Summ> CMed Basic	-0.290	Not Confirmed
H2.3a: Effectiveness	Theme Precision	CMed Categ> CMed Basic	0.073	Not Confirmed
	Theme Recall	CMed Categ> CMed Basic	0.307	Not Confirmed
H2.3b: Efficiency		CMed Categ> CMed Basic	0.022*	Confirmed

2.7.3 Results from Usability Questionnaire

The results of testing hypotheses H3 show that CMedPort was rated significantly higher than all benchmark systems. CMedPort achieved an average rating of 5.83 out of 7, while all three benchmark Chinese search engines were rated below 5. Among all the evaluation criteria, CMedPort consistently gained higher ratings than benchmark Chinese search engines. Users appreciated the easiness, effectiveness, and interface of CMedPort. Among the three benchmark systems, Yahoo HK and Openfind were rated higher than Sina. This is consistent with what I observed in the performance of search and browse tasks.

Questionnaires on easiness and usefulness of featured tools in CMedPort showed that the categorizer achieved 6.174 out of 7, the highest among the three features, while the cross-

regional search feature was rated 5.949 and the summarizer was rated 5.665. No paired t-tests were performed on individual tools in CMedPort because benchmark systems do not provide functions comparable to those in CMedPort.

Table 2.5.1: User Satisfaction Rating of CMedPort and Benchmark Systems

	Sina	Yahoo HK	Openfind	CMedPort (Avg)
Overall	4.040	5.069	4.857	6.033
Easiness	4.813	5.322	5.402	6.081
Effectiveness	4.078	4.372	4.804	5.890
Error Recovery	4.440	4.552	4.750	5.363
Interface	5.026	5.477	5.095	5.988
Functionality	4.083	4.643	4.571	5.626
Average	4.413	4.906	4.913	5.830
Cross-regional Search	-	-	-	5.949
Summarizer	-	-	-	5.665
Categorizer	-	-	-	6.174

Table 2.5.2: Hypotheses Testing for User Satisfaction

H3: User Satisfaction	p-value	Result
CMedPort > Sina	<0.001*	Confirmed
CMedPort > Yahoo HK	0.003*	Confirmed
CMedPort > Openfind	0.001*	Confirmed

2.7.4 Subjective Feedback

The results of the searching and browsing experiments and usability questionnaire show that CMedPort achieved significant improvement on three regional benchmark systems. Users' subjective feedback provides further information about the differences in the system performance. I summarize several aspects that received the most comments.

Information Quality and Coverage: Twenty-one out of 45 subjects indicated that CMedPort gave more relevant results compared to benchmark systems. They expressed that “it is easier to find useful information using CMedPort,” while “Sina provides a lot of commercial company information as top results.” One Taiwanese student said, “CMedPort is especially useful when looking for information from other regions (mainland China and Hong Kong).” Two users pointed out that the benchmark systems gave few results when they searched for “抗肿瘤药” (anti-tumor drug), while CMedPort gave plenty of useful results.

Categorizer: Among the 45 subjects, 32 subjects expressed that they liked the Categorization function. One subject said, “Categorizer is very powerful and useful. I like this function very much.” Others said, “It (categorizer) makes searching more quickly. I can skip a lot of irrelevant information in results and focus on relevant ones.” Four subjects mentioned that the categorizer topics were not very relevant to what they were looking for sometimes. I observed that category topics sometimes contain noise from advertisements.

Summarizer: Twenty-five subjects expressed their preference for CMedPort’s Chinese summarizer. For example, one subject claimed that “I also like summary function. It becomes more useful when the article is very long.” Four of them mentioned that they liked the feature that summary sentences are highlighted in the original Web page. Eight subjects gave neutral comments on the summarizer. One said, “summarizer is useful

sometimes, but not all the time.” Nine subjects complained about the processing speed of the summarizer, although some of them liked the idea of summarizing Web pages.

Cross-regional Search: Eleven subjects commented that cross-regional search is useful. They complained that “it is difficult to find information from other regions in Openfind.” Similar comments were made on the other benchmark regional search engines. One Hong Kong user said: “It is helpful to convert simplified Chinese into traditional Chinese.”

User-Interface: Seventeen users mentioned that they liked the clear user-interface of CMedPort. They commented that “CMedPort looks professional and is easy to understand,” while benchmark search engines “always have advertisements floating around and make me lose focus.”

Speed: Nine subjects commented that benchmark search engines had faster processing speed when compared to CMedPort, which saved them information seeking time. However, as CMedPort is still a research prototype and does not have a powerful server to support the process, it would not be difficult to improve the speed of the system.

In general, subjects’ overall opinion on CMedPort tended to be positive, although some complained about its speed. In contrast, benchmark systems received relatively fewer positive comments.

2.8 Conclusions and Future Directions

In this Chapter, I have proposed a general framework for supporting multilingual Internet searching and browsing on the Web. The framework is consistent with the architecture of most search engines. However, my approach features three extensions to this basic structure: generic language processing ability, integration of multiple information resources, and post-retrieval analysis.

I have discussed the development of CMedPort, a Chinese medical portal to serve the information seeking needs of Chinese users. A systematic evaluation has been conducted to study the effectiveness and efficiency of CMedPort in assisting human analysis. My experimental results show that CMedPort achieved significant improvement in searching and browsing performance compared to three benchmark regional search engines, Sina, Yahoo! Hong Kong, and Openfind. I believe that CMedPort's collection building method, meta-searching, and cross-regional searching contributed to the improvement in information seeking. Although post-retrieval analysis methods, such as categorizer and summarizer, did not further improve browsing performance significantly, users' subjective evaluation and verbal comments revealed that they appreciated these analysis functions. Overall, the experimental results are promising.

However, this research has several limitations. Since my study is based on subjects from different Chinese-speaking regions, a large sample size is infeasible. I therefore reported results based on a limited number of 45 subjects. Additionally, because there are few

medical domain Chinese search engines available, I was only able to compare with general Chinese search engines that contain medical content as my benchmark systems. Since these Chinese search engines do not have post-retrieval analysis components in them, I were only able to compare basic functions in CMedPort with existing search engines.

In the future I plan to study the semantic differences between Simplified and Traditional Chinese to provide cross-regional information search functionalities for Chinese users. I am also in the process of applying my integrated approach to search engines in more languages such as Spanish and Arabic. Meanwhile, I plan to add multilingual information retrieval function to these portals. In addition, a dynamic, graphic knowledge map that could categorize multilingual documents retrieved from the Web is in development.

CHAPTER 3

FACILITATING CROSS-LINGUAL WEB RETRIEVAL:

AN EXPERIMENT IN ENGLISH-CHINESE BUSINESS INTELLIGENCE

3.1 Introduction

As I discussed in Chapter 2, there are Web pages in almost every popular non-English language including various European, Asian, and Middle East languages. Consequently, it is often difficult for an English speaker to access non-English content on the Web and, in general, it is difficult for a user to retrieve documents written in a language that is not spoken by that user. As a result, retrieval from the Web of documents in different languages presents a very interesting and challenging research problem.

Cross-language information retrieval (CLIR), the study of retrieval information in one language through queries expressed in another language, appears to be a promising approach to addressing that problem. CLIR has been studied widely in different languages, such as English, Chinese, Spanish, and Arabic. Much research work has been reported and evaluation results have, in general, been satisfactory. Most systems have demonstrated performance similar to that for monolingual retrieval, i.e., traditional document retrieval in one language. Most CLIR research has used standard TREC

collections, predominately news articles, as their test set, but little research has investigated Web-based CLIR systems. Because, as several researchers (Kando, 2002; Oard, 2002) have suggested, operational applications will be the next step in CLIR research, a need to study how to integrate CLIR techniques into a multilingual Web retrieval system has arisen.

While traditional CLIR techniques are promising, they cannot be employed directly in Web applications. Several factors make multilingual Web retrieval different from traditional CLIR research. First, Web pages are more unstructured and are very diverse in terms of document content and document format (such as HTML or ASP). As a result, Web pages are typically much “noisier” than such standard collections as news articles and therefore need extensive work in document pre-processing. Second, traditional CLIR usually focuses on effectiveness, measured in recall and precision, whereas efficiency also is important to end users in Web retrieval scenarios. If a single search were to last for more than a few minutes, a user would probably lose interest. In addition, queries on Internet have an average length of 2.21 (Spink & Xu, 2000), which is considered as short queries in information retrieval. In this study, I posed the following research questions: 1) Can CLIR techniques achieve satisfactory performance for retrieving Web documents that are much “noisier” than traditional text collections? 2) Can I combine existing CLIR techniques to build a multilingual Web portal with both satisfactory effectiveness and efficiency?

To address these problems, my research has investigated the feasibility of applying CLIR techniques to Web applications by developing and evaluating an English-Chinese cross-language Web portal that utilizes various techniques. The rest of the Chapter is structured as follows. Section 3.2 reviews related research, including three fundamental approaches to CLIR, translation ambiguity problems and query expansion techniques. I also discuss problems in using existing CLIR techniques in Web applications and present my research questions. In Section 3.3, I propose my Web-based multilingual retrieval prototype. Section 3.4 discusses the system architecture and implementation details of a prototype English-Chinese Web portal called ECBizPort. I also show an example of how a search query will be translated and expanded by my system, using different CLIR techniques. Section 3.5 reports the setup and results of an experiment designed to evaluate the performance of the prototype. Finally, in Section 3.6 I conclude my work and suggest some future directions.

3.2 Literature Review

In this section, I review CLIR techniques that are related to my research. Because CLIR involves finding documents in languages other than the query language; it has relied heavily on different techniques for translating the search query from the source language to the target language. Most research approaches translate queries into the document language, and then perform monolingual retrieval. In section 3.2.1, I review three major query translation approaches, namely a machine-translation approach, a

corpus-based approach, and a dictionary-based approach. In section 3.2.2, I review several translation disambiguation techniques that have been used to reduce errors introduced during query translation. I also review, in Section 3.2.3, applications of CLIR in Web-based systems.

3.2.1 Query Translation Approaches

Salton (1972) discussed the “controlled vocabulary” approach, one of the earliest practical CLIR approaches, but its requirement to index a document collection manually makes it unsuitable for high-volume applications (Oard, 1997). Other than the use of controlled vocabulary, most research has studied free text retrieval systems, in which there are three main approaches: using a machine translation (MT) system, using a parallel corpus, or using a bilingual dictionary.

3.2.1.1 Machine Translation-based Approach

The machine translation-based (MT-based) approach uses existing machine translation techniques to provide automatic translation of queries. Sakai (2000) used MT Avenue, a free web-based Japanese-English translation service, and achieved reasonable effectiveness with the aid of pseudo-relevance feedback. Aljlal et al. (2002) used ALKAFI, a commercial Arabic-English MT system and studied the effects of query length on MT-based CLIR. This approach is simple to apply, but the current output

quality of machine translation is still not very satisfactory, especially for western and oriental language pairs, because typical Web search queries lack the contextual information which is necessary for MT to perform word sense disambiguation correctly (Sakai, 2000). Off-the-shelf MT systems may miss the correct translation for a word even when it is among the original candidates in the MT dictionary (Jones et al., 1999). This also affect the effectiveness of MT-based approach.

3.2.1.2 Corpus-based Approach

A corpus-based approach analyzes large document collections (parallel or comparable corpora) to construct a statistical translation model. Landauer and Littman (1991) developed a corpus-based technique called Cross-language Latent Semantic Indexing (CL-LSI), which is a language-independent approach. Although the approach is promising, the performance relied largely on the quality of the corpus. Davis and Dunning (1995) applied evolutionary programming on a parallel Spanish-English collection, and reported 75% of monolingual IR performance. Sheridan and Ballerini (1996) applied thesaurus-based query expansion techniques on a comparable Italian-English collection. More recently, in TREC 9, the BBN group used two parallel corpora (Hong Kong News and Hong Kong Law) to translate English query words into Chinese (Xu & Weischedel, 2000). A corpus-based approach does not depend on manually built bilingual dictionaries and is good for emerging domains where bilingual dictionaries are not available. However, a parallel corpus is very difficult to obtain, especially for

western and oriental language pairs such as English and Chinese. Even those that are available tend to be relatively small or to cover only a small number of subjects. To deal with this problem, Nie et al. (1999) investigated the possibility of automatically gather parallel text from the Web. Their Web mining approach showed the feasibility of using the Web as potential corpus.

3.2.1.3 Dictionary-based Approach

In a dictionary-based approach, queries are translated by looking up terms in a bilingual dictionary and using some or all of the translated terms. This is the most productive area in CLIR because of its simplicity and the wide availability of machine-readable dictionaries. Ballesteros & Croft (1996; 1997) investigated dictionary-based Spanish-English CLIR and reported that using both pre-and post-translation query expansion was more effective than using either one separately. Later, they applied co-occurrence analysis with a query expansion technique and achieved 91% of monolingual retrieval precision. Hull and Grefenstette (1996) studied the Spanish-English pair using structured queries. Oard and Wang (2001) discussed Pirkola's structured queries and balanced translation. Chen et al. (2000) focused on short-query translation by combining multiple English-Chinese sources. Dictionary-based approaches are relatively easy to implement, and bilingual machine-readable dictionaries (MRDs) are more widely available than parallel corpora. However, there are always unknown words that are not covered in dictionary and much research studied translations of these out-

of-vocabulary words (Lu et al., 2003). Researchers have also identified several challenges to this approach: (1) multiple definitions of a word could introduce noise into the translated query (a.k.a., ambiguity); (2) failure to translate technical/new terminology which is often not found in general dictionaries; (3) failure to translate multi-term concepts as phrases (Ballesteros & Croft, 1996).

3.2.2 Reducing Translation Ambiguities and Errors

It has been shown that when simple dictionary translations are used without addressing the problem of translation ambiguity, the effectiveness of CLIR can be 60% lower than that of monolingual retrieval (Ballesteros & Croft, 1998). Several techniques proposed to reduce the ambiguity and errors introduced during query translation have been used with each of the translation approaches discussed earlier. In the following, I briefly review three of them, i.e., phrasal translation, co-occurrence analysis, and query expansion.

3.2.2.1 Phrasal Translation

Phrasal translation techniques are often used to identify multi-word concepts in a query and to translate them as phrases. Hull and Grefenstette (1996) and Chen et al. (2002) showed that effectiveness of CLIR is significantly improved when phrases in queries are manually translated. It has also been reported that the effectiveness of CLIR can be

improved by using phrase information in machine-readable dictionaries (Davis & Ogden 1997; Ballesteros & Croft 1998). Kwok (2000) reported his successful experience in extracting phrase information from a Chinese/English bilingual wordlist. The major problem of using phrasal translation is that many phrases are not covered by dictionaries and thus cannot be translated correctly.

3.2.2.2 Co-occurrence Analysis

In order to improve the correctness of query translation, it is also popular to use co-occurrence statistics to select the best translation(s). The assumption here is that the correct translations of query terms tend to co-occur more frequently in target language documents than incorrect translations. Co-occurrence analysis techniques rely on corpora for target word selection. Some recent research has investigated using collection of Web search engines as the corpus (Wang et al., 2004) and achieved promising results. Co-occurrence analysis has been successfully used to resolve translation ambiguity in many previous studies (Ballesteros & Croft, 1998; Maeda et al., 2000; Gao et al., 2001; Sadat et al., 2002) and some improved co-occurrence analysis methods have been suggested (Nie et al., 1999). All previous studies using co-occurrence analysis disambiguation have reported dramatic improvement in CLIR performance. However, the heavy computational and storage requirements of co-occurrence analysis have limited its practical use in retrieval systems where efficiency is a major concern.

3.2.2.3 Query Expansion

Query expansion has been considered in many CLIR studies. The assumption of query expansion is that additional terms that are related to the primary concepts in a query are likely to be relevant and that adding these terms to the query can reduce the impact of incorrect equivalents generated during the translation (Ballesteros & Croft, 1996). McNamee and Mayfield (2002) studied the effectiveness of query expansion for various resource qualities. They strongly recommended using query expansion when high quality resources are not available. When such resources are available, however, query expansion does not help a lot.

Query expansion may be local, as in the local feedback method (also known as pseudo relevance feedback). The local feedback method is often used for query expansion in CLIR. It involves only the top ranked documents retrieved by the original query. The most frequently appearing terms and phrases from those top ranked documents are added to the query. Queries are both reweighed and expanded based on this information (Attar & Fraenke, 1977; Croft & Harper, 1979). Other query expansion methods also may use both local and global information, such as local context analysis (LCA) method (Xu and Croft, 1996).

Expansion may take place before query translation, when it is referred to as pre-translation query expansion, or after translation, when it is known as post-translation query expansion. In a combined pre- and post-translation query expansion, queries are

first expanded before translation, the expanded queries are then translated, and the translated queries are expanded again, after which the final expanded queries are used for retrieval. Research has arrived at different conclusions about query expansion. While some achieved significant improvement when using pre- and post-translation query expansion, others gained very little change for the better (Gey & Chen, 2000).

3.2.3 CLIR for Web applications

As discussed earlier, most research has focused on the study of technologies that improve retrieval precision on standard TREC collections, rather than on real-world, interactive Web retrieval applications.

In addition to CLIR systems designed for general text documents, some Web-based CLIR systems also are available. Some of them are Keizai, Arabvista, ECIRS, and MULINEX. Keizai, developed at the New Mexico State University, is an interactive Web-based CLIR system, which accepts English queries and returns Japanese and Korean documents (Ogden et al., 1999). It provides a user-aided translation disambiguation, which allows the user to select a translation from the candidates. ECIRS (http://www.cs.nmsu.edu/~sliu/cgi-bin/e-c_search/index_concord.pl) is an English-Chinese Web-based system with a relatively small collection of Chinese documents (Liu, 2001). It uses a simple dictionary-based approach without further translation disambiguation and query expansion support. Arabvista

(<http://www.arabvista.com/>) is a commercial search engine developed by Emirates Internet and Multimedia for Middle East users. With an English or Arabic query, it could retrieve Web pages in multiple languages, including Chinese, French, and German. However, the collection favors some languages. A simple query like “computer system” could not get any results from Chinese or Japanese. MULINEX is a comparatively more mature multilingual Web search and navigation tool for English, French and German, developed in DFKI Language Technology Lab (Capstick et al., 1998). It incorporates Web spiders, concept-based indexing, relevance feedback, translation disambiguation, document categorization, and summarization functionalities. It also translates retrieved documents into the users’ language such that the users can read them.

Among these systems, MULINEX uses a more comprehensive approach than the others. However, the major problem for most of these systems is that no systematic evaluations are available, leaving the effectiveness of these systems uncertain.

3.2.4 Summary

The Web has become a major information source world-wide for people in any knowledge field. Although the use of CLIR techniques in Web retrieval is expected to address the multilingual information needs of Web users. Each of the three CLIR translation approaches has some drawbacks, such as availability of required resources.

A corpus-based approach often suffers from a lack of high-quality parallel or comparable corpora. A machine translation-based approach has limited effectiveness especially when short queries are involved. A general dictionary-based approach cannot deal with new terminologies that often are used in Web documents.

Another major problem is that traditional CLIR techniques have not been widely used and evaluated in Web applications (Oard, 2002). Although a few systems exist, most provide only simple translation functions and do not provide comprehensive evaluation results. In addition, most previous studies were conducted using document collections provided by TREC or other similar organizations. These collections usually consist of documents carefully selected for evaluation purposes. Web pages are much more diverse and dynamic, distributed on many servers. They contain extensive html meta tags, however these meta data are not standardized and are often missing. For instance, many Web documents do not have meta information of the coding system it uses, which could cause problem in indexing. Different from traditional text documents, Web page typically contains many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements. Information contained in these noisy blocks can harm retrieval performance. Thus, making using of CLIR techniques on a highly diverse and sometimes “noisy” Web page collection is still questionable. It would be interesting to study the performance of different CLIR techniques in Web-based applications.

Based on my review, I believe CLIR techniques are a promising response to the challenge of development of practical multilingual Web retrieval systems and Web portals, especially when query translations are combined with various translation disambiguation techniques. In this study, I posed the following research questions:

- Can CLIR techniques achieve satisfactory performance for retrieving Web documents that are much “noisier” than traditional text collections?
- Can I combine existing CLIR techniques to build a multilingual Web portal with both satisfactory effectiveness and efficiency?

The remainder of the chapter presents my work in studying these two questions.

3.3 Proposed Approach to Multilingual Web Retrieval

Aiming to apply an integrated set of CLIR techniques to the Web environment, I propose an architecture for multilingual Web portal development. The proposed system architecture is shown in Figure 3.1.

My system architecture consists of five major components: (1) Web Spider and Indexer, (2) Pre-translation Query Expansion, (3) Query Translation, (4) Post-translation Query Expansion, and (5) Document Retrieval. In the following, I describe each component in detail.

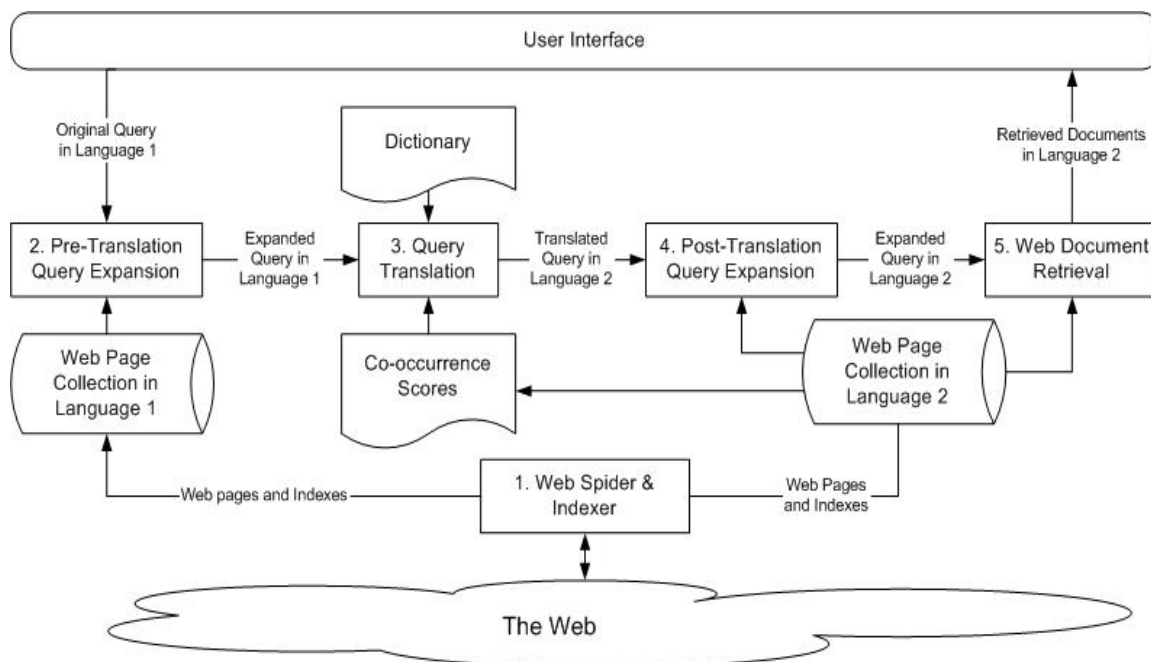


Figure 3.1: CLIR System architecture

3.3.1 Web Spider and Indexer

Web spiders, or Web crawlers, are programs that retrieve pages from the Web by recursively following URL links in pages using standard HTTP protocols (Cheong, 1996). The Web Spider component is responsible for document collection building. Document collections in two or more languages are needed for a multilingual Web portal not only as an information resource for users but also as a comparable corpus that can be used for translation disambiguation and query expansion. The second purpose is especially important because it is difficult to obtain parallel or comparable corpora for particular domains.

There are high requirements for quality of the collections; every Web page in the collection has to be highly relevant to the selected domain and, at the same time, must be diverse enough to cover multiple topics and interests in the domain. Focused Web spiders — spiders that focus on collecting pages in specific domains or Web sites — can be used to build domain-specific Web collections (Chau & Chen, forthcoming). However, although traditional focused Web crawling methods can be used to create Web page collections that meet the relevance requirement, the scope of collections built by these methods is usually restricted to the topics to which the starting URLs relate (Bergmark et al., 2002). As a result, simple focused crawlers often fail to provide comprehensive coverage of the different topics within the domain. To address this problem, I proposed a collection-building method that extends the capability of traditional focused crawling by meta-searching multiple large search engines. The process of my method can be described as follows: Similarly to traditional focused crawlers, I start the “probing” of the Web with a set of starting URLs and fetch relevant Web pages. At the same time, new starting URLs are being identified by querying multiple search engines (e.g., Google, Yahoo, AltaVista, etc.) and combining their top results. Using this method, the diversity of the collection meets the requirement by combining the top results from multiple search engines (Lawrence & Giles, 1998) while the relevance of the collection is retained.

Web pages also must be indexed differently from traditional text documents. Documents from the Web can be in various formats, such as HTML, ASP, or JSP. Web-

specific indexers are designed to work with specific Web page structures (e.g., removing markup tags from HTML documents).

3.3.2 Pre-translation Query Expansion

In my Web portal I undertook pre-translation query expansion to expand users' queries in the original language. As discussed earlier, there are two common ways to perform pre-translation query expansion, namely local feedback and local context analysis. I chose to use the local feedback method because of its higher efficiency, an important factor for Web applications. My approach followed the method reported by Ballesteros and Croft (1997). The Pre-translation Query Expansion component takes a search query and sends it to the local document collection to perform a search. The top n documents retrieved are analyzed. All terms from these documents are extracted and their $tf*idf$ scores calculated. $tf*idf$ is term frequency multiplied by inverse document frequency, a measure widely used in information retrieval applications. The expanded query is then reweighed with the Rocchio formula (Xu & Croft, 1996).

3.3.3 Query Translation

The Translation component is the core of the system. It is responsible for translating search queries in the source language into the target language. Among the three translation approaches, the dictionary-based approach seems to be most promising for

Web applications for two reasons. First, compared with the parallel corpora required by the corpus-based approach, MRDs used in dictionary-based CLIR are much more widely available and easier to use. The limited availability of existing parallel corpora cannot meet the requirements of practical retrieval systems in today's diverse and fast-growing Web environment. Second, compared with MT-based CLIR, the dictionary-based CLIR approach is more flexible, easier to develop, and easier to control. While it is impractical to build a complex MT system just for CLIR, existing commercial MT software is either packaged as a black box, leaving little space for users to modify it for their specific purposes, or it is too costly. According to a previous study (Gao et al., 2001), dictionary-based CLIR with a combination of disambiguation techniques can achieve even better performance than high-quality MT systems. I proposed to use a dictionary-based approach combined with phrasal translation and co-occurrence analysis for translation disambiguation.

In the dictionary lookup process, I first conducted *maximum phrase matching* on English queries. Sequence of English words that matches a dictionary entry was identified as a phrase. If such phrase was identified, it would be assigned a higher score than individual words. The longest sequence identified was assigned the highest score as a phrase. In addition, the entry with the smallest number of translations were preferred over other candidates, because such translation candidates are less ambiguous than entries with large number of translations. Translations containing more continuous keywords were ranked higher than those containing discontinuous keywords.

Co-occurrence analysis also was used to help choose the best translation among candidates. For each pair of terms $\{p, q\}$ in the query, all possible definition pairs $\{D_p, D_q\}$ in the dictionary were extracted such that D_p is a definition of query term p in the target language and D_q is a definition of query term q in the target language. Each pair was used as a query to retrieve documents in the indexed collections. The co-occurrence score between two definitions D_1 and D_2 then could be calculated as follows:

$$Co - occur(D_1, D_2) = \frac{N_{12}}{N_1 + N_2}$$

where N_{12} is the number of Web pages returned when performing an “AND” search using both D_1 and D_2 in the query and N_1, N_2 are the numbers of documents returned respectively when using only D_1 or D_2 in the query. My method is similar to that of (Maeda et al., 2000) in which they sent definition pairs to other search engines and used the number of returned documents to calculate the co-occurrence scores. What differentiates my proposed method from theirs is that they calculated the co-occurrence score “on the fly” which may greatly lower system efficiency; I calculated co-occurrence scores in advance to avoid affecting run time efficiency which are extremely important for Web applications.

3.3.4 Post-translation Query Expansion

The Post-translation Query Expansion component is responsible for expanding the query in the target language. Similarly to pre-translation expansion, I followed the

method described by Ballesteros and Croft (1997). The translated query is sent to the local document collection in the target language to retrieve the relevant documents. All terms from the top n documents are extracted and ranked by $tf*idf$ scores. The top terms are then combined with the translated query and reweighed to build the final query.

3.3.5 Document Retrieval

The Document Retrieval component is responsible for taking the query in the target language and retrieving the relevant documents from the text collection. This component can be designed based on similar the retrieval component in traditional information retrieval systems. Different ranking methods, such as frequency-based ranking or PageRank, also can be incorporated in this component (Arasu et al., 2001).

3.4 ECBizPort: An English-Chinese Web Portal for Business Intelligence

In this section, I report my experience in implementing a multilingual Web retrieval system using the dictionary-based CLIR approach. The Web portal, called ECBizPort, is an English-Chinese Web portal for business intelligence in the information technology (IT) domain. I found that the whole building process can be done relatively quickly and easily by making maximum use of monolingual retrieval system development techniques and tools. I will also discuss some important issues in multilingual Web retrieval system development.

Figure 3.2 shows two sample screenshots of the ECBizPort prototype. A user can enter a search query in the box provided and choose among different translation methods. The query will be passed to the system for query translation and query expansion. A set of relevant documents will be retrieved by the system and returned to the user. The translated and expanded query is also displayed to the user so he/she may use the terms to refine the query manually.

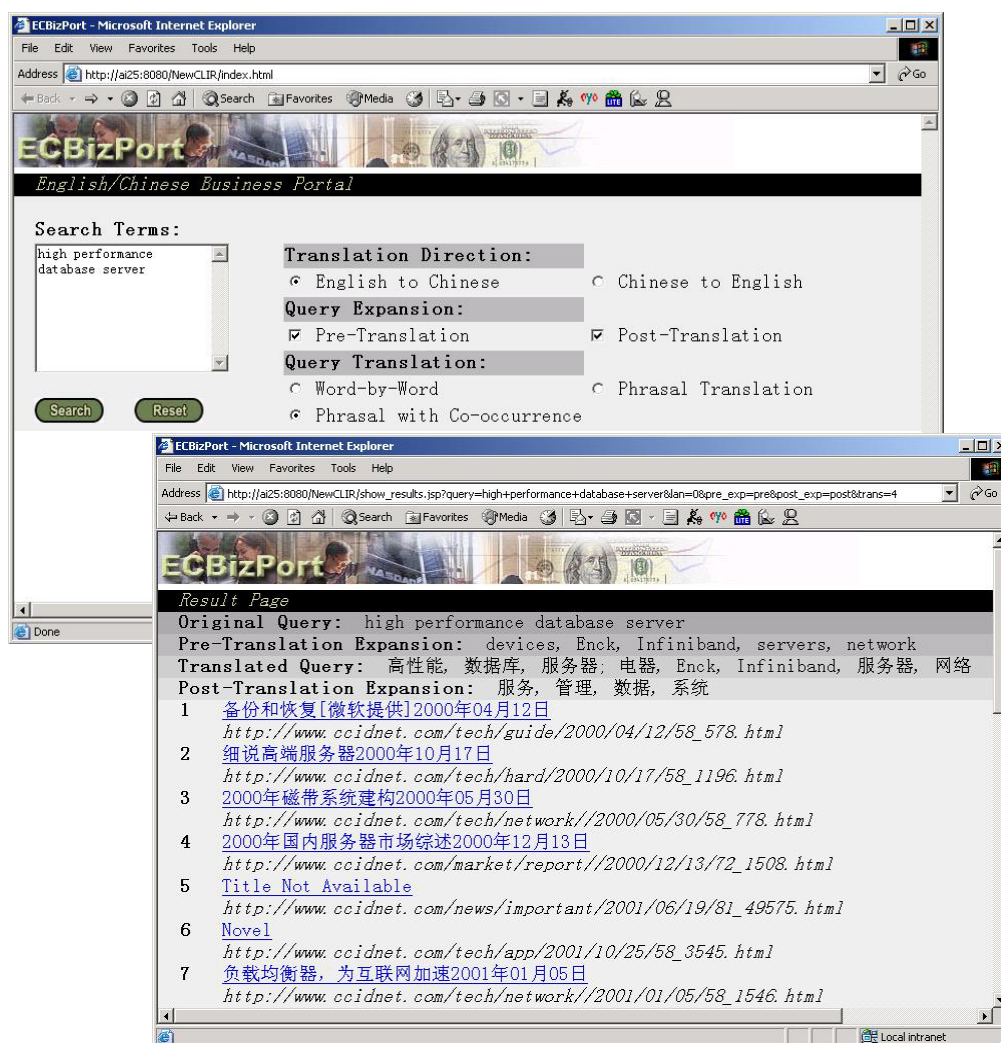


Figure 3.2: Sample screenshots of ECBizPort

3.4.1 Domain Selection

I decided to implement an English-Chinese Web portal for IT business intelligence because I believe that such a multilingual Web portal will be very useful; English and Chinese are the most popular languages on the Web and a strong business partnership between the U.S. and China has given rise to a great IT business information need between the two countries. Another reason is that Chinese is the second most popular language online. Chinese and Chinese users represent 10.8% of the Internet (Global Reach, 2002), making it desirable to study English-Chinese as a language pair in CLIR, where English queries are used to match against Chinese documents.

3.4.2 Web Spider and Indexer

The AI Lab SpidersRUs toolkit (<http://ai.bpa.arizona.edu/spidersrus/>), a digital library development tool developed by my research group, is used to build the English and Chinese collections for the Web portal. The toolkit contains components that support document fetching, document indexing, collection repository management, and document retrieval. It can also build collections in different languages and encodings.

To address the limitation of most existing focused crawlers which use local search algorithms in Web searching, I used a meta-search enhanced focused crawling approach (Qin et al., 2004) to build the ECBizPort collections. Similarly to traditional focused crawlers, my crawler starts with a set of starting URLs and fetches relevant pages back

based on the content- and link-based analysis results. Outgoing links in the relevant pages are extracted and put into the URL queue. At the same time, a meta-searching component keeps drawing queries from a domain-specific lexicon, retrieving diverse and relevant URLs by querying multiple search engines, and combining their top results. For the Chinese collection, I used 20 IT/Business-related starting URLs suggested by my domain expert, such as <http://www.zgcsc.com/> and <http://www.csdn.net/>. A domain lexicon of 100 typical IT/business-related keywords were created. During the spider process, these keywords identified by the expert were sent in Chinese to five major search engines, Google (<http://www.google.com/>), Yahoo China (<http://cn.yahoo.com/>), Sina (<http://www.sina.com.cn/>) Sohu (<http://www.sohu.com/>), and Baidu (<http://www.baidu.com/>). The top results returned from each search engine were used as new starting URLs. Spiders were set to fetch 100,000 pages. Running on a Pentium-4 PC, the spiders spent about six hours collecting 100,000 Chinese Web pages. Similarly to the Chinese collection, the English collection consisted of 100,000 pages and was built in about five hours with 16 expert-identified starting URLs, a 100-keyword lexicon and five major general search engines, namely Google, Yahoo (<http://www.yahoo.com/>), AltaVista (<http://www.altavista.com/>), Infoseek (<http://infoseek.go.com/>), and Hotbot (<http://www.hotbot.com/>). By using this meta-search enhanced spider algorithm, I obtained two IT/Business domain specific collections, one in English and the other in Chinese.

To support document retrieval, English Web pages in the ECBizPort were indexed using a word-based indexing approach; Chinese Web pages were indexed using a character-based indexing approach. In both approaches, the positional information on the words or characters within a Web page was captured and stored such that when the query was a phrase, Web pages containing the exact phrase could be retrieved and given higher ranking than pages with separated words.

3.4.3 Query Translation

3.4.3.1 English-Chinese Dictionary Translation

Query term translations were performed using the LDC (Linguistic Data Consortium) English-Chinese bilingual wordlists as dictionaries (http://www ldc.upenn.edu/Projects/Chinese/ LDC_ch.htm). The LDC wordlists include two specific lists: the English-to-Chinese wordlist (“ldc2ec”) and the Chinese-to-English wordlist (“ldc2ce”), each contains around 120,000 entries. The main reason for choosing the LDC wordlists was that the Chinese-to-English wordlist could be used as a comprehensive word dictionary as well as a phrase dictionary. Taking advantage of the phrasal translations, Kwok (2000) reported that using the Chinese-to-English wordlist alone improved the effectiveness of CLIR by more than 70%. Similar phrasal translation techniques were adopted in my Web portal.

Each entry in the dictionary was indexed. For example, the indexer could interpret the information of the English term “IT” having three Chinese translations “它”, “情报技术”, and “信息技术”:

它	/ <i>it</i> /
情报技术	/ (military) Intelligence Technology / <i>IT</i> /
信息技术	/ Information Technology / <i>IT</i> /

The relationships between the English term (IT) and the Chinese translations (“它”, “情报技术”, and “信息技术”) were captured and recorded. Some other important information, such as the number of English terms found in one dictionary entry and the positions of the term located in the entry, also was captured and stored for disambiguation purposes.

Given the indexed dictionary, definitions of English terms could be quickly and easily retrieved. The Web page ranking function of the retrieval component could be used to perform further disambiguation. For example, assume the follow dictionary entries have been indexed:

情报	/ intelligence /
情报技术	/ (military) Intelligence Technology / IT /
智慧	/ wisdom / knowledge / wits / intelligence /

信息	/ information /
技术	/ technique / technology /
信息技术	/ Information Technology /IT/
信息技术产业	/ Information Technology industry / IT Industry /

For the English term “intelligence”, three definitions were retrieved: “情报技术” (intelligence technology), “智慧” (wisdom, intelligence), and “情报” (spy intelligence). The definitions then were sorted according to the number of English terms found to be related to each definition. The Chinese definition with the smallest number of English translations was ranked first. In this way, “情报” was selected as the best definition of “intelligence” since each was the only translation for the other. Maximum phrase matching was also incorporated in my system by ranking Chinese translations containing more continuous key words higher than those containing discontinuous key words. For example, for the English terms “information technology”, the definition “信息技术” containing continuous keywords “information” and “technology” was selected as a phrase translation, rather than the two separated definitions “信息” and “技术”. Similarly, the English terms “information technology industry” would be translated into “信息技术产业”, a three-word phrase rather than three separated terms or a single word and a two-word phrase.

As discussed earlier, co-occurrence analysis also was incorporated in my system. It was implemented by extracting all the terms that appeared in my dictionary from the documents in the ECBizPort collections. The co-occurrence scores were calculated in a batch process and stored in a database.

3.4.3.2 Query Expansion

To get good and meaningful expansions, the two collections had to be indexed using a comprehensive and up-to-date lexicon to get as many good phrases as possible. As SpidersRUs uses word-based indexing (character-based indexing for Chinese) to avoid information loss, it did not capture phrases in either language during my general indexing process. This led to loss of semantic meaning since in most cases a meaningful term contains more than two characters in Chinese. Although the LDC wordlist can alleviate the problem by providing some phrases, it is not sufficiently up-to-date and comprehensive or the IT business domain, which involves a lot of new terminologies.

To address that problem, I decided to extract key phrases from my collection to build my own lexicon. Arizona Noun Phraser (AZNP), developed by my research group, was used to extract phrases from the English collection (Tolle & Chen, 2000). AZNP has three components: a word tokenizer, a part-of-speech tagger, and a phrase generation module. Its purpose is to extract all noun phrases from each document based on linguistic rules. The Mutual Information (MI) technique, also developed by my group,

was used to extract key phrases from the Chinese collection (Ong & Chen, 1999). The MI program uses a statistical PAT-tree approach to extract key phrases from Chinese documents. The English collection, with 100,000 Web pages, was sent to AZNP to build the English lexicon, while the Chinese collection of the same size was sent to MI to build the Chinese lexicon. These two collections were indexed based on their two respective lexicons. The indexed terms were used for both pre- and post-translation query expansion.

The local feedback method was implemented for both pre- and post-translation query expansion in my system. For pre-translation expansion, the top ten English Web pages were retrieved by the original English query, and the five English terms/phrases with the highest $tf*idf$ scores were added to the original query. In post-translation expansion, the top ten Chinese Web pages were retrieved, and the top five Chinese terms/phrases were added to the translated Chinese query. In both pre- and post-translation expansion, the terms in the expanded query were reweighed using the Rocchio formula (Xu & Croft, 1996).

3.4.4 Document Retrieval

The document retrieval component was supported by the AI Lab SpidersRUs toolkit and the design was relatively straightforward. After a target query had been built, it was passed to the search module of the toolkit. The search module searched the document

indexes and looked up the documents that were most relevant to the search query. The retrieved documents then were ranked by their $tf*idf$ scores and returned to the user through the Web-based interface.

3.5 An Example of Query Translation and Expansion

In this section I give an example of a typical user session in which a user tries to find some Chinese Web pages related to the English key phrases “IT industry” and “development environment”. So, he typed “IT industry development environment” into ECBizPort and clicked the “Search” button.

If the user had chosen the word-by-word translation method, the system would first look up all the translations for each English word. The results are shown as follows with each translation separated from the next by a space:

IT: 之 它 自称无所不知的 自己做 自己做方式的 重要的是 正好 应该说
 应战 由 由 由此可见 信息技术 兴 往常 巧 其 岂 恰好 莫非 莫非 莫如
 目前还不清楚 没关系 没有差别 雷峰塔 可见 可见 可谓 可惜 据说 据
 悉 看 看来 看来 看来 看上去 看上去 技术情报 假装博学多闻的人 简
 单的说 画蛇添足 过不去 对不对 对头 当然 传说 不客气 不了了之 不
 买账 不买账 不巧 不如 不谢 不屑 不言而喻 不言而喻 不要紧 不要紧

不依 不待说 不得不 不迭 不敢当 不好意思 不及 敝帚自珍 本来 板上
钉钉 板上钉钉 包干 抱薪救火 罢休 罢休 白饭 安土重迁

industry: 业界 子工业面 支柱产业 业界标准 行业 通讯行业 松下电气工业 食
品加工业 企业集团 矿业 建筑业 计算机工业 罐头工业 国营企业 工
商 工商界 工业 工业的巨头 大型企业 大型企业 产业

development: 发展 开发过程 开发环境 开发周期 开发周期 经济发展 技术发展 动
态 大力发展

environment: 周围 作业环境 运算环境 研制过程 虚拟环境 网路环境 实时操作环
境 联网环境 开发环境 环境 分布式环境 操作环境

The word-by-word translation method then picked the first match from each translation, resulting in the following query (the English explanations are given in parentheses):

之 (it) 业界 (industry) 发展 (development) 周围 (surroundings)

Although each translated Chinese term is one possible correct translation of the corresponding English word, none of them is the correct translation in the context of this query. For example, although “业界” can be translated into “industry”, it is not common to use this word as part of the translation of the phrase “IT industry” in Mandarin Chinese. Also, “周围” often refers to surroundings as in “natural surroundings” rather than “environment” as in “development environment”.

This translation method certainly did not satisfy the user. So, he tried again using the phrasal translation method. The returned translations were listed as follows:

之 (it) 业界 (industry) 开发环境 (development environment)

The phrasal translation method greatly improved the translation results and the two-word concept, “development environment”, in the original query had been successfully identified and translated as the phrase “开发环境”. However, the translation for “IT” still was incorrect, and the translation for “industry” still did not fit well into the context of the original query. Therefore, the user tried again, using the phrasal translation method with co-occurrence analysis disambiguation. The returned translation is shown below:

信息技术 (information technology) 产业 (industry) 开发环境 (development environment)

At this point, all the phrases have been correctly translated and fitted well into the context of the original search query. To further improve the retrieval performance, the user also chose to perform pre- and post- translation query expansion. The results are shown below:

Microsoft, database system, Operating System, software development, computer engineering

The pre-translation query expansion is incorporated into the original query which then could be translated together to get the following query:

信息技术 (information technology) 产业 (industry) 开发环境 (development environment) 微软 (Microsoft) 数据库系统 (database system) 操作系统 (operating system) 软件 (software) 开发 (develop) 计算机 (computer) 工程学 (engineering)

Then, the post-translation query expansion results also could be added into the query.

市场 (market), 公司 (enterprise, firm), 开发 (development), 电脑 (computer), 软件 (software)

Therefore, the final query was:

信息技术 产业 开发环境 微软 数据库系统 操作系统 软件 开发 计算机 工程学
市场 公司 开发 电脑

I can see that all new expanded query words were IT-related terminologies and could improve precision of retrieving Web pages using the translated query.

3.6 System Evaluation

In order to evaluate the performance of my system, an experiment was designed and conducted. In this section, I discuss the experimental and results of my study.

3.6.1 CLIR Evaluation Methodologies

CLIR evaluation aims at testing the effectiveness, measured by precision and recall, of retrieval systems. To make effectiveness comparable across systems, the tests usually are carried out on a common data set.

Three major evaluation workshops that have provided test collections for CLIR experiments are: the Cross-Language Evaluation Forum (CLEF) covering many European languages, the NTCIR Asian Language Evaluation covering Chinese, Japanese and Korean, and the TREC Cross Language Track from 1997-2002. In all workshops, the task was to match topics in one language against documents in another language and return a ranked list (Gey & Chen, 2000; Chen et al., 2002; Peters, 2002). In these tasks, a set of documents from the subject collection was pre-judged by human experts to be relevant to the original query. Sometimes, since it would have been

unrealistic to expect every document in the collection to be judged, precision is often of more interest than recall, and is usually reported at low recall levels. In other words, precision rate is often reported for the top n retrieved documents (n usually being between 5 and 1000). Once relevance judgments had been established, precision could be computed upon the ranked list of each entry.

Since CLIR always yields precision loss compared to traditional monolingual information retrieval, researchers are often interested in how well CLIR performs as compared with the corresponding monolingual run. In a monolingual run, original queries were manually translated into the target query, and retrieval was performed on this translated query. When precisions for both CLIR and monolingual retrieval are obtained, the precisions of CLIR and of the monolingual retrieval can be compared.

The average precision for simple CLIR runs, such as word-by-word query translation without using any translation disambiguation or query expansion, are often compared with runs under other conditions to show the superiority of different techniques (Ballesteros & Croft, 1996; Oard & Wang, 2001).

3.6.2 Experiment Design and Measure

In general, I followed the TREC evaluation process in my experiment design. However, because my study used Web pages instead of standard collections, no established

relevance judgment was available for precision and recall. Therefore, I decided to create relevance judgment by recruiting human experts and adopting the precision at top-n retrieved documents ($n = 10$ in my experiments) as my primary performance measure. Since I were particularly interested in how well these techniques would work for Web content in a business intelligence portal, I recruited experts in the business domain. Three business school graduate students, all fluent in both English and Chinese, served as domain experts. They identified seven Chinese queries of interest in the business/IT domain and translated these queries into English as the base queries. These 7 base queries were: China IT industry development, database management system, quality control, electronic signature, high speed Internet, hardware interface, and Web application. The average length of these queries is 2.43, which is pretty close to the average Web query length of 2.21 (Spink & Xu, 2000). The original Chinese queries were used to get monolingual runs. As discussed, such monolingual retrieval represents the performance of traditional information retrieval. The English base queries were used to get cross-lingual runs based on five settings: word-by-word translation (WBW), phrasal translation with co-occurrence analysis (Ph-Co), Ph-Co with pre-translation expansion, Ph-Co with post-translation expansion, and Ph-Co with both pre- and post-translation expansion.

The experts individually submitted each query to the system under the different settings. Because it was not practical for the experts to read all the 100,000 Web pages in the collection, I emphasized precision only for the top ten retrieved Web pages for each

query and setting, which is referred to as target retrieval (Eguchi et al., 2002). The results were compared with the two standard benchmark settings: (1) monolingual information retrieval (the best-case scenario), and (2) word-by-word translation (the worst-case scenario). Word-by-word translation picks the first translation in the dictionary and ignores all the other translation candidates. With seven queries and six different settings, each expert performed a total 42 searches using the system. Each expert went through the top 10 Web pages returned for each query and gave each page a score of 0 or 1, with 0 meaning irrelevant and 1 meaning relevant to the search. The time spent for retrieval was also recorded as a measure of the efficiency of the system.

3.7 Experimental Results and Discussions

Table 3.1 shows the experimental results. The different settings and methods are shown in the first column of the table. The second column shows the top-10 precision scores for each method averaged across the 21 searches performed by each of the three experts. The third column shows a method's performance compared with that of the monolingual retrieval. The fourth column shows how much each method's precision was improved in comparison with the WBW translation. The last column shows the average time used for the retrieval.

Table 3.1: Precision and time

Method	Average top-10 Precision	Performance compared with Monolingual	Improvement compared with WBW	Time Used (Sec)
Monolingual	0.671	100.0%	-	7.1
WBW	0.283	41.8%	0.00%	14.0
Phr-Co	0.491	73.1%	74.6%	25.1
Phr-Co-Pre	0.491	73.1%	74.6%	45.1
Phr-Co-Post	0.500	74.5%	78.0%	52.1
Phr-Co-Pre-Post	0.500	74.5%	78.0%	71.8

Monolingual: monolingual retrieval

WBW: word-by-word translation

Phr-Co: phrasal translation with co-occurrence disambiguation

Phr-Co-Pre: phrasal translation with co-occurrence disambiguation and pre-translation query expansion

Phr-Co-Post: phrasal translation with co-occurrence disambiguation and post-translation query expansion

Phr-Co-Pre-Post: phrasal translation with co-occurrence disambiguation and both pre- and post-translation query expansion

3.7.1 Precision

In Table 3.1, it can be seen that all methods except word-by-word translation achieved over 70% of the performance level of the monolingual system. In other words, when using English queries to retrieve Chinese documents, the experts were able to achieve a top-10 precision rate of more than 70% of the top-10 precision obtained when using Chinese queries to search for Chinese documents. The results were encouraging and comparable with what others have reported for traditional CLIR systems and improvement in the range of about 60% to 90%. The results demonstrated that CLIR techniques assisted users in searching for documents in a different language in a noisy Web portal setting.

In order to identify any significant differences among the performances of the various translation techniques, paired *t*-tests were performed for each pair of methods. The statistical results (p-values) are shown in Table 3.2.

Table 3.2: Paired *t*-test results

vs.	Phr-Co	Phr-Co-Pre	Phr-Co-Post	Phr-Co-Pre-Post
WBW	0.0002*	0.0002*	0.0001*	<0.0001*
Phr-Co		1.0000	0.1623	0.6487
Phr-Co-Pre			0.6657	0.1623
Phr-Co-Post				1.0000

*The difference is statistically significant at the 0.1% level.

As shown in Table 3.2, all four methods based on phrasal translation and/or query expansion performed significantly better than word-by-word translation, the baseline translation method. However, there were no significant differences among the four methods. Phrasal and co-occurrence disambiguation performed much better than word-by-word translation, achieving a 74.6% improvement, more than I expected. This probably resulted because co-occurrence disambiguation in the focused business IT domain is likely to perform better than it does with general news articles, although it achieves significant improvement with both sets. Using phrasal and co-occurrence disambiguation, 72% of the query words were correctly translated.

Surprisingly, when combined with phrasal and co-occurrence disambiguation, pre-translation expansion did not further improve performance. Post-translation expansion, used alone or combined with pre-translation expansion, slightly improved performance, but the improvement was not as significant as I had expected. I suspect the noisy factor

of Web pages might caused the limitation of query expansion results. Commercial content, advertisement, menu bars, etc. are mixed with the relevant business intelligence content on each page, the expanded words were not from the page content, but those advertisements. This will affect the retrieval performance. I observed that Chinese Web pages were usually noisier than English Web pages, which made it difficult to obtain high-quality terms for expansion. Acronyms of companies and products that appeared frequently in English Web pages were often added to the query, but most of them remained un-translated and did not improve performance.

3.7.2 Efficiency

Efficiency is another important aspect of Web retrieval. Long system response time (time elapsed between the moment when the search button is clicked and the results finally appear on the screen) can cause users to lose their patience and thus lower user satisfaction. To investigate the effect of CLIR techniques on system efficiency, I conducted a simulation in which system response times for performing various CLIR tasks were recorded and compared. As system response time also depends on factors such as hardware performance and network traffic, I analyzed the processes of different CLIR techniques and made a baseline estimation of their effect on system efficiency. My results showed that phrasal translation with co-occurrence disambiguation took 3.5 times longer than monolingual translation. When both pre- and post-translation expansions were used, the retrieval time increased to 10 times that of monolingual

retrieval, which reached 70 seconds. It should be noticed that my prototype was run on a personal computer which is much less powerful than machines used in commercial search engines. The retrieval time would be much shorter on a powerful machine in a real Web retrieval system. With most calculations done during indexing time, the efficiency of the prototype is satisfactory.

In summary, the prototype multilingual Web retrieval system achieved 74% to 78% performance improvement when compared with word-by-word translation. Phrasal translation with co-occurrence disambiguation greatly improved precision, while query expansion translation did not further improve the performance. In terms of efficiency, a multilingual retrieval took 25 to 71 seconds. This result was not as good as had been expected, but I believe it could be greatly improved if the system were to be run on a high-performance server.

3.8 Conclusions and Future Directions

Relatively large-scale test collections for CLIR experiments are available for evaluation of different retrieval approaches. However, few Web-based systems for online cross-lingual information retrieval are available. In this Chapter, I have presented my experiences using an English-Chinese multilingual Web retrieval system in the business IT domain that combines our knowledge of Web retrieval, system building, and CLIR techniques to address the need for multilingual Web retrieval. An experiment was

conducted to measure the effectiveness and efficiency of my Web portal following TREC evaluation procedures. My results showed that my system's phrasal translation and co-occurrence disambiguation led to great improvement in performance, while query expansion techniques did not improve results further. The Web portal was reasonably efficient on a PC and should achieve better efficiency on a more powerful machine. In sum, my study demonstrated the feasibility of applying CLIR techniques to Web applications and the experimental results are encouraging.

I plan to expand my research in several directions. First, I plan to integrate more CLIR techniques into the Web portal to make it more robust. My techniques in ECBizPort is my first step to study Web-based CLIR. I am also investigating how to improve the speed of the system to achieve faster response time, which is necessary for a Web portal. In addition, I plan to expand the Web portal to more languages, such as Spanish and Arabic. Such expansion will allow us to study whether CLIR techniques will perform differently for a multilingual Web portal when more than two languages are involved. Because I believe that different domains might have different effects on the performance of CLIR techniques, I am interested in testing my approach in other domains, such as medicine. Lastly, I plan to extend my system evaluation and user study. On the system performance aspect, a more systematic comparison on the performance of Web-based CLIR systems versus a traditional CLIR system is undergoing. User study on usability and information accessibility of interactive Web-based systems will be conducted in the future.

CHAPTER 4

DEVELOPING A MULTILINGUAL WEB RETRIEVAL SYSTEM: EXPERIMENTS ACROSS WESTERN AND EASTERN LANGUAGES

4.1 Introduction

In Chapter 3, I present an English-Chinese cross-lingual Web retrieval system. In this Chapter, I extend it to a multilingual Web retrieval system which involves five languages: English, Chinese, Japanese, German and Spanish. Multilingual Information Retrieval (MLIR) is the study of responding to a query by searching for documents in other languages (Hull and Grefenstette, 1996). MLIR is an extension of Bilingual Information Retrieval (BLIR), where target documents are in a single language that is different from the query (Chen et al., 2002). Both multilingual and bilingual retrieval are sometimes referred to in a broad sense as Cross-lingual Information Retrieval (CLIR). While bilingual information retrieval has been studied more extensively in different-language pairs such as English-Spanish, English-Chinese, and English-Arabic, multilingual information retrieval has not been widely considered. While much MLIR research has focused either on European or Asian languages, retrieval between European and Asian languages is made more challenging by different linguistic structures. Such an East-West language retrieval issue has not been well explored.

In this research I aim at exploring research and system issues of relevance to multilingual Web retrieval. The Chapter is structured as follows. Section 4.2 reviews related research, including fundamental approaches to MLIR: addressing translation ambiguity and linguistic resource problems and issues related to designing a Web-based MLIR system. In Section 4.3 I discuss problems that arise from using existing MLIR techniques in Web applications and present my research questions. In Section 4.4 I propose the Web-based multilingual retrieval system design. Section 4.5 discusses the system architecture and implementation details of a prototype Multilingual Web Portal in the business domain. Section 4.6 reports the setup and results of an experiment designed to evaluate the performance of the prototype. In Section 4.7 I discuss the findings of my experiments and finally, in Section 4.8, I conclude my work and suggest some future directions.

4.2 Literature Review: MLIR on the Web

Similar to what I reviewed in Chapter 3, in multilingual information retrieval, most MLIR research is based on free text retrieval systems, in which three main approaches were adopted: using a machine translation (MT) system, using a parallel corpus, or using a bilingual dictionary. All three approaches relies on some kind of linguistic resources. In this section I review resources that support MLIR and Web-based MLIR systems.

4.2.1 Resources to Support MLIR

4.2.1.1 Availability of Resources

Query translation and translation disambiguation often require extensive machine translation or linguistic resources. Automatic machine translation systems are well developed between English and the world's major languages, such as Chinese, French, German, Italian, Japanese, Portuguese, and Spanish. However, such systems between other pairs of languages are rare (Larson et al., 2002).

Of the linguistic resources, bilingual dictionaries between major languages are more prevalent than parallel texts in a general domain. However, even relatively widely available bilingual dictionaries exist only for certain language pairs (in most cases between English and another language). Very often, the available dictionaries have different vocabulary coverage for different language pairs, which significantly affects translation quality (McNamee & Mayfield, 2002).

4.2.1.2 Addressing Issues of Scarce Resources

Several efforts have been made to investigate the scarce resources problem in MLIR.

Obtaining Resources from the Web. The Web is becoming the largest data repository in the world. In a new trend arising in natural language processing, some breakthroughs

have resulted from effectively using Web data for linguistic purposes (Brill et al, 2001). Recently several research groups have looked at the WWW as a source for multilingual corpora. Lu et al. (2004) experimented with anchor-text-based Web mining to extract equivalent word pairs in different languages. Nagata et al. (2001) collected a partial parallel corpus between English and Japanese on the web to create a translation dictionary. Other researchers also have taken various Web mining approach and machine learning algorithms to construct parallel corpus (Young et al., 2001; Nie et al., 1999; Yang and Li, 2003; Resnik and Smith, 2003). Although the Web has become a promising resource for MLIR research, the diversity of Web pages means significant work is necessary to construct reasonable MLIR resources from Web collections.

Combining Available Resources. Previous research has shown that by combining multiple resources, MLIR achieves higher precision than that of any single resource. Chen et al. (2000) used two large dictionaries and the Yahoo Chinese search engine to translate Chinese queries into English and achieved great improvement in query translation with the combined resources. Kwok (1999) found that using a bilingual wordlist and found the bilingual wordlist complemented the machine translation. In TREC 2001, the BBN group achieved considerable improvement when combining three Arabic-English bilingual wordlists (one derived from a parallel corpus) (Fraser et al., 2001). Qu et al. (2005) combined various resources for their Japanese-English retrieval, including proper name translations extracted from parallel corpus, bilingual dictionaries

and hand coded person names. Different MLIR resources often complement each other and could improve MLIR system performance.

4.2.2 MLIR for Web Applications

4.2.2.1 Features of Web-based MLIR Systems

As discussed earlier, much research on the technologies that improve retrieval precision has used standard TREC collections, rather than on real-world, interactive Web retrieval applications. While traditional MLIR techniques are promising, they cannot be adopted directly in Web applications. Web-based MLIR differs from traditional MLIR in the following aspects.

Collection building: Traditional MLIR systems are often tested on standard, readily available collections (mostly news articles). Web-based MLIR requires an extensive crawling (spidering) process to build multilingual collections.

Collection size: Traditional MLIR systems are often tested on smaller collections (usually less than 1 GB of data), while Web-based MLIR usually deals with larger collections (more than several GBs). Taking the document collection in TREC 2002 as an example, the collection for the Cross-language Track is 869 MBs, but the Web Track in TREC contains 18.1 GBs of data.

Text format: Traditional MLIR uses standard collections, where all the documents are tagged in structured data format. Web-based MLIR needs to deal with different document formats of documents, including HTML, ASP, PDF, PS, Word, etc.

Efficiency: Traditional MLIR usually focuses on effectiveness, but efficiency is often ignored. However, efficiency is important for end users in Web retrieval scenarios.

Query length: Traditional MLIR often uses long query texts, a sentence description, or sometimes a narrative paragraph. Queries on the Internet are much shorter and have an average length of 2.21 words (Spink & Xu, 2000). Short queries offer less context information for translation disambiguation and thus are more challenging in MLIR research.

4.3 Research Questions

The Web has become a major information source for people worldwide in any knowledge field. The use of MLIR techniques in Web retrieval is expected to address the multilingual information needs of Web users; however, the diverse and fast-growing nature of the Web makes some techniques unsuitable. Traditional MLIR techniques have not been widely used and evaluated in Web applications (Oard, 2002). Although a few systems exist, most provide only simple translation functions and do not provide comprehensive evaluation results. Most research has been based on document collections

carefully selected for evaluation purposes. There is a need to study the performance of MLIR in Web-based applications.

Each of the three MLIR translation approaches has some drawbacks and the performance of each approach is limited by the availability of natural language resources. As the largest knowledge repository, the Web offers extensive resources that could be used in MLIR. Some researchers have experimented with Web mining approaches to compensate for limited existing resources. For non-English language pairs, pivot languages translation shows potential to achieve reasonable performance, but it has not been widely studied yet. MLIR research that involves more than two languages often has studied either European or Asian languages, but not both.

Based on my review, I believe MLIR techniques offer a promising solution to the problems of practical multilingual Web retrieval systems and Web portals, especially when query translations are combined with various translation disambiguation techniques. In this study, I posed the following research questions:

1. How can I develop a generic approach for a multilingual Web retrieval system that incorporates both European and Asian languages?
2. How can I mine the Web for useful linguistic resources to improve multilingual Web retrieval performance?

3. Can a multilingual Web retrieval system achieve satisfactory performance (in effectiveness and efficiency) in the Web context?

4.4 A Proposed Multilingual Web Retrieval (MWR) Approach

Aiming to apply an integrated set of MLIR techniques in the Web environment, I propose a framework for multilingual Web portal development. The proposed system architecture is shown in Figure 4.1.

My MWR system architecture consists of three major components: (1) Multilingual Collection Building, (2) Query Translation, and (3) Document Retrieval. In the following, I describe each component in detail.

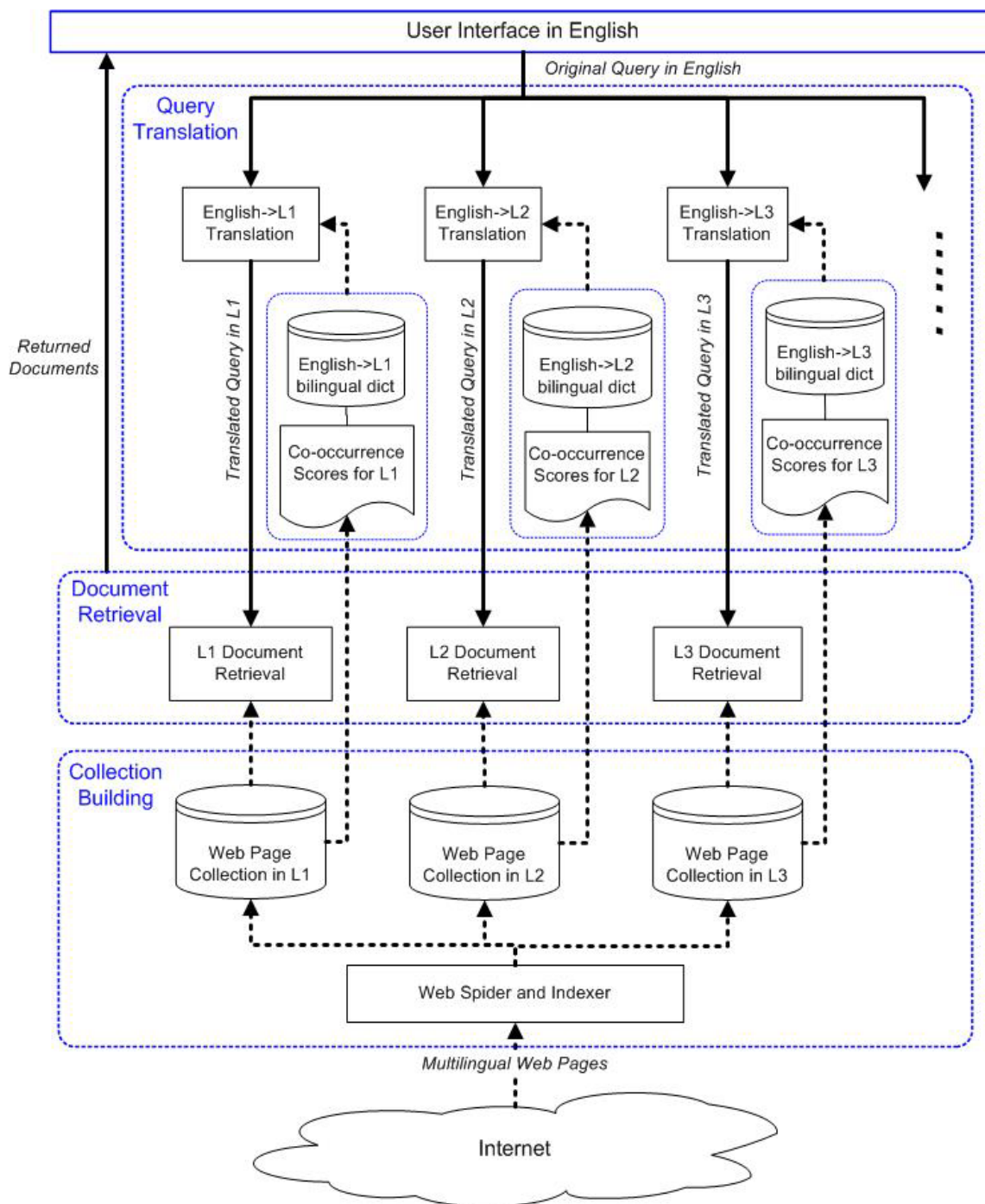


Figure 4.1: Proposed architecture for a multilingual Web retrieval system (the MWR system)

4.4.1 Multilingual Collection Building

The multilingual collection building process has rarely been addressed in traditional MLIR research. However, it is a crucial component if a Web-based MLIR system is to leverage the increasing availability of information in different languages on the Web. The collection building process consists of two parts, the Web spider and indexer.

4.4.1.1 Web Spider

Web spiders, or Web crawlers, are programs that retrieve pages from the Web by recursively following URL links in pages using standard HTTP protocols (Cheong, 1996). The Web Spider component is responsible for document collection building. Document collections in two or more languages are needed for a multilingual Web portal, not only as the information resources provided by the portals to their users but also as a comparable corpus that can be used for translation disambiguation and query expansion. The second purpose is especially important because it is difficult to obtain such parallel or comparable corpora for particular domains.

There are high quality requirements for the collections. Every Web page in the collection has to be highly relevant to the selected domain and, at the same time, they have to be diverse enough to cover multiple topics and interests in the domain. Focused Web spiders — spiders that collect pages in specific domains or Web sites — can be used to build domain-specific Web collections (Chau & Chen, 2003). Although traditional focused

Web crawling methods can be used to create Web page collections that meet the relevance requirement, the scope of collections built by these methods is usually restricted to the topics to which the starting URLs relate (Bergmark et al., 2002). As a result, simple focused crawlers often fail to provide comprehensive coverage of the different topics within a domain. To address this problem, I have proposed a collection-building method that extends traditional focused crawling by meta-searching multiple large search engines. My method can be described as follows. Similarly to traditional focused crawlers, I start my “probing” of the Web with a set of starting URLs and fetch relevant Web pages. At the same time, new starting URLs are being identified by querying multiple search engines (e.g., Google, Yahoo, AltaVista, etc.) and combining their top results, thereby combining the top results from multiple search engines (Lawrence & Giles, 1998) while maintaining the relevance of the collection. Documents in different languages are collected separately to allow language specific processing in other components.

4.4.1.2 Indexer

Web pages must be indexed differently from traditional text documents. Because documents from the Web can be in various formats, such as HTML, ASP, JSP, PDF, or MSWord, Web-specific indexers are designed to work with a specific Web page structure (e.g., removing markup tags from HTML documents).

Encoding is another problem to be considered when indexing multilingual documents. Hundreds of different encoding systems represent diverse languages. Different encoding systems even exist for the same language. For example, EUC-JP and Shift-JIS are both used in Japanese Web pages; Big5 and GB2312 are both used in Chinese Web pages. These encoding systems often conflict. The indexer needs to be informed of the Web page encoding in order to perform indexing accordingly.

For languages with a rich morphology, stemming is frequently used in indexing to improve retrieval performance. The Porter stemmer is a standard tool for English which achieves normalization by chopping off suffixes (Porter, 1980). Some languages, such as German, have more complex morphologies and a simple stemming algorithm does not achieve successful performance. In these cases a more comprehensive algorithm, such as language decompounding, is needed (Chen et al., 2002).

In general, the proposed multilingual indexer needs to deal with different Web page formats, different encoding systems, and language specific stemming algorithms. Indexes of Web pages serve two purposes: 1) to generate a co-occurrence analysis matrix for translation disambiguation; 2) to generate $tf*idf$ scores for document retrieval.

4.4.2 Query Translation

The Query Translation component is the core of the system. It is responsible for translating search queries in the source language into each of the target languages. Among the three translation approaches, the dictionary-based seems to be most promising for Web applications for two reasons. First, compared with the parallel corpora required by the corpus-based approach, MRDs used in dictionary-based MLIR are much more widely available and easier to use. The limited availability of high-quality parallel corpora makes the approach problematic for multilingual Web retrieval. Second, compared with MT-based MLIR, the dictionary-based MLIR approach is more flexible, easier to develop and easy to adopt. It is often impractical to build a complex MT system just for MLIR. Existing commercial MT software is either packaged as a black box, leaving little space for users to modify it for their specific purposes, or it is available only for certain language pairs. According to a previous study (Gao et al., 2001), a dictionary-based MLIR with a combination of corpus-based disambiguation techniques can achieve better performance than high-quality MT systems. In my research, I propose a dictionary-based approach that combined with phrasal translation and corpus-based co-occurrence analysis for translation disambiguation. Figure 4.2 illustrates the four major steps in query translation: 1) Dictionary Lookup, 2) Maximum Phrase Matching, 3) Phrasal Translation, and 4) Co-occurrence Analysis. An example of how an English query “Security Intelligence Technology” is translated into a Chinese query “安全 情报技术” is given on the right-hand side of Figure 4.2.

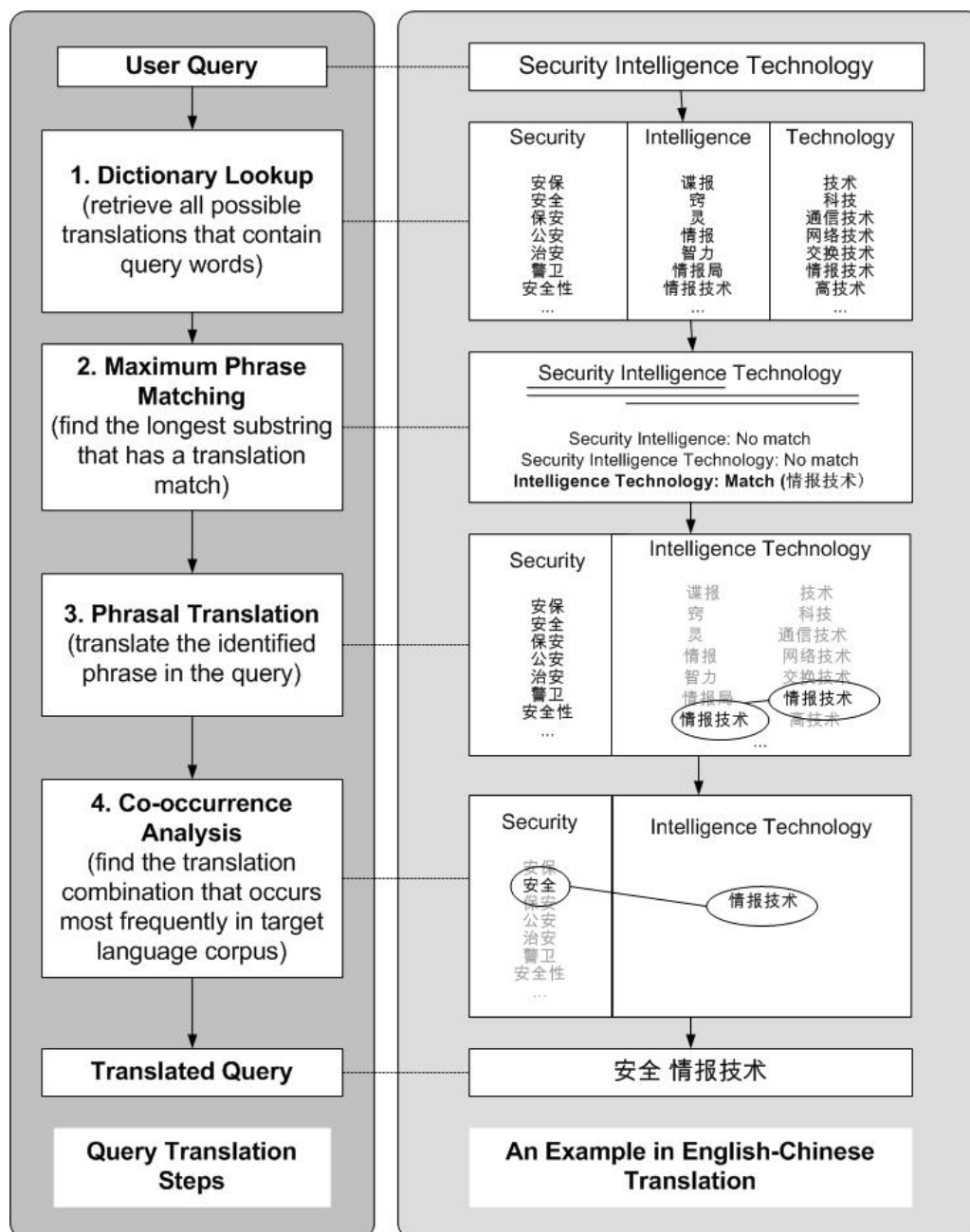


Figure 4.2 Query translation steps with an example in English-Chinese translation

4.4.2.1 Dictionary Lookup

I first prepare the dictionary in the [query language -> target language] format. The dictionary is indexed such that information about the number and type (word or phrase) of translations is recorded for each word. In the dictionary lookup process, I retrieve all possible translations of each query word. Note that this includes both word-to-word translations and phrase translations that are associated with the original query. Since each word is often associated with 5 or even 10 different translations, the next step focuses on translation disambiguation. The example in Figure 5.2 shows that each of the individual query words generates more than 7 translations, which brings a high level of ambiguity.

4.4.2.2 Maximum Phrase Matching and Phrasal Translation

Maximum phrase matching is the process of identifying the longest substrings in the query that can be translated into the target language as a whole. In this process, all query words are parsed from left to right and possible substrings of different length are generated. The program identifies the substrings with all adjacent query words sharing the same translation. These substrings are then treated as a phrase, and phrasal translation is performed. The remaining words still keep their word-by-word translation from the first step. The example in Figure 2 illustrates that maximum phrase matching identified “intelligence technology” as a phrase translation from the original query, and that phrase is translated as a whole into “情报技术” rather than the two separate translations. Phrasal

translation is an effective way to eliminate translation ambiguity. However, its performance relies heavily on the completeness and quality of the dictionary.

4.4.2.3 Corpus-based Co-occurrence Analysis

Existing MLIR approaches for query translation often require a target-language corpus for the disambiguation of translated query terms. However, such domain-specific resources are not always available. I propose to retrieve Web documents as a corpus and use co-occurrence information between terms within that corpus. The corpus extracted from the Web could provide a good basis for translation disambiguation.

Co-occurrence analysis is used to help choose the best translation among candidates. A co-occurrence matrix was calculated from the extracted corpus beforehand. The indexer program described in Section 4.1.2 first extracts all terms from the corpus. For each pair of terms $\{Q_i, Q_j\}$, all possible translation pairs $\{T_m, T_n\}$ in the bilingual dictionary are extracted. The co-occurrence score between two translations T_m and T_n then can be calculated as follows:

$$Co-occur(T_m, T_n) = \frac{N(T_m, T_n)}{N(T_m) + N(T_n) - N(T_m, T_n)}$$

where $N(T_m, T_n)$ is the number of documents that contain both T_m and T_n , and $N(T_m)$, $N(T_n)$ are the numbers of documents contain T_n or T_m in the query. My method is similar to that of (Maeda et al., 2000), in which they sent translation pairs to other search engines and used the number of returned documents to calculate co-occurrence scores. What differentiates my proposed method from theirs is that they calculated a co-occurrence score “on the fly,” which may greatly lower system efficiency. I calculated co-occurrence scores in advance to avoid affecting run-time efficiency, which is extremely important for Web applications.

In the last step, I retrieve the highest-scored translation for all possible translation combinations and select the final translation. When there are more than two words in the query, co-occurrences analysis is performed on all possible word pairs. The combination that achieves highest summed score is then selected. In the example in Figure 2, the second translation for “security,” the second translation “安全” achieved the highest score with the phrase translation result of “intelligence technology.” Thus, this translation of “安全” is selected over other translations.

4.4.3 Document Retrieval

The Document Retrieval component uses the translated queries in each target language and retrieves documents from each collection. This component is similar to the retrieval component in monolingual information retrieval systems. Different ranking methods,

such as frequency-based ranking or PageRank, can also be incorporated in this component (Arasu et al., 2001). The retrieved results are grouped by languages when presented to the end user. I do not investigate result merging strategies in this Chapter .

4.5 A Multilingual Web Portal for Business Intelligence: An Experiment

In this section, I report my experience in implementing a multilingual Web retrieval system using the proposed MLIR approach. The prototype system five major representatives of Western and Eastern languages: English, Chinese, Japanese, German and Spanish, which are the five most popular languages on the Web. I chose international IT as the testbed because of its emerging roles and globalization trend.

4.5.1 Multilingual Collection Building

The AI Lab SpidersRUs toolkit (<http://ai.bpa.arizona.edu/spidersrus/>), a digital library development tool developed by my research group, was used to build multilingual collections for the Web portal. The toolkit contains components that support document fetching, document indexing, collection repository management, and document retrieval. It can also build collections in different languages and encodings.

4.5.1.1 Spidering

As mentioned earlier, to ensure the relevance and coverage of the Multilingual Business Intelligence Portal collections, a meta-search-enhanced focused spidering method was used. For each collection, a list of business-related starting URLs and a list of typical business-related queries were selected by domain experts. During the spidering process, pages were fetched from the Web by recursively following URL links. At the same time, the queries identified by the experts were sent to four major search engines, Google (<http://www.google.com/>), Yahoo! (<http://www.yahoo.com/>), AltaVista (<http://www.altavista.com/>), and HotBot (<http://www.hotbot.com/>). These four search engines were chosen for their ability to search documents in the chosen languages. The spider program was set to stop after collecting 100,000 pages to make collections comparable in size. Running on a Pentium-4 PC, the spiders spent about 6-10 hours collecting 100,000 IT/business-related Web pages for each language. Table 4.1 illustrates my spidering details.

Table 4.1: Spider program settings and number of pages collected for each language

Language	# of Starting URLs	# of Business Keywords	# of Web pages Spidered
English	31	242	100,000
Chinese	40	135	100,000
Japanese	33	164	100,000
Spanish	20	115	100,000
German	33	127	100,000

4.5.1.2 Indexing and Stemming

Indexing

My collections were indexed in two ways: first by employing a character-based/word-based index, and then using dictionary translations as indexing terms.

To support document retrieval, English, Spanish, and German Web pages in the Multilingual Business Intelligence Portal were indexed using a word-based indexing approach; Chinese and Japanese Web pages were indexed using a character-based indexing approach. In both approaches, positional information on the words or characters within a Web page was captured and stored such that when the query was a phrase, Web pages containing the exact phrase could be retrieved and given higher ranking than pages with separated words.

Using word-based indexing and character-based indexing during my general indexing process avoided information loss, but did not capture phrases in any of the languages. This led to loss of semantic meaning, especially in Chinese and Japanese, since in most cases in Chinese and Japanese a meaningful term contains more than one character. These semantic meanings are useful in translation disambiguation, where the system needs to pick the right translation phrase among several candidates. Therefore, I indexed all the pages against their analogous dictionaries. The indexed terms are potential translations from bilingual dictionaries and would be used in co-occurrence calculation for translation disambiguation purposes.

Stemming

Word normalization will lead to much greater improvement in retrieval effectiveness for morphologically rich and lexically complex languages. The indexing procedure uses stemming algorithms for English, Spanish, German, and Arabic.

As a standard, the Porter stemmer is used for the English collection (Porter, 1980). For Spanish, I implemented the Snowball stemming algorithm, a description of which is available at <http://snowball.tartarus.org/spanish/stemmer.html>.

In German, compound words are widely used and this causes more difficulties than English compound words. For example, the word *Entwicklungsumgebung* (development environment) is derived from *Entwicklung* (development) and *Umgebung* (environment). According to Chen (2002), including both compounds and their composite parts during indexing would improve the performance. I took a completely dictionary-based approach to German word normalization. If a word was not found in the dictionary, I would then search for substrings of the word to see if I could find a match for the word through matching a series of substrings.

In Chinese and Japanese, noun phrases do not have morphological variations, so no stemming algorithm was applied to these two languages.

4.5.2 Query Translation

Query term translations were performed using bilingual dictionaries. Table 4.2 summarizes the dictionaries I used for each language pair.

Table 4.2: Bilingual dictionaries used in query translation

Language Pair	Bilingual Dictionary Used	Provider	# of Entries in Dictionary
Chinese/ English	LDC Wordlist	LDC	120,000
Japanese/English	EDICT	Monash University	106,012
English/Spanish	EFN Wordlist	EFN Organization (http://www.efn.org/)	25,535
English/German	TravLang Dictionary	TravLang	18,554

As I proposed in Section 4.4.2.1, each entry in these dictionaries was first indexed. The relationships between the English term and the Chinese translations were captured and recorded. For further disambiguation purposes some other important information, such as the number of English terms found in one dictionary entry and the positions of the term located in the entry, also was captured and stored. Given the indexed dictionary, definitions of English terms could be quickly and easily retrieved. Maximum phrase matching was also incorporated in my system by ranking Chinese translations containing more continuous key words higher than those containing discontinuous key words. Word co-occurrence information trained from a target language text collection was used to disambiguate the translations of query terms. Co-occurrence analysis was implemented by extracting all the terms that appeared in corresponding dictionaries from the documents in the Multilingual Portal collections. The co-occurrence scores were

calculated in a batch process based on the procedure described in Section 4.4.2.3 and stored in a database.

4.5.3 Document Retrieval

The document retrieval component was performed as in monolingual retrieval. It was supported by the AI Lab SpidersRUs toolkit. After a target query had been built, it was passed to the toolkit's search module, which searched the document indexes and looked up the documents that were most relevant to the query. The retrieved documents then were ranked by their $tf*idf$ scores and returned to the user through the Web-based interface.

4.5.4 Sample User Sessions

Figure 4.3 shows sample user sessions on the Multilingual Business Intelligence Web Portal prototype. The interface is in UTF-8 encoding. On the search page (Figure 4.3.a), users can choose among five languages as their query language by checking the boxes. They can input multiple keywords in the search box above. The left-side of the search page provide five target languages. Under each language, users can choose among three query translation approaches: word-by-word translation, phrasal translation, a combined phrasal and co-occurrence analysis approach.

The query is passed to the system for query translation. A set of documents is retrieved by the system. On the result pages (Figure 4.3.b-4.3.d), I display top 20 returned results from each target language. Results are organized according to their languages. The title and a one-sentence summary of each document are displayed. Users can browse the set of results from a particular language. The translated queries are also displayed to the user so he/she may use the terms to refine the query manually.

a. Search Interface

Address: http://ai25.bpa.arizona.edu:8080/newMLWR/index.html

Multilingual Web Retrieval

Multilingual Web Retrieval Experiment System

Search Terms:
database management system

Choose Source Language:
 English
 Chinese
 Spanish
 German
 Japanese
 Arabic

Choose Target Languages:
 English
 German
 Chinese
 Spanish
 Japanese

Word-by-Word
 Phrasal Translation
 Phrasal with Co-occurrence

Search Reset

b. Chinese Results

English Query: IT information service
 Chinese Translation: 高性能服务器

- 戴尔公司 (Dell Inc.) - 笔记本电脑, 台式机, 工作站, 服务器和存储
<http://www1.ap.dell.com/content/default.aspx?c=en&l=zh&s=gen>
- 戴尔惠普竞争白热化 昔日盟友变成今日对手 - IT - 21CN.COM
<http://it.21cn.com/itnews/news/2002-10-21/804876.htm>
- 清华大学继续教育学院办公案例分析 - IT - 21CN.COM
<http://it.21cn.com/prnews/2002-11-22/943604.htm>

c. Japanese Results

English Query: database management systems
 Japanese Translation: データベース | 管理 | システム

- データベース導入を促す日本IBM専業基研研究所 活動内容公開 - CNET Japan
<http://www.japan.cnet.com/news/ent/story/0,2000047623,20060333,00.htm>
- Linux技術者1000人を育成 - 日本オタク丸。全国で技術セミナーを開催 - CNET Japan
<http://www.japan.cnet.com/news/ent/story/0,2000017623,20060219,00.htm>
- サイボウズ ガルーンとNECソフトウェア九州の電子法システムが連携動作 - CNET Japan
<http://www.japan.cnet.com/news/ent/story/0,2000047623,20060275,00.htm>
- 顧客ニーズの変化にオープンソースに変わるデータベース業界 - CNET Japan
<http://www.japan.cnet.com/news/special/story/0,2000047679,20064489,00.htm>
- マイクロソフト、インテルと対簿強ひに主簿 | 世界ITビジネス争奪戦の光と影 - CNET Japan
<http://www.japan.cnet.com/news/loop/story/0,2000047670,20052882,00.htm>
- NTT、長崎、東横、ケーブル.....「地デジ」で放送に切り替えるか - CNET Japan
<http://www.japan.cnet.com/news/loop/story/0,2000047670,20054584,00.htm>

d. Spanish Results

English Query: database management systems
 Spanish Translation: libre | comercio | acuerdo

- Acuerdos de Libre Comercio - Acuerdo de Libre Comercio de las Américas - ALCA
http://www.nic.gov.co/Marco_internacional/Acuerdos_libre_comercio_alca.htm
- Acuerdos de Complementación Económica, Libre Comercio e Intercambio Preferencial y Alcance Parcial
http://www.nic.gov.co/Marco_internacional/Acuerdos_complementacion_economica.htm
- ALCA - FTAA - ZLEA - Sitio Oficial del Proceso del Área de Libre Comercio de las Américas (ALCA)
http://www.ftaa-alca.org/alca_e.asp
- RED MEXICANA DE ACCION FRENTE AL LIBRE COMERCIO
<http://www.rmalc.org/mal/>
- EE UU y América Central negocian acuerdo libre comercio
<http://usembassy.state.gov/colombia/www/a03.shtml>
- Acuerdo de Libre Comercio entre EE.UU. y Centroamérica
<http://usembassy.or.cr/Ca/la/alca.html>
- SICE-Panamá-Quetzalteno - Acuerdo de Libre comercio e intercambio preferencial
<http://www.nice.oas.org/tra-del/efredra/delpangas.asp>

Figure 4.3: User interface of Multilingual Business Intelligence Portal

4.6 System Evaluation

I designed and conducted an experiment to evaluate the performance of my multilingual system. In this section, I discuss MLIR evaluation methodologies, experiment design and measures, and my hypotheses.

4.6.1 MLIR Evaluation Methodology

MLIR evaluation aims at testing the effectiveness of retrieval systems, often measured by precision and recall. To make effectiveness comparable across systems, such tests usually are carried out on a common data set.

Three major evaluation workshops that have provided test collections for CLIR experiments are: the Cross-Language Evaluation Forum (CLEF) covering many European languages, the NTCIR Asian Language Evaluation covering Chinese, Japanese, and Korean, and the TREC Cross Language Track from 1997-2002. In all these workshops, the task was to match topics in one language against documents in another language and return a ranked list (Gey & Chen, 2000; Chen et al., 2002; Peters, 2002). A set of documents from the subject collection was pre-judged by human experts to be relevant to the original query. Since it would have been unrealistic to expect every document in the collection to be judged, precision often was considered of more interest than recall and usually was reported at low recall levels. In other words, precision rate

often was reported for the top n retrieved documents (n usually being between 5 and 1000). Once relevance judgments had been established, precision could be computed upon the ranked list of each entry.

Since MLIR always yields precision loss compared with traditional monolingual information retrieval, researchers are often interested in how well MLIR performs in comparison with a corresponding monolingual run. In a monolingual run, original queries were manually translated into the target query, and retrieval was performed on this human-translated query. When precisions for both MLIR and monolingual retrieval had been obtained, the precisions of MLIR and of the monolingual retrieval could be compared.

Besides multilingual and monolingual comparison, comparisons are often made among different translation approaches. The average precision for simple MLIR runs, such as word-by-word query translation without using any translation disambiguation or query expansion, are often compared with runs under other conditions to show the superiority of different techniques (Ballesteros & Croft, 1996; Oard & Wang, 2001).

4.6.2 Experiment Design and Measures

In general, I followed the TREC evaluation process in my experiment design (Voorhees, 1998). However, because my study used Web pages instead of standard collections, no

established relevance judgment was available for precision and recall. I decided to create relevance judgment by recruiting human experts. Since I was particularly interested in how well these techniques would work for Web content in a business intelligence portal, I recruited five bilingual business school students as domain experts, all fluent in English and one of the target languages (Chinese, Japanese, Spanish, and German). They identified 10 English-based queries of interest in the business/IT domain and translated these queries into the target language as a standard translation. These English queries contained 2-4 words (2.4 words on average) and resembled queries often submitted by an end-user of a Web search engine in terms of length. The human-translated queries were used to retrieve *monolingual runs*. As discussed, such monolingual retrieval represents the best-case scenario performance of multilingual information retrieval, where all the words are perfectly translated.

To obtain *multilingual runs*, the original English queries were sent to my multilingual system based on four settings: 1) word-by-word translation (WBW), 2) phrasal translation, 3) co-occurrence analysis translation, and 4) phrasal translation with co-occurrence analysis (Ph-Co). Word-by-word translation picked the first translation in the dictionary and ignored all the other translation candidates, which was considered as a worst-case benchmark for the multilingual runs.

During the experiment, the experts individually submitted each query to the system under the five different settings (one for monolingual and four for multilingual). The results of

the target retrieval were compared with the two benchmark settings: (1) monolingual information retrieval (the best-case scenario), and (2) word-by-word translation (the worst-case scenario). With ten queries and five different settings, each expert performed a total of 50 searches using the system. Each expert went through the top 10 Web pages returned for each query and gave each page a score of 0 (irrelevant) or 1.0 (relevant to the search). To compare the effectiveness of the system, I used precision only for the top 10 retrieved documents for each query and setting, a measurement referred to as target retrieval in the NTCIR workshop (Eguchi et al., 2002). Precision measures how well the system can find relevant results for a user's query. It is calculated as

$$Precision = \frac{\text{number of relevant documents retrieved by the system}}{\text{number of all documents retrieved by the system}}$$

Efficiency refers to the amount of time the system took to retrieve documents. During the experiment it was recorded as the time needed, since users hit the search button until the results were returned.

4.6.3 Hypotheses

I tested three groups of hypotheses in my multilingual Web retrieval system.

H1: Web-based multilingual information retrieval could achieve a performance level reasonably comparable to that of monolingual retrieval.

H2.1: Co-occurrence analysis trained from Web documents could significantly improve multilingual retrieval effectiveness over word-by-word translation.

H2.2: Phrasal translation could significantly improve multilingual retrieval effectiveness over word-by-word translation.

H3.1: When phrasal translation and co-occurrence analysis are applied together, effectiveness is further improved over using co-occurrence analysis translation alone.

H3.2: When phrasal translation and co-occurrence analysis are applied together, effectiveness is further improved over using phrasal translation alone.

In Hypothesis 1, I believed the overall performance of the Web-based multilingual retrieval system could achieve $2/3$ of monolingual retrieval performance, which is considered to be reasonable performance in multilingual information retrieval (Grefenstette, 1998).

For Hypotheses 2.1 and 2.2, I compared the extent to which each technique could push performance in multilingual Web retrieval.

For Hypothesis 3.1 and 3.2, I believed that by applying both co-occurrence analysis and phrasal translation, the ambiguity created by simple translation is reduced and the effectiveness could be further improved.

I did not emphasize system efficiency in my hypotheses because it is obvious that monolingual retrieval was most efficient, while applying both co-occurrence analysis and phrasal translation was the least efficient. However, I provided experiment results on efficiency measures in the next section to study whether these techniques are practical in real world applications.

4.7 Experiment Results and Discussions

In this section, I describe and analyze the results of my study comparing five experimental settings in four different language pairs. Table 4.3 summarizes the average precision of each experiment setting for each language pair. The language pairs are shown in the first column and the different settings and methods are shown in the second column of the table. The third column shows the top-10 precision rates of each method averaged across the 10 searches performed by each of the 5 experts. The fourth column shows a method's performance compared with that of the monolingual retrieval. The fifth column shows how much each method's precision was improved when compared with WBW translation. Table 4.4 shows the p -value of various t -tests in testing the hypotheses.

Hypotheses tested are listed in the first column and the next four columns are p -values of t-tests for four different language translation pairs.

Table 4.3: Summary of system effectiveness performance

	Experiment Settings	Average Precision	Performance compared with Monolingual	Improvement compared with WBW
English-Chinese	Monolingual	0.68	100%	
	WBW	0.13	19.12%	-
	Co-occurrence	0.41	60.29%	215%
	Phrasal	0.39	57.35%	200%
	Co+Phr	0.52	76.47%	300%
English-Japanese	Monolingual	0.54	100%	
	WBW	0.04	7.41%	-
	Co-occurrence	0.37	68.52%	825%
	Phrasal	0.33	61.11%	725%
	Co+Phr	0.46	85.18%	1005%
English-Spanish	Monolingual	0.58	100%	
	WBW	0.31	53.45%	-
	Co-occurrence	0.36	62.07%	16.13%
	Phrasal	0.34	58.62%	9.67%
	Co+Phr	0.36	62.07%	16.13%
English-German	Monolingual	0.75	100%	
	WBW	0.07	10.29%	-
	Co-occurrence	0.53	70.59%	586%
	Phrasal	0.07	10.29%	0%
	Co+Phr	0.53	70.59%	586%

WBW: Word-by-word translation

Co-occurrence: Co-occurrence analysis-based translation

Phrasal: Phrasal-based translation

Co+Phr: Combined co-occurrence analysis and phrasal-based translation

Bold figures indicates the best performance achieved during multilingual retrieval

Table 4.4: Summary of system effectiveness performance

Hypotheses	English-Chinese	English-Japanese	English-Spanish	English-German
H1: Multilingual > 2/3 Monolingual	<0.001*	<0.001*	0.060	0.001*
H2.1: Coo > WBW	<0.001*	0.002*	0.060	0.001*
H2.2: Phr > WBW	0.008*	0.020*	0.172	1.000
H3.1: Co-Phr > Coo	0.030*	0.060	1.000	1.000
H3.2: Co-Phr > Phr	0.006*	0.030*	0.080	0.001*

An asterisk (*) means that t-test results was significant at a 95 percent confidence level.
Sample size: 50

4.7.1 Overall Comparison Between Multilingual and Monolingual

Table 4.5 compares the precision achieved in monolingual retrieval and the best multilingual retrieval performance in my experiments. On average, multilingual performance achieved 73.57% of monolingual performance. Monolingual precision level for the four languages varied between 0.54 (Japanese) and 0.75 (German). Because I used a generic word-based (character-based for Japanese and Chinese) index and a generic tf*idf ranking schema for all languages, this variation range is considered to be reasonable. Japanese yielded the lowest precision in monolingual retrieval, and I believe it is because of its morphological complexity and lack of standard orthography (Halpern, 2003). The precision can be improved if these problems are addressed. The best monolingual performance was a 0.75 precision of German. I believe that the decomposing stemming algorithm I used for the German language contributed to the precision performance.

When comparing multilingual performance with monolingual performance, the best multilingual retrieval performed from 62.07% (Spanish) to 85.18% (Japanese) of the corresponding monolingual precision. *HI* was supported for *English-Chinese, English-Japanese, and English-German multilingual retrieval, but was not supported for English-Spanish retrieval*. Although English-Spanish retrieval was not as good as other language pairs, it is comparable to two-third of the monolingual precision with a p-value of 0.060 (see Table 4.4). In general, I conclude that Web-based multilingual information retrieval could achieve a performance level reasonably comparable to that of monolingual retrieval when appropriate techniques are used.

Table 4.5: Comparison between monolingual and best multilingual performance

Language Pair	Monolingual Precision	Best Multilingual Performance Precision	Performance Compared with Monolingual
English-Chinese	0.68	0.52	76.47%
English-Japanese	0.54	0.46	85.18%
English-Spanish	0.58	0.36	62.07%
English-German	0.75	0.54	70.59%
Average	0.64	0.47	73.43%

4.7.2 Phrasal and Co-occurrence Translation Comparison

4.7.2.1 English-Chinese Language Pair

For the English-Chinese language pair, I found that both phrasal translation and co-occurrence analysis significantly improved performance over word-by-word translation.

By applying phrasal translation alone the performance improved from 19.12% to 57.35%

of monolingual retrieval, and co-occurrence analysis alone improved the performance to 60.29%. Both improvements were statistically significant at the 95% confidence level (Table 4.4). Thus I concluded that *H2.1 and H2.2 were both confirmed* for the English-Chinese language pair.

When phrasal translation and co-occurrence analysis were combined, the performance was boosted to 76.47%, significantly better than using phrasal translation alone or using co-occurrence translation alone. Both *H3.1 and H3.2 were confirmed*.

4.7.2.2 English-Japanese Language Pair

Similar to the English-Chinese pair, *H2.1 and H2.2 were confirmed* for the English-Japanese language pair: The multilingual retrieval resulted in significantly better performances when phrasal translation or co-occurrence analysis translations were used in comparison to word-by-word retrieval. The two methods reached 61.11% and 68.52% of monolingual performance, which greatly exceeded the 7.41% effectiveness obtained from the word-by-word approach.

When looking at the combined approach, H3.2 was confirmed, but H3.1 was not. I concluded that using phrasal and co-occurrence translation together performed significantly better than using phrasal translation alone, but not significantly better than using co-occurrence alone. Compared to other language pairs, a combined approach for

English-Japanese retrieval achieved the highest improvement over word-by-word translation. I believe that phrasal translation and co-occurrence analysis are extremely helpful when morphologically complex or orthographically irregular languages are involved.

4.7.2.3 English-Spanish Language Pair

There was no significant difference between using simple word-by-word translation and phrasal translation. Using co-occurrence analysis alone did not significantly improve the performance; neither did the combination of the two techniques. I observed that the word-by-word translation in this language pair performed quite well: 53.45% of monolingual retrieval while the other three languages achieved less than 20% of monolingual retrieval. I believe that because Spanish is closer to English a word-by-word translation could often pick the correct translation. However, because of the good performance in word-by-word translation, the co-occurrence analysis approach did not significantly improve the performance, despite the fact that I still observed a 16.13% increase when co-occurrence analysis is used. Phrasal translation enhanced the word-by-word translation by only 9.67%.

Using phrasal and co-occurrence analysis together did not significantly improve the effectiveness over using phrasal or co-occurrence analysis alone. *Thus H3.1 and H3.2 were not confirmed for English-Spanish retrieval.* That the phrasal-based approach had

little impact on English-Spanish cross-lingual retrieval is mainly due to the fact that the English-Spanish dictionary contains very few phrase entries. Unlike co-occurrence analysis which can be built on Web collections, phrasal translation is dependent on the coverage of the bilingual dictionary. The English-Spanish dictionary is much smaller than the dictionaries for the two Asian languages, with only one-fourth the number of entries and most of them are word-to-word translations.

4.7.2.4 English-German Language Pair

In my experiments with the English-German language pair, I found that *H2.1 was not confirmed*: phrasal translation did not perform better than word-by-word translation. While phrasal translation did not improve the effectiveness, co-occurrence analysis significantly enhanced the effectiveness from 10.29% to 70.50% of monolingual retrieval. Again, I noticed that the English-German dictionary that I used contains limited phrases, resulting in the failure of phrasal translation. The decompounding stemming approach successfully segmented most German compounds during indexing such that during co-occurrence analysis, the translated German words can be identified from my index of the Web collection.

In testing the combined approach performance I concluded that while *H3.1 was not confirmed*, *H3.2 was supported again*. Because no phrases were identified in the query,

phrasal translation did not have a positive impact on the effectiveness. Therefore a combined approach benefited from co-occurrence analysis but not phrasal translation.

4.7.2.5 Discussion

I observed performance differences between European languages (Spanish and German) and Asian languages (Chinese and Japanese). For the two Asian languages, phrasal translation alone and co-occurrence alone both significantly improved performance, and using both co-occurrence and phrasal translation further improved performance. For the two European languages phrasal translation alone did not significantly improve the performance, while co-occurrence significantly improved German translation but not Spanish translation.

This result could be explained by looking at the different resources used for each language pair. English-Chinese and English-Japanese dictionaries are more comprehensive and contain significantly more phrase information than German and Spanish dictionaries. The English-Chinese (E-C) dictionary contains 120,000 entries and the English-Japanese (E-J) dictionary contains 106,012 entries, compared with 18,554 entries in the English-German (E-G) dictionary and 25,535 entries in the English-Spanish (E-S) dictionary.

In all cases, co-occurrence analysis quite consistently improved translation performance. I found improvement larger than that in traditional MLIR, which could have resulted from the high quality of my Web page collections. In traditional MLIR, general news articles are used as the co-occurrence training set and the query terms and their translations are less sensitive to that general training set. In a domain specific multilingual Web retrieval, the corpus is built to be highly relevant to the domain. This helps co-occurrence analysis assign high scores to translations that are most relevant to the domain. My experiment results showed that in domain-specific multilingual Web retrieval, corpora mined from the Web provide a good training set for co-occurrence analysis. These comparable corpora have potential to replace some linguistic resources that are not widely available and could serve in various corpus-based approaches.

4.7.3 Efficiency

Besides effectiveness, efficiency is another important aspect of Web retrieval. Long system response time (time elapsed between the moment when the search button is clicked and the results' final appearance on the screen) can cause users to lose patience and thus lower user satisfaction. To investigate the effect of MLIR techniques on system efficiency, system response times for performing various MLIR tasks were recorded and compared. System response time depends on factors such as hardware performance and network traffic. Since my experiments were run on PCs, I emphasize the response time of multilingual retrieval in comparison to monolingual retrieval. Thus, I consider

monolingual retrieval as my baseline performance. Table 4.6 summarizes the average time spent under each system setting.

Table 4.6: Efficiency of Multilingual Business Intelligence Portal

	English-Chinese (seconds)	English-Japanese (seconds)	English-Spanish (seconds)	English-German (seconds)	Average (seconds)	Comparison with monolingual
Monolingual	6.87	7.12	4.32	5.05	5.84	1
WBW	8.14	8.87	5.86	6.13	7.25	1.24
Co-occurrence	20.47	23.18	10.89	11.34	16.47	2.82
Phrasal	8.98	9.76	6.55	6.99	8.07	1.38
Co+Phr	22.04	24.26	11.47	12.15	17.48	2.99

My results showed that compared to monolingual retrieval, simple word-by-word translation took 1.24 times as long as monolingual retrieval. When phrasal translation was involved, the retrieval time increased to 2.82 times that of monolingual retrieval. However, phrasal translation did not decrease the efficiency much, at 1.38 times the response time of monolingual retrieval. When phrasal translation and co-occurrence analysis were used together, the performance dropped to 2.99 times the response time of monolingual systems. Co-occurrence analysis was the component which took the most computation and efficiency could drop a lot when it was applied.

I also observed that although Chinese and Japanese performed better than Spanish and German in effectiveness, the efficiency was lower. There are two possible reasons for the slower response of the two Eastern languages. First, Chinese and Japanese words are multi-character words with an average word length of 2-3 characters. When Japanese

words are spelled in Katakana or Hiragana, the average length is more than 2-3 characters. I adopted a character-based index in order to capture those multi-character words precisely. However, this approach could slightly hurt the efficiency compared to a word-based index. Second, the dictionaries I obtained for Chinese and Japanese contained many more entries than the ones for Spanish and German, which resulted in a much larger co-occurrence table for the two Eastern languages. This also increased the computational time in co-occurrence analysis.

It should be noticed that my prototype was run on a personal computer that is much less powerful than machines used in commercial search engines. The retrieval time would be much shorter on a powerful machine in a real Web retrieval system. With most calculations done during indexing time, the efficiency of the prototype is satisfactory.

4.8 Conclusions and Future Directions

Relatively large-scale test collections for MLIR experiments are available for evaluation of different retrieval approaches. However, few Web-based systems for online cross-lingual information retrieval are available. In this Chapter, I presented my experience in using a multilingual Web retrieval system with five languages (English, Chinese, Japanese, Spanish, and German) in the business IT domain. The system combines my knowledge of Web retrieval, system building, and MLIR techniques to address the need

for multilingual Web retrieval. An experiment was conducted to measure the effectiveness and efficiency of my Web portal, following TREC evaluation procedures. The prototype multilingual Web retrieval system achieved 62% to 85% of the performance level of monolingual retrieval. I also found that co-occurrence analysis improved performance for translation between English and any other language, while phrasal translation only improved performance for two Asian languages that had comprehensive bilingual dictionaries. In terms of efficiency, a multilingual retrieval took 6.13 to 12.15 seconds, or 1.24 to 2.99 times that of monolingual retrieval. The Web portal was reasonably efficient run on a PC and should achieve better efficiency on a more powerful machine. In summary, my study demonstrated the feasibility of applying MLIR techniques in Web applications and the experimental results are encouraging.

I plan to expand my research in several directions. First, I plan to conduct an interactive user evaluation of the usefulness of this multilingual Web retrieval system to real users. In such an interactive user evaluation, all the retrieved documents will be translated into the user's familiar language using a commercial machine translation product. I am also investigating how the speed of the system can be improved to achieve faster response time, which is necessary for a Web portal. In addition, I plan to expand the Web portal to more languages. Such expansion will allow us to study whether MLIR techniques will perform differently for a multilingual Web portal when more than two languages are involved. Lastly, because I believe that different domains might have different effects on

the performance of MLIR techniques, I am interested in testing my approach in other domains, such as public health and bioinformatics.

CHAPTER 5

NAME TRANSLITERATION BY COMBINING THE PROBABILITY MODEL AND THE WEB MINING MODEL

5.1 Introduction

The previous two Chapter described development of multilingual Web retrieval systems using a dictionary-based approach. However, in multilingual retrieval most proper names are unknown words that cannot be found in dictionaries, known as out-of-vocabulary (OOV) terms (Chen and Lee, 1998). Proper names, such as organizations, company names, product names, and person names, play an important role in search queries (Bian & Chen 2000). It was reported that 67.8%, 83.4%, and 38.8% of queries to the Wall Street Journal, Los Angeles Times, and Washington Post respectively involved name searching (Thompson and Dozier 1997). Those OOV proper names are some of the most difficult phrases to translate because they come from nowhere and are often domain specific (Al-Onaizan and Knight, 2001). During translation between language pairs employing the same alphabets (e.g., English/Spanish), proper names stay the same. For language pairs employing different alphabets (e.g., English/Arabic), proper names are translated phonetically, referred to as *transliteration*. For example, President “George Bush” is transliterated into Chinese as “乔治 布什” and the company name “SONY” is

transliterated into Arabic as “سوني.” Being able to identify correct transliterations of proper names as well as identify the origin of transliterated words would largely affect the precision of multilingual Web retrieval and would also be beneficial in machine translation systems or Question Answering systems. While the identification of proper names has received significant attention, transliteration of proper names has not (Al-Onaizan and Knight, 2002).

In this Chapter, I aim to develop a generic approach to enable automatic transliteration of Arabic proper names which combines an enhanced Hidden Markov Model (HMM) and a Web mining model. The rest of the Chapter is structured as follows: Section 5.2 reviews related research, in automatic transliteration and provides a taxonomy of existing approaches. In Section 5.3, I identify research gaps and present my research questions. In Section 5.4, I propose my transliteration framework, *ArizonaNameTran*. Section 5.5 discusses my experiment design and measures. In Section 5.6 I report and discuss experiment results. Finally, in Section 5.7 I conclude my work and suggest some future directions.

5.2 Related Works

5.2.1 Transliteration Problem

Transliteration is the representation of a word or phrase in the closest corresponding letters or characters of a language with different alphabet so that the pronunciation is as

close as possible to the original word or phrase (AbdulJaleel and Larkey, 2003). It can be classified in two directions: forward transliteration and back transliteration (Lin & Chen 2002). Consider a name pair (s, t) where s is the original proper noun in the source language and t is the transliterated word in the target language. Forward transliteration is the process of phonetically converting s into t . Back transliteration is the process of correctly finding or recovering s given t . Forward transliteration is a one-to-many mapping. Forward transliteration is a one-to-many mapping. For example, the Arabic name “مهند” can be back-transliterated into “Muhammed,” “Mohammed,” “Muhamed,” etc. Some transliterations might be more popular than others, but it is difficult to define one “correct” transliteration. On the other hand, back transliteration is a many-to-one mapping and has been identified as a more difficult task than forward transliteration for some language pairs (Stalls & Knight, 1998).

Furthermore, both directions have been explored with different language pairs in previous research. When (s, t) are both alphabet-based languages, such as English/Arabic, mapping can be applied from $s \rightarrow t$ directly. When (s, t) contains both alphabet-based and character-based languages, such as English/Chinese, a two-stage mapping is needed and a phonetic representation t' is often introduced. Transliteration is first performed from $s \rightarrow t'$, and then from $t' \rightarrow t$. For example, when translating the English name “Clinton” into Chinese, it is first transliterated into Chinese Pinyin form as “Ke Lin Dun” (Pinyin is a type of phonetic representation of Chinese). “Ke Lin Dun” is then transliterated into

Chinese characters as “克林顿.” Table 5.1 classifies previous research according to transliteration directions and type of languages studied.

Table 5.1: Transliteration problems studied in previous research

Direction	Forward transliteration	Back transliteration
Process	Phonetically convert to a foreign language	Recover the original name
Feature	One-to-many	Many-to-one
Examples	<p>Clinton->克林顿 ->柯林頓</p> <p>مهند -> Muhammed -> Mohammed</p> <p>قاعدة -> Al Qa'ida -> Al Qaeda -> Al Quieda</p>	<p>克林顿 -> Clinton 柯林頓</p> <p>Al Qa'ida -> قاعدة Al Qaeda -> Al Quieda -></p> <p>Muhammed -> مهند Mohammed</p>
Previous Research	<p>Arabic->English Arbabi et al. (1994)</p> <p>English->Arabic AbdulJaleel & Larkey (2002) Darwish et al. (2001) Al-Onaizan & Knight (2002)</p> <p>English->Chinese Wan & Verspoor (1998) Virga & Khudanpur (2003)</p>	<p>Arabic->English Stalls & Knight (1998)</p> <p>Thai->English Kawtrakul et al. (1998)</p> <p>Japanese->English Knight & Graehl (1997) Goto et al. (2001)</p> <p>Chinese->English Lin & Chen (2002)</p>

5.2.2 Transliteration Models Overview

Previous transliteration models can be categorized into four approaches: a rule-based approach, a machine learning approach, a statistical approach and a Web mining approach.

5.2.2.1 Rule-based Approach

A rule-based approach maps each letter or a group of letters in the source language to the closest sounding letter or letters in the target language according to pre-defined rules or mapping tables. Darwish et al. (2001) described a hand-crafted English to Arabic transliteration system. Each English letter was mapped to the closest sounding Arabic letter or letters. All the mappings rules were decided manually. Kawtrakul et al. (1998) presented a Thai-English back transliteration using an English phonetic dictionary. Wan and Verspoor (1998) described a two-step English to Chinese transliteration, which maps English into Pinyin and then map Pinyin into Chinese characters through table lookup.

The rule-based approach is straight forward and easy to implement. It does not rely on any training data. However, it requires manual identification of **all** transliteration rules and heuristics, which is a time-consuming process and sometimes error-prone (Darwish et al., 2001). Transliteration accuracy depends on the completeness of the rules. Due to the ambiguity of some rules, noise is often introduced. Moreover, this approach is not expandable to different languages pairs.

5.2.2.2 Machine Learning Approach

The machine learning approach has been adopted in previous research to improve rule-based mapping by filtering out unreliable translations trained from target language patterns. Arbabi et al. (1994) used a hybrid neural network and knowledge-based system approach in forward transliteration of Arabic personal names into the Roman alphabet. The neural network was trained on Arabic name samples, and it protects against inaccurate names generated by the rule-based system.

The machine learning approach helps eliminate some ambiguity in transliteration and can be generalized to multiple languages. However, transliteration improvement is often achieved based on a rule-based system. Although some ill-formed transliterations can be removed, it occasionally filters out good transliterations.

5.2.2.3 Statistical Approach

A statistical approach is the most promising approach. Instead of relying on a large set of language heuristics, a statistical approach obtains translation probabilities from a training corpus: pairs of transliterated words. This step also requires alignment of training pairs before calculating the probability model. Once the model is trained, on arriving at a new word, the statistical approach picks the transliteration candidate with the highest transliteration probability to generate as the correct transliteration.

Phoneme-based approach. Most previous statistical-based research used phoneme-based transliteration, relying on a pronunciation dictionary. Letter sequences in the source language are first mapped to a phonetic representation acquired from a dictionary, then mapped to letter sequences in the target language. Knight and Graehl (1997) described a phoneme-based probabilistic model for an English-Japanese back-transliteration system. Their probability model first transformed written English into English pronunciation, then to a Japanese sound inventory, and finally into written Japanese words (katakana). Using a similar approach, Stalls and Knight (1998) developed a probabilistic model of English->Arabic transliteration. The phoneme-based approach fails when such a dictionary is not available. Meng et al. (2001) reported 47.5% syllable accuracy during English-Chinese transliteration where 2,233 name pairs were used as the training corpus. More recently Virga and Khudanpur (2003) relied on a text-to-speech system to obtain phonemic pronunciation of each English name in English-Chinese name transliteration. Their training sample size was the same as in Meng's work.

Phoneme-based mapping is quite effective when a pronunciation dictionary is available. It handles multi-letter combinations successfully. However, only words with known pronunciation can be produced and it cannot deal with OOV terms. It could fail in back transliteration, since many foreign names, such as Muhammed, are not likely to be in a dictionary (Al-Onaizan and Knight, 2002).

Grapheme-based approach. The grapheme-based approach directly maps letter sequences in a source language into letter sequences in the target language with a probability model. This approach is often used for transliterations between two alphabet-based languages, such as English/Arabic, English/Russian, etc. Al-Onaizan and Knight (2002), in a study involving Arabic-English transliteration, showed that a grapheme-based model achieved better accuracy than a state-of-the-art phoneme-based model, and the mixed phoneme- and grapheme-based approach only slightly improved the accuracy over the grapheme-based approach. To filter out ill-formed name strings, they added a Web-based filtering step which eliminated candidates with zero Web counts. However, their transliteration model did not consider the context information of alphabets, which could harm performance. AbdulJaleel and Larkey (2002) also presented a grapheme-based statistical method for English to Arabic forward transliteration. They concluded that a bigram model outperformed a unigram model in English->Arabic transliteration, because the bigram model considers the context to some degree. They used 5,000, 10,000 and 50,000 name pairs respectively as training data, and reached 43.4% accuracy with a training sample of 50,000. But no significant differences were found with varied training sample sizes.

Unlike the phoneme-based approach, the grapheme-based approach does not require a phonetic dictionary or linguistic rules. However, it is likely that a given letter sequence in a source language might generate an ill-formed phoneme sequence in a target language in a solely grapheme-based mapping. It remains unknown whether the grapheme-based

statistical approach can be applied between alphabet-based and character-based languages, such as English-Chinese.

5.2.2.4 Web Mining-based Approach

The Web mining-based approach takes a very different view of the transliteration problem. This approach does not rely on transliteration heuristics or probability model. Instead, it searches the Internet for transliteration using relevant context words of the source name. Goto et al. (2001) proposed such an Internet-based technique for finding English equivalents for Japanese names. They first searched the Internet for relevant context words of the original name, and then used the translated context words as a query to obtain relevant Web documents. The assumption here was that the two name equivalents should share similar relevant context words in their languages. Correct transliteration is then extracted from the closest matching proper nouns. Similarly, Lu, Chien and Lee (2004) presented an approach to finding translation equivalents of query terms and constructing multilingual lexicons through the mining of Web anchor texts and link structures, which was shown to be effective on English-Chinese Web documents.

The Web mining approach is applicable to any pairs of languages. No rules, dictionaries, or training corpora are needed. However, the performance depends on the ability to identify proper names and accuracy in translating relevant context words. This approach works well for hotspots in news articles, but not normal names.

5.2.3 A Taxonomy of Transliteration Research

Proper name transliteration is an important problem in many applications. However, it was not widely studied. Based on my previous review, I present a taxonomy of transliteration approaches in Tables 5.2.1. A statistical approach shows the most promise. It does not rely on rules or heuristics; training data can be obtained fairly easily; and it achieved reasonably good accuracy in previous research.

Table 5.2.2 summarizes major research in statistical approach transliteration. The phoneme-based method was often used; however it fails when a pronunciation dictionary is not available. Al-Onaizan and Knight showed that a grapheme-based approach always outperformed a phoneme-based approach in Arabic-English transliteration.

Table 5.2.1: Taxonomy of Transliteration Research

Models	Resources	Descriptions	Examples
Rule-based	Mapping heuristics and knowledge	Transliteration is based on heuristics of source and target languages	Darwish et al. (2001) Wan & Verspoor (1998) Kawtrakul et al. (1998)
Machine learning enhanced	Training samples of words in target language	Machine learning algorithms such as Neural Network are used to filter out ill-formed transliterations	Arbabi et al. (1994)
Statistical approach	Training samples (list of transliteration pairs)	Translation probabilities are learned from a training sample of transliterated words in two languages	See Table 5.2.2
Web mining approach	Comparable Web context of proper names in both languages	Extract proper names from relevant context in both languages, and then compare their pronunciation similarity to match transliterations	Goto et al. (2001) Lu, Chien and Lee (2003)

Table 5.2.2: Taxonomy of Transliteration Research using Statistical Approach

Statistical Approach			
Models	Resources	Descriptions	Examples
<i>Grapheme-based</i>	Pairs of transliteration samples	Directly maps letter sequences in source language into letter sequences in target language	AbdulJaleel & Larkey (2002) Al-Onaizan & Knight (2002)
<i>Phoneme-based</i>	Phonetic dictionary; Pairs of training samples	Letter sequences in source language are mapped to their phonemic representations acquired from a dictionary first, and then mapped to letter sequences in target language	Virga & Khudanpur (2003) Knight & Graehl (1997) Stalls & Knight (1998) Meng et al.(2001)

5.3. Research Questions

Based on my review, several research gaps have been identified. Statistical approaches are the most promising, but little of the research has considered context information in the transliteration model. Although Al-Onaizan & Knight (2002) used Web counts to filter out unreliable transliteration, it remains unknown how and to what extent a Web mining model could enhance the probability model. It is a challenge to develop a generic approach for name transliteration to support knowledge discovery in multilingual content. I propose the following research questions.

1. How can I build a generic model for proper name transliteration for different language pairs?
2. Can context information and a Web mining component improve the transliteration performance (measured in accuracy)?
3. Can a generic name transliteration model achieve similar performance with different types of languages (character-based vs. alphabet-based)?

5.4 Proposed Framework: *Arizona NameTran*

Aiming to develop a generic framework with less human intervention and more easily obtained resources, I propose to adopt a grapheme-based statistical approach in proper name transliteration. Most previous research used a simple statistical approach with

independent probability estimation, assuming that transliteration of letters is context-independent. Correct transliteration is dependent on both source and target word context. I propose to use the Hidden Markov Model, which is one of the most popular probability models and has been used in speech recognition, the human genome project, consumer decision modeling, etc. (Rabiner, 1989), yet has seldom been explored in proper name transliteration. HMM fits the transliteration problem well. Since the model translates the current grapheme based on the observation of the previous grapheme transliterated, it captures context information. Furthermore, by examining the popularity of all possible transliterations on the Internet, bad transliterations can be filtered and their online popularity can serve as an indicator of transliteration correctness.

The proposed framework makes improvements in three aspects: 1) incorporating a simple phonetic transliteration knowledge base, 2) incorporating a bigram and a trigram HMM, and 3) incorporating a Web mining model to identify the most popular transliteration. It is composed of a training process and a transliteration process as shown in Figures 5.1 and 5.2. I explain the detailed components in each process in Sections 5.4.1 and 5.4.2.

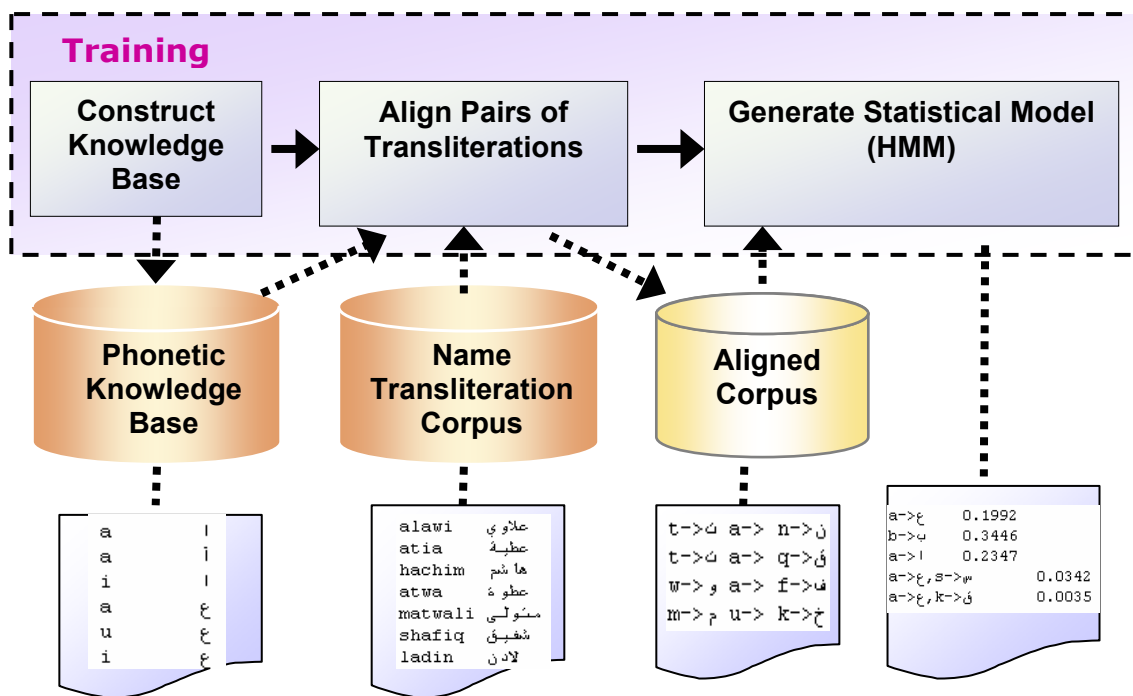


Figure 5.1: Training Statistical Model

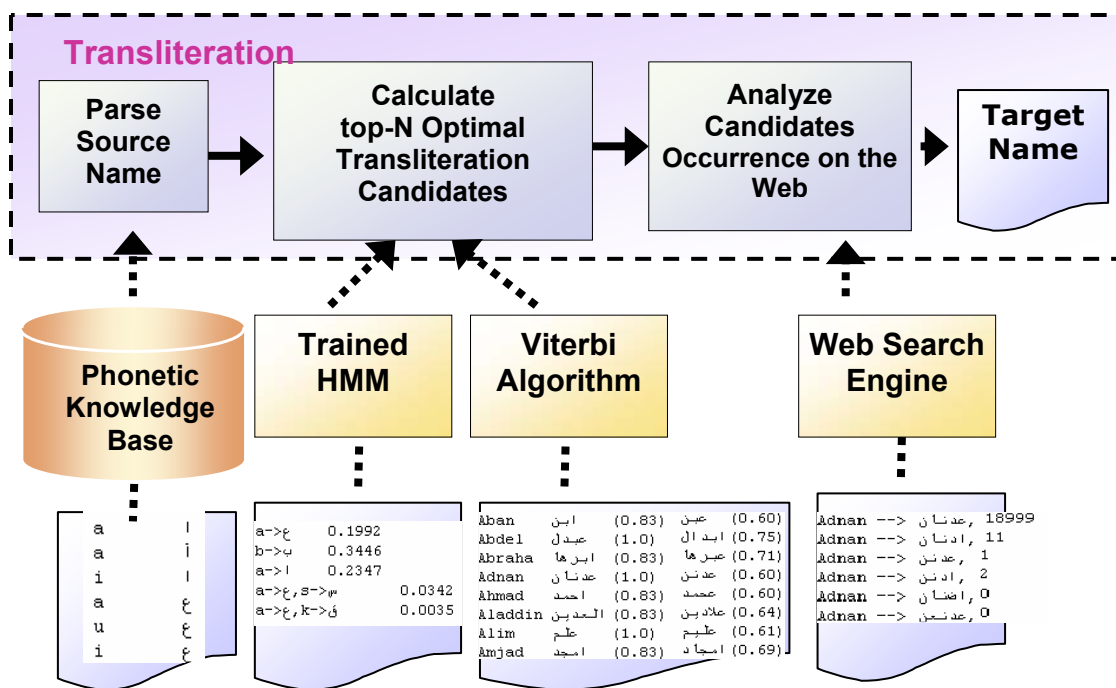


Figure 5.2: Transliteration Process

5.4.1 Training Statistical Model

The **training process** generates transliteration probabilities based on a training corpus (Figure 5.1). There are three steps in the training process: 1) to Construct a Knowledge Base, 2) to Align Pairs of Transliteration, and 3) to Generate the Statistical Model.

5.4.1.1 Phonetic Knowledge Base

The first step in training is to *Construct a Simple Phonetic Knowledge Base (KB)*, which consists of general phonetic rules for name parsing and alignment. In this step, multi-letter phonemes are identified as one transliteration unit. For example, “ow,” “th,” and “ee” are multi-letter phonemes in English. Restriction rules for alignment are also identified. For example, the English letter “a” can map to “ا,” “ع,” “ة,” or “ى” in Arabic. This is the component where some language specific features are captured. Note that the knowledge base is much less complex than what is used in a rule-based system.

5.4.1.2 Alignment

The second step is to *Align Pairs of Transliterations*. *Alignment* is a drawing that connects each letter or transliteration unit in the source language with a letter or transliteration unit in the target language. Different text alignment approaches have been proposed in Machine Translation and Cross-lingual Information Retrieval research, such as Finite State Automata (FSA), backtracking methods, the EM-algorithm, etc. Most of

them deal with complex linguistic context. Since word context is less complex than that of texts, I use a simple and efficient left-to-right, one-step-backtracking method to produce optimal alignment.

The alignment step starts from the first letter (or letter group) in the source name and assumes a mapping with the first letter (or letter group) in the target name if no restrictions are found in the KB. If KB violations are found, the program either jumps to the second letter in the target name for a potential mapping with the first letter in the source name, or it jumps to the second letter in the source name for a potential mapping with the first letter in the target name, and so on. An unmapped letter in the target name is considered to be an omitted pronunciation during transliteration, and an unmapped letter in the source name is considered to be an over-generalized pronunciation during transliteration.

```

For each pair of transliteration S, T {
  if (multi-letter phonemes are identified in knowledge base) {
    Tokenize transliteration pairs against multi-letter phonemes
  }
  let i=1, j=1 //Start from first letter s1 in s
  for each letter/unit in s {
    if ((map s(i) to t(j)) && no restrictions in KB are violated){
      Record s(i)→t(j) //aligned
    }
    elseif ((map s(i) to t(j+1)) && no restrictions in KB are
violated) {
      Record s(i)→ t(j+1); ε∧ t(j) //overgeneralized
pronunciation
    }
    elseif ((map s(i++) to t(j)) && no restrictions in KB are
violated) {
      Record s(i)→ ε; s(i++)∧t(j) //omitted pronunciation
    }
    i++; j++
  }
}

```

Figure 5.3: Pseudo-codes for word alignment process

5.4.1.3 Statistical Model

The last step in the training process is to *Generate the Statistical Model*, or the probability model. The model is derived from frequency counts of letter mappings observed in the aligned training corpus. Most previous research used a simple statistical model with independent probability estimation and I use this approach as my benchmark. I also investigate three more advanced statistical models: a bigram HMM, a trigram HMM and a combination of bigram and trigram HMM.

All statistical models try to find the candidate transliteration with the highest transliteration probabilities:

$$\arg \max P(t | s) = \arg \max P(t_1 t_2 \dots t_n | s_1 s_2 \dots s_m)$$

Where s is the source name to be transliterated, which contains letter string $s_1 s_2 \dots s_m$;

t is the target name, which contains letter string $t_1 t_2 \dots t_n$

In a simple statistical model, transliteration probability is estimated as:

$$P(t_1, t_2, t_3, \dots, t_n | s_1, s_2, s_3, \dots, s_n) = P(t_1 | s_1) P(t_2 | s_2) \dots P(t_n | s_n)$$

Where
$$P(t_n | s_n) = \frac{\# \text{ of times } s_n \text{ translates to } t_n \text{ in corpus}}{\# \text{ of times } s_n \text{ appears in corpus}}$$

The bigram HMM improves the simple statistical model in that it incorporates context information into a probability calculation. The transliteration of the current letter is dependent on the transliteration of ONE previous letter (one previous state in HMM).

Transliteration probability is estimated as:

$$P(t_1, t_2, t_3, \dots, t_n | s_1, s_2, s_3, \dots, s_n) = P(t_1 | s_1) P(t_2 | s_2, t_1) P(t_3 | s_3, t_2) \dots P(t_n | s_n, t_{n-1})$$

Where
$$P(t_n | s_n) = \frac{\# \text{ of times } s_n \text{ translates to } t_n}{\# \text{ of times } s_n \text{ occurs}}, \text{ and}$$

$$P(t_n | s_n, t_{n-1}) = \frac{\# \text{ of times } s_n \text{ translates to } t_n \text{ given } s_{n-1} \rightarrow t_{n-1}}{\# \text{ of times } s_{n-1} \text{ translates to } t_{n-1}}$$

In some cases, the translation probability of the current letter depends not only on one state before the current state (or one letter/transliteration unit before the current character), but on two or more states. The trigram HMM intends to capture even more context information by translating the current letter dependent on the TWO previous letters. Transliteration probability is estimated as:

$$P(t_1, t_2, t_3, \dots, t_n | s_1, s_2, s_3, \dots, s_n) = P(t_1 | s_1) p(t_2 | s_2, t_1) P(t_3 | s_3, t_2, t_1) \dots p(t_n | s_n, t_{n-1}, t_{n-2})$$

Where
$$P(t_n | s_n) = \frac{\# \text{ of times } s_n \text{ translates to } t_n}{\# \text{ of times } s_n \text{ occurs}},$$

$$P(t_n | s_n, t_{n-1}) = \frac{\# \text{ of times } s_n \text{ translates to } t_n \text{ given } s_{n-1} \rightarrow t_{n-1}}{\# \text{ of times } s_{n-1} \text{ translates to } t_{n-1}}$$

and
$$P(t_n | s_n, t_{n-1}, t_{n-2}) = \frac{\# \text{ of times } s_n \text{ translates to } t_n \text{ given } s_{n-1} \rightarrow t_{n-1} \text{ and } s_{n-2} \rightarrow t_{n-2}}{\# \text{ of times } s_{n-1} \text{ translates to } t_{n-1} \text{ and } s_{n-2} \text{ translates to } t_{n-2}}$$

The combined bigram and trigram model estimates the probability based on a weighed bigram and trigram probability.

5.4.2 Transliteration Process

The transliteration process transliterates proper names using the probability model obtained from the training process (Figure 5.2). It contains three steps: 1) to Parse Source

Names, 2) to Generate Top-N Transliteration Candidates and 3) to Analyze Candidates' Occurrences on the Web (Web mining approach).

5.4.2.1 Source Name Parsing

The first step in the transliteration process is to *Parse Source Names*. Source names are first tokenized against letters or multi-letter phonemes identified in the Phonetic Knowledge Base. These tokenized unites, most of which are single letters, are used as input for the statistical model.

5.4.2.2 Top-N Optimal Transliterations

The next step is to *Calculate Top-N Optimal Transliteration Candidates* based on trained probabilities. When feeding the model with a new proper name in the source language, the most probable transliteration is a letter sequence path that maximizes $P(t|s)$. As I described in Section 5.4.1, $P(t|s)$ is evaluated as a sequence of consecutive letter mappings and the conditional probability of each letter mapping can be estimated from the training corpus. In other words, $P(t|s)$ is calculated as the multiplication of sequences of conditional probabilities according to the statistical model used.

However, calculating all the possible sequences with such a large number of parameters is overwhelming. Thus, I use Viterbi's search algorithm for finding the most likely

sequence of target transliteration letters that result in a sequence of given source names. Viterbi's algorithm is a dynamic programming algorithm which is often used in the context of Hidden Markov Models (Viterbi, 1967). Instead of keeping one optimal path, I keep the top-N optimal paths as my transliteration candidates.

5.4.2.3 Web Occurrence Analysis

To boost the transliteration performance I propose to use the Web mining approach, which *Analyzes Candidates' Occurrence on the Web*. Each one of the top-N transliterations obtained from the previous step is sent to a Web search engine using a meta-search program which records the number of documents retrieved, referred to as Web frequency. This information is an indicator of the candidate transliteration's online popularity. The more often the candidate transliteration appears in online documents, the more likely it is a correct transliteration.

Unlike Al-Onaizan and Knight's work (2002), I do not throw away candidates with zero Web counts. Both Web frequency information and transliteration probability of top-N candidates contribute to the final score formula that is used to rank transliteration candidates.

$$\textit{Final score} = \alpha * \textit{normalized probability score} + \beta * \textit{normalized Web frequency},$$

$$\textit{where } \alpha + \beta = 1.$$

This final rank of transliterations is derived from a weighed score of the normalized Web frequency and a probability score. On the one hand, even though I am using Web mining for disambiguation, I do not want to treat all the top-N transliteration candidates equally. Instead, I retain information from the probability model. In this way, if two transliterations have a similar Web frequency score (e.g. 128,000 vs. 128,001) their probability scores will play a major role in selecting the best transliteration. On the other hand, I still want to distinguish between different Web frequency counts if the difference is big enough. In transliteration the occurrence difference between 128,000 and 1 should have a much bigger effect than the difference between 128,000 and 127,000,, in which case the Web frequency score will play a more important role in the final ranking score. In my framework, I chose $\alpha=0.5$ and $\beta=0.5$ to generate the final score. All the transliteration candidates are then ranked by their final scores.

5.5 Experiment Design and Hypotheses

I designed experiments to study the performance of my proposed research framework using different statistical models and using a Web mining model. In this section, I present the hypotheses and experimental design.

5.5.1 Hypotheses

I am interested in the performance of five experimental settings: 1) A simple statistical approach, 2) A bigram HMM approach, 3) A trigram HMM approach, 4) A hybrid HMM approach (bigram + trigram) and 5) A Web-mining-enhanced approach (all four previous settings + Web mining).

In H1.1- H1.3, I studied the performance of the probability model alone. A simple statistical approach has been adopted in previous transliteration research, and I used it as my benchmark. A bigram HMM is a traditional HMM, which predicts the grapheme transliteration based on the conditional probability of one previous grapheme transliteration observed in training data. I believe that incorporating the Hidden Markov Model would improve performance. A trigram HMM is an improved HMM which integrates a more complex conditional probability model and captures two previous grapheme transliterations. It provides a stronger relation between word graphemes. Furthermore, I believe that a combined bigram and trigram model could complement each other and further improve the performance over a trigram model alone. Thus, I hypothesized that:

H1.1: A bigram HMM approach performs better than a simple statistical approach.

H1.2: A trigram HMM approach performs better than a bigram HMM approach.

H1.3: A hybrid HMM approach performs better than a trigram HMM or a bigram HMM alone.

In H2.1-2.4, I looked at the effect of integrating a Web mining model with probability models. A Web mining model provides additional information on transliteration online popularity. I believed that a combined model would always outperform a single probability model. Thus, I hypothesized that:

H2.1: Integrating a Web mining model improves a simple statistical approach significantly.

H2.2: Integrating a Web mining model improves a bigram HMM approach significantly.

H2.3: Integrating a Web mining model improves a trigram HMM approach significantly.

H2.4: Integrating a Web mining model improves a hybrid HMM approach significantly.

5.5.2 Experiment Measure

5.5.2.1 Rigid and Relaxed Accuracy

Previous transliteration research has used “accuracy” to measure performance which is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Transliterations}}{\text{Total Number of Transliterations}}$$

As I discussed in Section 5.2.1, back transliteration is a many-to-one mapping. Although one Arabic name can have as many as forty English variations, a given English transliteration of Arabic name only has one origin in Arabic. Thus, there is only one correct transliteration for a given name in back transliteration.

However, it is difficult to judge correct transliteration during forward transliteration, which is a one-to-many mapping. One English name could have more than five Chinese transliterations that are acceptable to human. In this case, there are two ways to define “correct transliterations” (Al-Onaizan and Knight, 2002). In a *rigid accuracy*, there is one and only one correct transliteration. A transliteration is considered correct if it matches the pre-defined gold-standard. Gold-standard can be translations extracted from a dictionary. Or it can be judged by human experts as the most widely used or most acceptable transliteration. In a *relaxed accuracy*, a transliteration is considered correct if it is acceptable to human experts. A relaxed accuracy allows multiple correct answers. However, it is often difficult to generate all the possible translations in advance. Relaxed accuracy can be done by judging the machine-generated transliterations afterwards.

For example, President Clinton is transliterated as “克林顿” in most Chinese news articles, which is considered a gold-standard transliteration. The transliteration system could generate a different name “柯林顿”, which has the same pronunciation as the gold-standard, and is actually used in many contexts as well, yet not as dominant. My expert could decide that it was acceptable. In a rigid accuracy judgment, the machine-generated

“柯林顿” will not be considered as a correct transliteration, while in a relaxed accuracy it will be treated as a correct one.

Since rigid accuracy is a more stringent test which avoids human judgment variations, I chose to use rigid accuracy for my measure. Note that a relaxed accuracy is always higher than a rigid accuracy, and a rigid accuracy can be viewed as the lower bound of system accuracy.

5.5.2.2 Top-N Accuracy

Besides measuring the accuracy for the highest ranked transliteration, identifying a set of top-N transliteration candidates are of interest. Top N accuracy is defined as the percentage of names whose selected top N transliterations include correct transliterations.

Among all the ranked transliteration candidates, top-N accuracy is defined as

$$\text{Top-N Accuracy} = \frac{\text{Number of Times Correct Transliterations appeared in the first } N \text{ Candidates}}{\text{Total Number of Transliterations Performed}}$$

In my experiments, I chose $N=1, 2, 4, 8$.

5.5.3 Language Pairs and Dataset

In order to study languages with diverse features, I selected two pairs of languages, English-Arabic back transliteration and English-Chinese forward transliteration. English

and Arabic are both alphabet-based languages and Arabic is well known for its large number of transliteration variations. I chose back transliteration in this language pair because it is more challenging and of practical interest. On the other hand, Chinese is a character-based language. When transliterating from an alphabet-based language to a character-based language, two-phase transliteration is conducted: English->Chinese phonetic representation (Pinyin) followed by Chinese phonetic representation->Chinese character.

My English-Arabic transliteration dataset is a list of 1000 unique Arabic names extracted from <http://www.ummah.net/family/masc.html>. An Arabic-speaking expert manually translated all Arabic names into English transliterations according to his knowledge. Because I focus on back transliteration from English to Arabic, each English transliteration was translated into one and only one correct Arabic origin. My English-Chinese dataset is a list of 2000 unique English names and their transliterations extracted from a bilingual dictionary. These dictionary-provided translation are considered to be gold-standard in my experiments. Both datasets are unaligned.

5.5.4 Experiment Methodology

I used the 10-fold cross validation method, commonly used in testing data mining algorithms and models, to test system accuracy. I first randomly divided the data into 10 subsets of equal size. I trained the model 10 times, each time leaving out one of the

subsets, to compute the system's top-N accuracy. Accuracy scores obtained from each subset were then averaged.

5.6 Experiment Results

In this section I describe and analyze the results of my experiments. I present results from English-Arabic back transliteration and English-Chinese forward transliteration respectively.

5.6.1 English-Arabic back transliteration

Table 5.3 presents the overall transliteration performance results (measured by top-N accuracy) under five experiment conditions and their improvement over a simple statistical model. The best performance was achieved using a combined hybrid HMM and Web mining model (column 9), a 0.38 top-1 accuracy and a 0.72 top-8 accuracy. Bigram HMM, hybrid HMM, and a combined hybrid HMM and Web mining model enhanced a simple statistical approach by 20.87%, 62.09%, and 79.05% respectively for top-1 accuracy. Surprisingly, trigram HMM degraded the performance by 81.59%. Improvements in the top-2, top-4 and top-8 accuracy were not as tremendous as that of the top-1, ranging from 3.88% (top-8 accuracy for bigram) to 25.88% (top-8 accuracy for Web mining enhanced).

Table 5.3: Summary of system performance (accuracy) with different models and their improvement over a Simple Statistical model (English-Arabic)

	Simple	Bigram	Impr. over Simple	Trigram	Impr. over Simple	Hybrid	Impr. over Simple	Web mining enhanced	Impr. over Simple
Top-1	0.21	0.26	20.87%	0.04	-81.59%	0.34	62.09%	0.38	79.05%
Top-2	0.41	0.44	5.20%	0.08	-79.84%	0.50	20.54%	0.52	25.88%
Top-4	0.57	0.59	3.09%	0.12	-78.58%	0.63	9.94%	0.64	11.34%
Top-8	0.66	0.69	3.88%	0.22	-66.72%	0.72	8.26%	0.72	8.26%

Simple: simple statistical model

Bigram: bigram Hidden Markov Model

Trigram: trigram Hidden Markov Model

Hybrid: Hybrid Hidden Markov Model (bigram + trigram)

Web-mining-enhanced: Hybrid Hidden Markov Model + Web mining model

Impr. Over Simple: Improvement achieved over simple statistical model

5.6.1.1 Comparison of Probability Models

Table 5.4 reports top-N (N=1, 2, 4, 8) accuracy achieved from four different probability models. I provide my paired *t*-test results in testing and Figure 5.3 illustrates the difference among all four approaches.

The results for my hypotheses showed that H1.1 and H1.3 were supported, but H1.2 was not. There were significant improvements from the simple statistical approach to the bigram HMM approach. However, the performance significantly decreased when using trigram HMM alone. There are two possible causes for drop in accuracy. First, I believe that overall bigram HMM is a better model for English-Arabic transliteration. Most Arabic name transliteration processes depend on just one letter ahead of the current one, instead of two letters ahead. Second, I observed that because trigram HMM is a strong

relation, it needs a large training dataset to obtain all possible triple-letter sequences. My training data of 900 Arabic names might not be sufficient. The hybrid model performed significantly better than a bigram model, which implied that when trigram probability existed in training data it helped improve the performance. I concluded that a hybrid HMM which combines bigram and trigram information yielded best performance in my experiments and the top-8 accuracy reached 0.72 with a relatively small training dataset.

Table 5.4: Summary of average accuracy achieved and *t*-test results (English-Arabic)

Summary of results (Probability models)				
	Simple	Bigram	Trigram	Hybrid
Top-1	0.21	0.26	0.04	0.34
Top-2	0.41	0.44	0.08	0.50
Top-3	0.57	0.59	0.12	0.63
Top-4	0.66	0.69	0.22	0.72
Paired <i>t</i> -test (2 tail, $\alpha=0.05$)				
P_{simple}		1.09E-05	2.2E-20	2.13E-11
P_{bigram}			1.29E-22	1.14E-10
$P_{trigram}$				9.62E-28

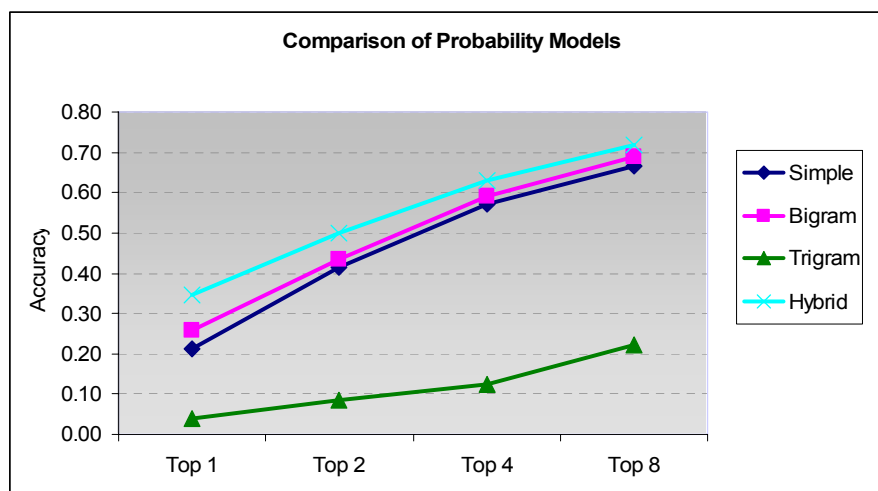


Figure 5.3: Performance comparison of probability models (accuracy) (English-Arabic)

5.6.1.2 Performance of Web Mining Model

Table 5.5 summarizes the average accuracy achieved and paired *t*-test results of comparing probability model alone and a combined probability and Web mining model. Figure 4 illustrates the improvements obtained from the Web mining model.

As I hypothesized in H2.1 to H2.4, Web mining always advanced the performance of the probability model significantly, no matter which probability model I use. H2.1-H2.4 were all supported. This confirmed that online occurrence information obtained from search engines is an effective way to identify the correct transliterations. There is a pattern of larger enhancement on lower accuracy, and smaller enhancement on higher accuracy. The improvements achieved on top-1 and top-2 accuracy were more obvious than that obtained on top-4 and top-8 accuracy. Similarly, the boosting effects on simple, bigram and trigram models were more noticeable than that in a hybrid model.

Table 5.5: Summary of average accuracy achieved and *t*-test results using combined Probability and Web mining model (English-Arabic)

Summary of results (Probability + Web mining models)								
	Simple	Simple +Web mining	Bigram	Bigram +Web mining	Trigram	Trigram +Web mining	Hybrid	Hybrid +Web mining
Top-1	0.21	0.34	0.26	0.37	0.04	0.13	0.34	0.38
Top-2	0.41	0.48	0.44	0.51	0.08	0.18	0.50	0.52
Top-4	0.57	0.62	0.59	0.63	0.12	0.21	0.63	0.64
Top-8	0.66	0.68	0.69	0.69	0.22	0.22	0.72	0.72
Paired <i>t</i>-test (2 tail, $\alpha=0.05$)								
		H2.1		H2.2		H2.3		H2.4
<i>P value</i>		2.44E-08		2.76E-08		2.76E-08		0.0001

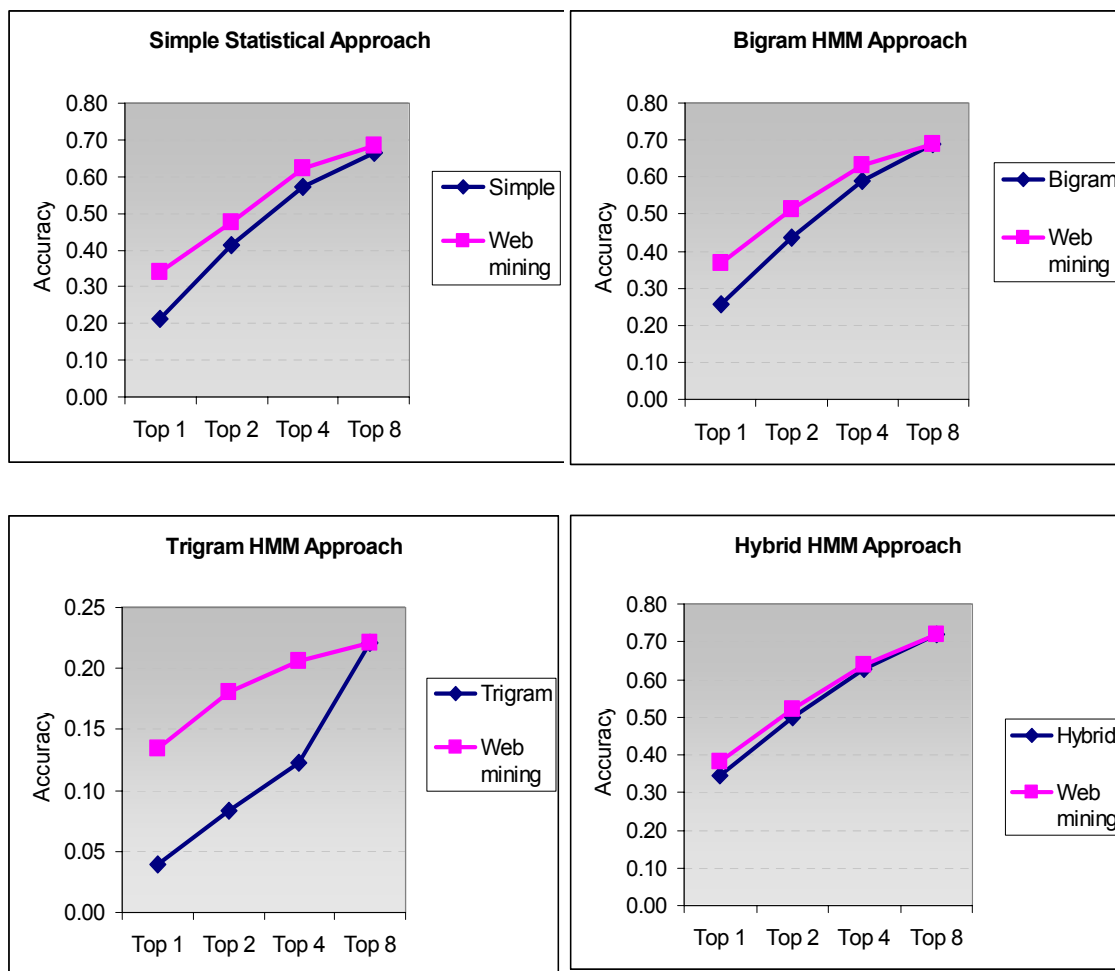


Figure 5.4: Performance comparison of combined probability and Web mining models (accuracy) (English-Arabic)

5.6.2 English-Chinese forward transliteration

Table 5.6 presents the overall transliteration performance results (measured by top-N accuracy) under five experiment conditions and their improvement over a simple statistical model. English-Chinese transliteration achieved extremely good performance.

The best performance was achieved using a combined hybrid HMM and Web mining model (column 9), a 0.64 top-1 accuracy and a 0.96 top-8 accuracy. Bigram HMM, Trigram HMM, hybrid HMM, and a combined hybrid HMM and Web mining model enhanced a simple statistical approach by 244.08%, 325.81%, 414.24% and 415.01% respectively for top-1 accuracy. Improvement reached in top-2, top-4 and top-8 accuracy were not as tremendous as that of top-1, yielded from 69.58% (top-8 accuracy for bigram) to 142.43% (top-8 accuracy for Web mining enhanced).

Table 5.6: Summary of system performance (accuracy) with different models and their improvement over a Simple Statistical model (English-Chinese)

	Simple	Bigram	Impr. over Simple	Trigram	Impr. over Simple	Hybrid	Impr. over Simple	Web mining enhanced	Impr. over Simple
Top-1	0.12	0.43	244.08%	0.53	325.81%	0.64	414.24%	0.64	415.01%
Top-2	0.20	0.55	170.21%	0.65	218.97%	0.75	267.48%	0.80	293.37%
Top-4	0.30	0.63	109.99%	0.73	144.00%	0.81	169.70%	0.91	202.61%
Top-8	0.40	0.67	69.58%	0.77	94.76%	0.85	114.30%	0.96	142.43%

5.6.2.1 Comparison of Probability Models

English to Pinyin and Pinyin to Chinese Characters

English to Chinese transliteration is a two-step process, which involves 1) English to Pinyin and 2) Pinyin to Chinese character transliterations. Tables 5.7 and 5.8 report top-N (N=1, 2, 4, 8) accuracy achieved from the four different probability models during English to Pinyin and Pinyin to Chinese character transliteration. H1.1, H1.2 and H1.3 were all confirmed in these two individual steps (Bigram+Trigram>Trigram>Bigram).

All enhancements were significant. Figures 5.5 and 5.6 illustrate the accuracy performance. During English to Pinyin transliteration, Bigram, Trigram and Hybrid Hidden Markov Model achieved great performance improvements over simple statistical approach. But during Pinyin to Chinese character transliteration, the difference between those models were not as obvious. This is due to the fact that alphabets mapping between English and Pinyin has a stronger correlation between letter sequences. The correlation between letter to character mapping and character sequence is weaker.

Table 5.7: Summary of average accuracy achieved and t -test results (English to Pinyin)

Summary of results (Probability models)				
	Simple	Bigram	Trigram	Hybrid
Top-1	0.20	0.40	0.47	0.54
Top-2	0.32	0.53	0.60	0.67
Top-3	0.41	0.61	0.68	0.75
Top-4	0.50	0.65	0.73	0.81
Paired t -test (2 tail, $\alpha=0.05$)				
P_{simple}		7.7E-30	5.14E-37	1.55E-38
P_{bigram}			3.2E-26	3.2E-26
$P_{trigram}$				3.2E-26

Table 5.8: Summary of average accuracy achieved and t -test results (Pinyin to Chinese)

Summary of results (Probability models)				
	Simple	Bigram	Trigram	Hybrid
Top-1	0.58	0.83	0.87	0.91
Top-2	0.80	0.95	0.96	0.98
Top-3	0.91	0.99	0.99	1.00
Top-4	0.97	1.00	1.00	1.00
Paired t -test (2 tail, $\alpha=0.05$)				
P_{simple}		2.47E-11	8.99E-11	2.84E-10
P_{bigram}			1.63E-05	1.63E-05
$P_{trigram}$				1.63E-05

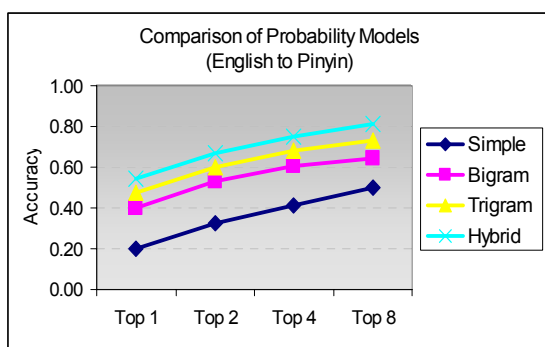


Figure 5.5: Performance comparison of probability models (accuracy) (English-Pinyin)

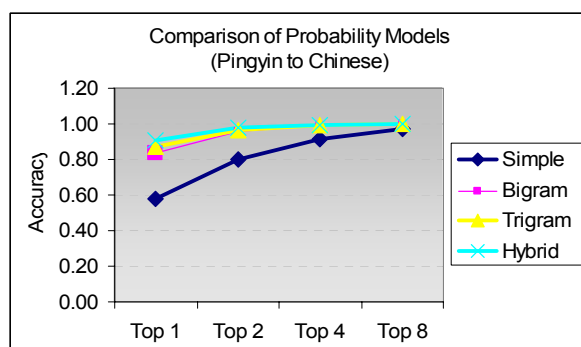


Figure 5.6: Performance comparison of probability models (accuracy) (Pinyin-Chinese)

English to Chinese Characters

Table 5.9 reports top-N (N=1, 2, 4, 8) accuracy achieved from four different probability models. I provided my paired t -test results in Table 5.10. Figure 7 illustrates the difference among all four approaches.

The results of 10-fold experiments again showed that H1.1, H1.2 and H1.3 were all supported. There were significant improvements from simple statistical approach to bigram HMM approach, from trigram to bigram HMM, and from Hybrid HMM to trigram HMM alone. I concluded that a hybrid HMM which combines bigram and trigram information yielded best performance in my experiments.

Table 5.9: Summary of average accuracy achieved and t -test results (English to Chinese)

Summary of results (Probability models)				
	Simple	Bigram	Trigram	Hybrid
Top 1	0.12	0.43	0.53	0.64
Top 2	0.20	0.55	0.65	0.75
Top 3	0.30	0.63	0.73	0.81
Top 4	0.40	0.67	0.77	0.85
Paired t -test (2 tail, $\alpha=0.05$)				
P_{simple}		6.14E-36	2.15E-41	4.4E-44
P_{bigram}			2.74E-54	8.79E-32
$P_{trigram}$				8.32E-21

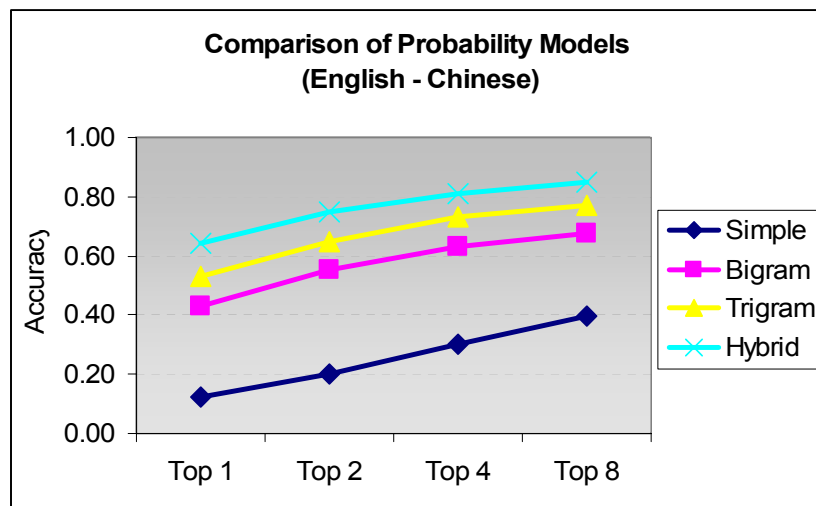


Figure 5.7: Performance comparison of probability models (accuracy) (English-Chinese)

5.6.2.2 Performance of Web Mining Model

Table 5.10 summarizes average accuracy achieved and paired *t*-test results of comparing probability model alone and a combined probability and Web mining model. Figure 5.8 illustrates the improvements obtained from Web mining model.

I observed in English-Arabic back transliteration that no matter which probability model I used, Web mining always advanced the performance of probability model significantly. H2.1-H2.4 were all supported. This again confirmed that online occurrence information obtained from search engines is an effective way to identify the correct transliterations.

Table 5.10: Summary of average accuracy achieved and *t*-test results using combined Probability and Web mining model (English-Chinese)

Summary of results (Probability + Web mining models)								
	Simple	Simple +Web mining	Bigram	Bigram +Web mining	Trigram	Trigram +Web mining	Hybrid	Hybrid +Web mining
Top 1	0.12	0.17	0.43	0.47	0.53	0.58	0.64	0.64
Top 2	0.20	0.25	0.55	0.61	0.65	0.71	0.75	0.80
Top 4	0.30	0.37	0.63	0.70	0.73	0.81	0.81	0.91
Top 8	0.40	0.49	0.67	0.76	0.77	0.86	0.85	0.96
Paired <i>t</i>-test (2 tail, $\alpha=0.05$)								
<i>P value</i>	2.29E-15		1.38E-15		1.41E-16		4.37E-10	

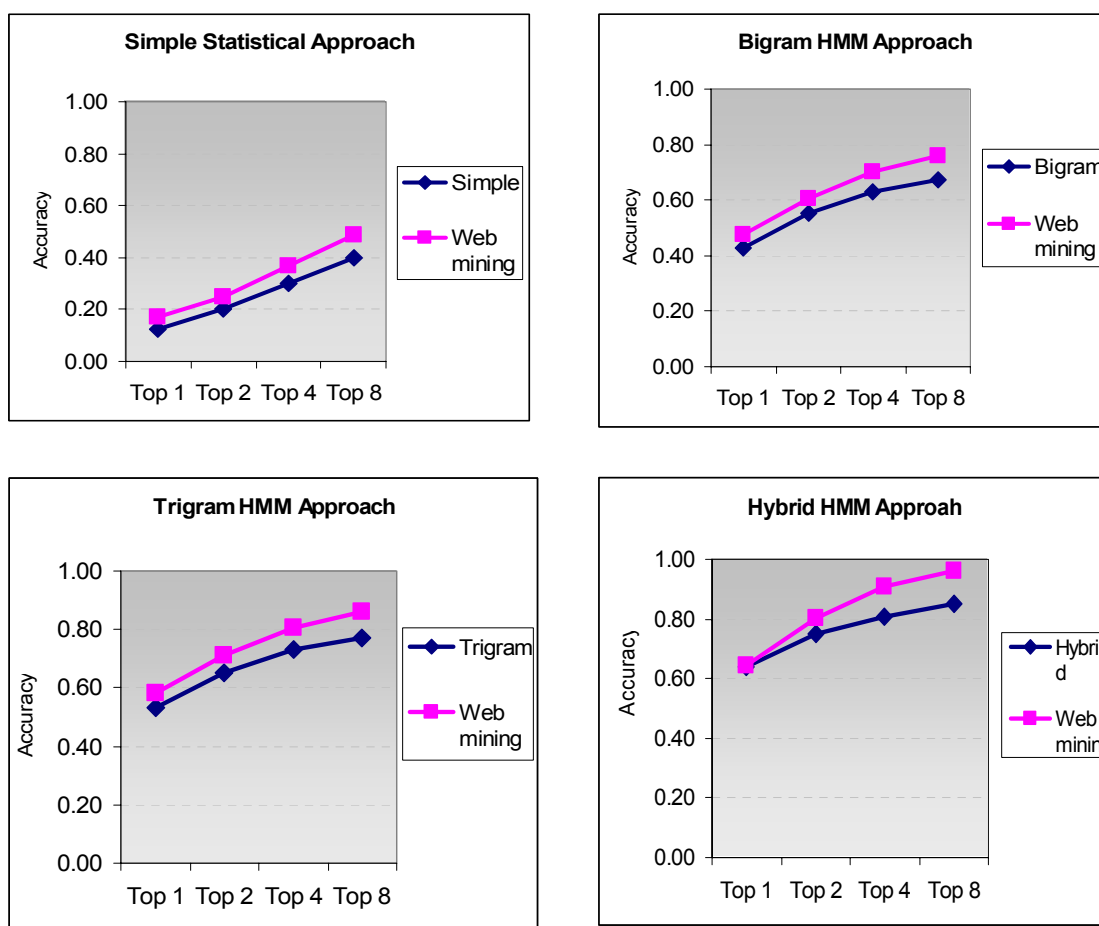


Figure 5.8: Performance comparison of combined probability and Web mining models (accuracy) (English-Chinese)

5.6. Conclusions and Discussions

In this research I proposed a generic proper name transliteration framework which incorporated the Hidden Markov Model and Web mining approaches. I evaluated the framework in two pairs of languages, English-Arabic and English-Chinese. For English-Arabic transliteration, I found that when using HMM alone, a combination of a bigram and a trigram HMM method performed best. While a bigram model achieved fairly good performance, a trigram model alone did not. The combined HMM and the Web mining approach boosted the performance of simple statistical model by 79.05% for top-1 accuracy. For English-Chinese transliteration, again I concluded that a combination of the bigram and the trigram HMM method performed best when using the HMM approach alone. The combined HMM and Web mining approach boosted the performance over simple statistical model by 415.01% for top-1 accuracy. Trigram was a better model for Chinese transliteration, out-performing the bigram approach. Frequency information obtained from the Web proved an effective way to identify the correct transliteration. However, the boosting effect was not as big as in English-Arabic back transliteration. Compared to previous research, my framework achieved better performance.

My framework of transliteration has several practical applications. For example, it could improve the performance of current multilingual Web retrieval by transliterating out-of-vocabulary proper nouns. It could also be adopted in machine translation systems. In the future, I plan to test my framework on more language pairs and incorporate a transliteration component into multilingual Web retrieval systems.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

The Internet provides the largest knowledge repository across disciplines, countries, and languages. It is desirable for researchers, managers, and government agencies to access, understand, analyze and share such multilingual information. This dissertation investigates effective and efficient approaches that support multilingual Internet searching and browsing. In this Chapter, I summarize the main conclusions and contributions of this dissertation, discuss the relevance of this research to Management Information Systems research, and suggest future directions.

6.1 Conclusions

The first case study of this dissertation explores an effective framework to support non-English Web retrieval. The framework is consistent with the architecture of most search engines. It features three extensions to this basic structure: generic language processing ability; integration of multiple information resources; and post-retrieval analysis. Various techniques and algorithms that could facilitate Internet searching and browsing were adopted and evaluated in a Chinese search engine in the medical domain. A systematic evaluation has been conducted to study the effectiveness and efficiency of CMedPort in assisting human analysis. The experimental results show that CMedPort achieved

significant improvement in searching and browsing performance compared to three benchmark regional search engines, Sina, Yahoo! Hong Kong, and Openfind. CMedPort's collection building method, meta-searching and cross-regional searching contributed to the improvement in information seeking. Although post-retrieval analysis methods, such as categorizer and summarizer, did not further improve browsing performance significantly, users' subjective evaluation and verbal comments revealed that they appreciated these analysis functions.

The second study of the dissertation investigates a more complicated problem: Web retrieval across two languages, English and Chinese. Web retrieval and Cross-lingual information retrieval techniques were adopted in an English-Chinese multilingual Web retrieval system in the business IT domain. An experiment was conducted to measure the effectiveness and efficiency of my Web portal following TREC evaluation procedures. My results showed that my system's phrasal translation and co-occurrence disambiguation led to great improvement in performance, while query expansion techniques did not improve results further. The Web portal was reasonably efficient on a PC and should achieve better efficiency on a more powerful machine. In sum, my study demonstrated the feasibility of applying CLIR techniques to Web applications and the experimental results are encouraging.

The third study of the dissertation further extended the Web-based Cross-lingual retrieval into five languages: English, Chinese, Japanese, Spanish and German. The system again

combines the knowledge of Web retrieval, system building, and Cross-lingual information retrieval techniques to address the need for multilingual Web retrieval. The prototype multilingual Web retrieval system achieved 62% to 85% of the performance level of monolingual retrieval. I also found that co-occurrence analysis improved performance for translation between English and any other language, while phrasal translation only improved performance for two Asian languages that had comprehensive bilingual dictionaries. This study provides a comprehensive study of multilingual Web retrieval applications. The system framework can be extended to other languages and other topical domains.

The fourth study of the dissertation investigates an important problem in cross-lingual Web retrieval: transliteration of out-of-vocabulary (OOV) proper nouns. I proposed a generic proper name transliteration framework which incorporated a Hidden Markov Model and a Web mining model. The framework has been evaluated with two pairs of languages, English-Arabic and English-Chinese. For English-Arabic transliteration, I found that when using HMM alone, a combination of a bigram and a trigram HMM method performed the best. While the bigram model achieved fairly good performance, the trigram model alone did not. For English-Chinese transliteration, again, a combination of the bigram and trigram HMM method performed the best among all probability models. Trigram is a better model for Chinese transliteration, out-performing the bigram approach. In both language pairs, a Web mining model greatly boosted the

performance of simple statistical model. This work contributes to the field of name transliteration by combining a Hidden Markov Model and a Web mining model.

6.2 Future Directions

I plan to expand my research in several directions. First, I plan to continue development of effective and efficient techniques and algorithms that support multilingual systems. In particular, the out-of-vocabulary problem in multilingual Web retrieval can be further investigated. Secondly, I plan to conduct large scale interactive user evaluations of such multilingual Web retrieval systems. This is an area that has not been widely investigated. I am also interested in studying the Human-Computer Interaction (HCI) design issues and examining the behavioral, organizational, and social impacts of multilingual knowledge discovery systems. In addition, I plan to expand the Web portal to more languages, especially languages that have not been widely studied in previous research. Lastly, I will experiment with my techniques in more application domains. This dissertation covers only a few domains where multilingual Web retrieval techniques can apply. I will apply the techniques in other domains such as bioinformatics.

6.3 Relevance to Business and MIS Research

The major application domain of this research is in business intelligence. As discuss in the introduction, there are a wide variety of circumstances in which a user totally

unfamiliar with the language of the document collection might find multilingual retrieval useful. For instance, intelligence agencies seeking global intelligence, national security agencies seeking terrorism information, researchers seeking to determine who has conducted research on a particular topic, companies seeking international business communications and opportunities, and so on. My research framework can contribute to various needs of multilingual business applications. It is especially important to businesses relying on the Internet. Managers of international organizations may find a number of new opportunities for business and management by employing multilingual Web retrieval techniques presented in this dissertation.

This research also contributes to the MIS research in general by introducing additional reference disciplines including cross-lingual information retrieval and Web mining to business and information systems applications. A wide range of complex systems in the business world can benefit from such multilingual research.

REFERENCES

- AbdulJaleel, N. and Larkey, L. (2003), "Statistical Transliteration for English-Arabic Cross-Language Information Retrieval" in *CIKM '03*, pp. 139-146.
- Al-Onaizan, Y. and Knight, K. (2001). Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp. 400 – 408.
- Al-Onaizan, Y. and Knight, K. "Translating Named Entities Using Monolingual and Bilingual Resources," In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2002.
- Aljlal, M., Frieder, O., and Grossman, D. (2002). "On bidirectional English-Arabic Search," *Journal of the American Society for Information Science and Technology*, 53(13), 1139-1151.
- Arbabi, M., Fischthal, S. M., Cheng, V. C. and Bart, E. (1994), "Algorithms for Arabic name transliteration," *IBM Journal of Research and Development*, v.38 n.2, p.183-194, March 1994.
- Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan S. (2001). "Searching the Web," *ACM Transactions on Internet Technology*, 1(1), 2-43.
- Ballesteros, L. and Croft, B. (1996). "Dictionary Methods for Cross-lingual Information Retrieval," in *Proceedings of the 7th DEXA Conference on Database and Expert Systems Applications*, Zurich, Switzerland, September 1996, pp. 791-801.
- Ballesteros, L. and Croft, B. (1997). "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval," in *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, July 1997, pp. 84-91.
- Ballesteros, L. and Croft, B. (1998). "Resolving Ambiguity for Cross-language Retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998, pp. 64-71.
- Ballesteros, L. (2000), 'Cross-Language Retrieval via Transitive Translation', In Croft, W. B. (ed.) *Advances in Information Retrieval: Recent Research from the CII*, pp. 203-234. Kluwer Academic Publishers.

Belkin, N. J., Marchetti, P. G., and Cool, C. (1993). "BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval," *Information Processing and Management* 29 (3), 325-344.

Bergmark, D., Lagoze, C., et al. (2002). "Focused Crawls, Tunneling, and Digital Libraries," in *Proceedings of the European Conference on Digital Libraries*, Rome, Italy, September 2002

Bian, G. W. & Chen, H. H (2000). Cross-language information access to multilingual collections on the internet. *Journal of the American Society for Information Science*, Special Issue: Digital Libraries, 51(3) pp. 281-296

Blair, D. C. and Maron, M. E. (1985). "An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System," *Communications of the ACM*, 28(3), 289–299.

Capstick, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Cagno, F., Gadaleta, G., Hernandez, J. A., Korte, R., Leisenberg, A., Leisenberg, M., & Christ, O. (1998). "MULINEX: Multilingual Web Search and Navigation," in *Proceedings of Natural Language Processing and Industrial Applications*, Moncton, Canada, 1998.

Carmel, E., Crawford, S. and Chen, H. (1992). "Browsing in Hypertext: A Cognitive Study," *IEEE Transactions on System, Man and Cybernetics*, 22(5), 865–884.

Chakrabarti, S., van den Berg, M. and Dom, B. (1999). "Focused Crawling: A New Approach to Topic-specific Web Resource Discovery," In *Proceedings of the 8th International World Wide Web Conference*, Toronto, Canada.

Chau, M., Chen, H., Qin, J., Zhou, Y., Qin, Y., Sung, W. and McDonald, D. (2002). "Comparison of Two Approaches to Building a Vertical Search Tool: A Case Study in the Nanotechnology Domain," In *Proceedings of JCDL'02*, Portland, Oregon, USA, 135–144, ACM Press.

Chau, M. and Chen, H. (2003). "Comparison of Three Vertical Search Spiders," *IEEE Computer*, 36(5), 56-62.

Chen, A., Jiang, H., and Gey, F. (2000). "Combining Multiple Sources for Short Query Translation in Chinese-English Cross-language Information Retrieval," in *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 17-23.

Chen, A. and Gey, F. (2004). "Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decomposing," *Information Retrieval*. Volume 7, Numbers 1-2.

Chen, K.-H., Chen, H.-H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S. H., Kishida, K., Eguchi, K., and Kim, H. (2002). "Overview of CLIR Task at the Third NTCIR Workshop," in *Proceedings of the Third NTCIR Workshop*, Tokyo, Japan, 2002.

Chen, H., Houston, A. L., Sewell, R.R. and Schatz, B.R. (1998). "Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques," *Journal of the American Society for Information Science*, 49(7), 582–603.

Chen, H., Fan, H., Chau, M. and Zeng, D. (2001). "MetaSpider: Meta-Searching and Categorization on the Web," *Journal of the American Society for Information Science and Technology*, 52(13), 1134–1147.

Chen, H., Lally, A., Zhu, B. and Chau, M. (2003a). "HelpfulMed: Intelligent Searching for Medical Information over the Internet," *Journal of the American Society for Information Science and Technology*, 54(7), 683–694.

Chen, H., Fan, H., Chau, M., and Zeng, D. (2003b). "Testing a Cancer Meta Spider," *International Journal of Human-computer Studies*, 59(5), 755-776.

Chen, H. H. and Lee, J. C. (1998), "Proper Name Translation in Cross-language Information Retrieval," In *Proceedings of the 36th conference on Association for Computational Linguistics (ACL)*- Volume 1, 1998 , Montreal, Quebec, Canada.

Cheong, F. C. (1996). *Internet Agents: Spiders, Wanderers, Brokers, and Bots*. 1996. New Riders Publishing, Indianapolis, Indiana, USA.

Chien, L. and Pu, H. (1996). "Important Issues on Chinese Information Retrieval," In *Proceedings of Computational Linguistics and Chinese Language*, 1(1), 205–221.

Chien, L. (1997). "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," In *Proceedings of the 1997 ACM SIGIR*, Philadelphia, PA, USA, 50-58.

Cho, J., Garcia-Molina, H. and Page, L. (1998). "Efficient Crawling through URL Ordering," In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia.

Church, K. and Hanks, P. (1989). "Word Association Norms, Mutual Information, and Lexicography," In *Proceedings of the 27th Annual Meeting of Association for Computational Linguistics*, Vancouver, BC, Canada, 76-83.

CNNIC (2003). "Statistical reports on the Internet development in China, The 12th Survey Report (2003/6/15)," [Online]. Available at <http://www.cnnic.net.cn/download/manual/en-reports/12.pdf>

Davis, M. and Dunning, T. (1995). "A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval," in *Proceedings of the Fourth Text Retrieval Evaluation Conference*, NIST, November 1995.

Davis, M. W. and Ogden, W. C. (1997). "Free Resources and Advanced Alignment for Cross-language Text Retrieval," in *Proceedings of the Sixth Text Retrieval Conference*, NIST, 1997.

Darwish, K, Doermann, D., Jones, R., Oard, D. and Rautiainen, M. (2001). "TREC-10 experiments at Maryland: CLIR and video." In *Proceedings of TREC 2001*. Gaithersburg: NIST.

Eguchi, K., Oyama, K., et al. (2002). "Evaluation Design of Web Retrieval Task in the Third NTCIR Workshop," in *Proceedings of the 11th International World Wide Web Conference (WWW2002)*, Honolulu, Hawaii, USA.

Ellis, D. (1989). "A Behavioral Approach to Information Retrieval Systems Design," *Journal of Documentation*, 45(3).

Fraser, A., Xu, J., and Weischedel, R. (2001). "TREC 2001 Cross-lingual Retrieval at BBN," In *Proceedings of the 10th Text Retrieval Conference*, NIST, 2001.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. (2001). "Improving Query Translation for Cross-language Information Retrieval Using Statistical Models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, 2001, pp. 96-104.

Gey, F. and Chen. A. (2000). "TREC-9 Cross-language Information Retrieval (English-Chinese) Overview," in *Proceedings of the Ninth Text Retrieval Conference*, NIST, 2000.

Global Reach (2003). "Global Internet Statistics," available at: <http://global-reach.biz/globstats/index.php3>

Goto, I., Uratani, N. and Ehara, T. (2001) "Cross-language Information Retrieval of Proper Nouns using Context Information," In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan.

Greene, S., Marchionini, G., Plaisant, C. and Shneiderman, B. (2000). "Previews and Overviews in Digital Libraries: Designing Surrogates to Support Visual Information Seeking," *Journal of the American Society for Information Science*, 51(4), 380–393.

Halpern, J (2003). The Challenges of Intelligent Japanese Searching. Available at: <http://www.cjk.org/cjk/joa/joapaper.htm>

Hearst, M. A. (1994). "Multi-paragraph Segmentation of Expository Text," In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 9-16.

Hearst, M. and Pedersen, J. O. (1996). "Reexamining the Cluster Hypothesis: Scatter/gather on Retrieval Results," In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*, 76–84, New York, ACM Press.

Hovy, E. and Lin, C. Y. (1999). "Automated Text Summarization in SUMMARIST," *Advances in Automatic Text Summarization*, 81–94, MIT Press.

Hull, D. A. and Grefenstette, G. (1996). "Querying across Languages: A Dictionary-based Approach to Multilingual Information Retrieval," in *Proceedings of 19th ACM SIGIR International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996, pp. 49-57.

Jones, G., Sakai, T., Collier, N., Kumano, A., and Sumita, K. (1999). "Exploring the Use of Machine Translation Resources for English-Japanese Cross-language Information Retrieval," in *Proceedings of the Post-Conference Workshop on Machine Translation for Cross Language Information Retrieval at AAMT Machine Translation Summit*, September 1999, pp. 15-22.

Kando, N. (2002). "Evaluation - the Way Ahead: A Case of the NTCIR," in *Proceedings of the ACM SIGIR Workshop on Cross-Language Information Retrieval: A Research Roadmap*, Tampere, Finland, August 2002.

Kawtrakul, A., Deemagarn, A., Thumkanon, C., Khantonthong, N. and McFetridge, Paul (1998), "Backward Transliteration for Thai Document Retrieval," *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAD)*, pp. 128-135.

Knight, K. and Graehl, J. (1997), "Machine Transliteration," *Proceedings of ACL*.

- Kuhlthau, C. C. (1991). "Inside the Search Process: Information Seeking from the User's Perspective," *Journal of the American Society of Information Science*, 42(5), 361-371.
- Kwok, K. (1997). "Comparing Representations in Chinese Information Retrieval," In *Proceedings of ACM SIGIR*, Philadelphia, PA. 34-41.
- Kwok, K.L., (1999). 'English-Chinese Cross-language Retrieval Based on a Translation Package', In *Machine Translation Summit VII workshop on Machine Translation for Cross Language Information Retrieval*, Kent Ridge Digital Laboratories, Singapore, 1999.
- Kwok, K. L. (2000). "Exploiting a Chinese-English Bilingual Wordlist for English-Chinese Cross Language Information Retrieval," in *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 173-179.
- Landauer, T. K. and Littman, M. L. (1991). "A Statistical Method for Language-independent Representation of the Topical Content of Text Segments," in *Proceedings of the 11th International Conference on Expert Systems and Their Applications*, Avignon, France, 1991.
- Lawrence, S. and Giles, C. L. (1998). "Searching the World Wide Web," *Science*, 280, pp. 98-100.
- Lawrence, S. and Giles, C. L. (1999). "Accessibility of Information on the Web," *Nature*, 400, 107-109.
- Leroy, G. and Chen, H. (2001). "Meeting Medical Terminology Needs: The Ontology-enhanced Medical Concept Mapper," *IEEE Transactions on Information Technology in Biomedicine*, vol. 5 (4), 261 - 270.
- Leroy, G. and Chen, H. (2002). "MedTextus: An Ontology-enhanced Medical Portal," In *Proceedings of the Workshop on Information Technology and Systems (WITS)*, Barcelona.
- Lehtokangas., R. and Airio, E. (2002). "Translation via a Pivot Language Challenges Direct Translation in CLIR," In *Proceedings of the ACM SIGIR Workshop on Cross-Language Information Retrieval: A Research Roadmap*, Tampere, Finland, August 2002.
- Lewis, J. R. (1995). "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," *International Journal of Human-Computer Interaction*, 7(1), 57-78

- Lin, W.C., Chen., H.H. (2003). 'Description of NTU Approach to NTCIR3 Multilingual Information Retrieval', In *Proceedings of the Third NTCIR Workshop*.
- Lin, W., Chen, H. H.(2002), "Backward Machine Transliteration by Learning Phonetic Similarity", In *Proceedings of The 6th Workshop on Computational Language Learning (CoNLL2002)*, pp. 139-145, Taipei, Taiwan, Aug. 31-Sep. 1, 2002.
- Lin, W. C., Yang, C. and Chen, H. H., "Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval," In *Proceedings of 4th Workshop of the Cross-Language Evaluation Forum (CLEF2003)*, Trondheim, Norway, Aug. 21-22, 2003.
- Liu, S. (2001). "ECIRS: an English-Chinese Cross-language Information-retrieval System," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2001, vol. 2, pp. 954 –959.
- Liu, Y., Jin, R., Chai, J. Y. (2005). "Cross-language: A maximum coherence model for dictionary-based cross-language information retrieval," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*.
- Lu, Wen-Hsiang, Chien, Lee-Feng, Lee, His-Jian (2004). "Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach," *ACM Transactions on Information Systems*, 22, 1-28, 2004.
- Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000). "Query Term Disambiguation for Web Cross-language Information Retrieval using a Search Engine," in *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 173-179.
- Marchionini, G. and Shneiderman, B. (1988). "Finding Facts vs. Browsing Knowledge in Hypertext Systems," *IEEE Computer*, 21(1), 70–80.
- Marchionini, G. (1995). "Information Seeking in Electronic Environments," Edited by J. Long. 10 vols. Vol. 9, Cambridge Series on Human-Computer Interaction. Cambridge: Cambridge University Press.
- Markó,K., Schulz, S., Medelyan, O. and Hahn, U. (2005). "Cross-language: Bootstrapping dictionaries for cross-language information retrieval," In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*.

- McCallum, A., Nigam, K., Rennie, J. and Seymore, K. (1999) "Building Domain-specific Search Engines with Machine Learning Techniques," In *Proceedings of AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*.
- McDonald, D. and Chen, H. (2002). "Using Sentence Selection Heuristics to Rank Text Segments in TXTRACTOR," In *Proceedings of JCDL'02*, Portland, Oregon. ACM/IEEE-CS, 28–35.
- McLellan, P., Tombros, A., Jose, J., Ounis, I. and Whitehead, M. (2001). "Evaluating Summarisation Technologies: A Task Oriented Approach," In *Proceedings of the 3rd European Conference on Digital Libraries*, Paris, France. 198-214.
- McNamee, P. and Mayfield, J. (2002). "Comparing Cross-language Query Expansion Techniques by Degrading Translation Resources," in *Proceedings of the 25th ACM SIGIR International Conference on Research and Development in Information Retrieval*, Tampere, Finland, August 2002.
- Meho, L.I. & Tibbo, H.R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology*, 54(6), 570-587.
- Meng, W., Liu, K., Yu, C., Wu, W. and Rishe, N. (1999). "Estimating the Usefulness of Search Engines," In *Proceedings of 15th International Conference on Data Engineering (ICDE'99)*, Sydney, Australia, 146-153.
- Meng, W., Wu, Z., Yu, C. And Li, Z. (2001). "A Highly Scalable and Effective Method for Metasearch," *ACM Transactions on Information Systems (TOIS)*, 19(3): 310–335.
- Meng, H., W. K. Lo, B. Chen and K. Tang (2001), "Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval," *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, December 2001.
- Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). "Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, United States, August, 1999, pp. 74-81.
- Nielsen (2003). "Global Internet Population Grows an Average of Four Percent Year-over-year," [Online] Available at:
http://www.nielsen-netratings.com/pr/pr_030220.pdf

Nunamaker, J. F., Chen, M., and Purdin, T. D. M. (1991). "System Development in Information Systems Research," *Journal of Management Information Systems*, 7(3), 89-106.

Oard, D. (1997). "Cross-language Text Retrieval Research in the USA," in *Proceedings of the 3rd ERCIM DELOS Workshop*, Zurich, Switzerland, March 1997.

Oard, D. (2002). "When You Come to a Fork in the Road, Take It: Multiple Futures for CLIR Research," in *Proceedings of the ACM SIGIR Workshop on Cross-language Information Retrieval: A Research Roadmap*, Tampere, Finland, August 2002.

Oard, D. and Wang, J. (2001). "NTCIR-2 ECIR Experiment at Maryland: Comparing Structured Queries and Balanced Translation," in *Proceedings of the Second National Institute of Informatics (NII) Test Collection Information Retrieval (NTCIR) Workshop*, Tokyo, Japan, 2001.

Ogden, W.C., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H., Sharples, N. (1999): "Keizai: An Interactive Cross-language Text Retrieval System," in *Proceedings of Workshop on Machine Translation for Cross Language Information Retrieval*, available at: <http://crl.nmsu.edu/Research/Projects/tipster/ursa/Papers/MTsummit.pdf>

Ong, T. and Chen, H. (1999). "Updatable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: A Linguistic Foundation for Knowledge Management," In *Proceedings of the Second Asian Digital Library Conference*, Taipei, Taiwan, 63 – 84.

Peters, C. (2002). "The Contribution of Evaluation," in *Proceedings of the ACM SIGIR Workshop on Cross-language Information Retrieval: A Research Roadmap*, Tampere, Finland, August 2002.

Porter, M. F. (1980). "An algorithm for suffix stripping", *Program*, 14(3), 130-137.

Qu, Y., Hull, D., Grefenstette, G., Evans, D., Ishikawa, M., Nara, S., Ueda, T., Noda, D., Arita, K., Funakoshi, Y., Matsuda, H.(2005). "Towards effective strategies for monolingual and bilingual information retrieval: Lessons learned from NTCIR-4," *ACM Transactions on Asian Language Information Processing (TALIP)*, Volume 4, Issue 2, pp78-110.

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, 77 (2), 257-286.

Resnik, P. and Smith, N. A. (2003). "The Web as a parallel corpus," *Computational Linguistics*, Volume 29, Issue 3, Pages: 349 - 380.

- Rijsbergen, C. J. van (1979). "Information Retrieval (2nd ed.)," London: Butterworths.
- Rush, J. E., Salvador, R. and Zamora, A. (1964). "Automatic Abstracting and Indexing: Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria," *Journal of the American Society for Information Science*, 22(4), 260—274.
- Sadat, F., Maeda, A., Yoshikawa, M., and Uemura, S. (2002). "A Combined Statistical Query Term Disambiguation in Cross-language Information Retrieval," in *Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02)*, Aix-en-Provence, France, September 2002, pp. 251-255.
- Sakai, T. (2000). "MT-based Japanese-English Cross-language IR Experiments Using the TREC Test Collections," in *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China, 2000, pp. 181-188.
- Salton, G. (1972). "Experiments in Multi-lingual Information Retrieval," *Technical Report TR 72-154*, Computer Science Department, Cornell University.
- Savoy, J. (2005). "Comparative study of monolingual and multilingual search models for use with asian languages," *ACM Transactions on Asian Language Information Processing (TALIP)*, Volume 4, Issue 2, Pages: 163 - 189
- Selberg, E. and Etzioni, O. (1995). "Multi-service Search and Comparison Using the MetaCrawler," In *Proceedings of the 4th World Wide Web Conference*, Boston, Mass, USA, 195 – 208.
- Selberg, E. and Etzioni, O. (1997). "The MetaCrawler Architecture for Resource Aggregation on the Web," *IEEE Expert*, 12(1), 8-14.
- Sheridan, P. and Ballerini, J. P. (1996). "Experiments in Multilingual Information Retrieval Using the SPIDER System," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 1996, pp. 58-65.
- Shortliffe, E. H. (1998). "The Evolution of Health-care Records in the Era of the Internet," *Medinfo*, vol. 9, 8 – 14.
- Spink, A. and Xu, J. (2000). "Selected Results from a Large Study of Web Searching: the Excite Study," *Information Research*, 6(1), available at: <http://InformationR.net/ir/6-1/paper90.html>.

Stalls, B. G. and Knight, K. (1998), "Translating Names and Technical Terms in Arabic Text," In *Proceedings of the COLING/ACL workshop on Computational Approaches to Semitic Languages*.

Thompson, P. & Dozier, C. (1997). Name searching and information retrieval. In *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island.

Tolle, K. and Chen, H. (2000). "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *Journal of the American Society for Information Science*, 51, 352-370.

Virga, P. and Khudanpur S. (2003), "Transliteration of Proper Names in Cross-Lingual Information Retrieval," In *Proceedings of SIGIR 2003*, Toronto, Canada.

Viterbi, A. J. (1967), "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Transactions on Information Processing*, 13:260-269.

Voorhees, E. and Harman, D. (1997). "Overview of the Sixth Text REtrieval Conference," In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, USA, 1-24.

Wan, S. and Verspoor, C. M. (1998), "Automatic English-Chinese name transliteration for development of multilingual resources," In *Proceedings of COLING-ACL'98*, Montreal, Canada.

Wang, J. H., Teng, J. W., Cheng, P. J., Lu, W. H. and Chien, L. F. (2004). "Translating Unknown Cross-lingual Queries in Digital Libraries Using a Web-based Approach," in *Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries*, Tucson, Arizona, June 2004.

Xu, J. and Croft, B. (1996). "Querying Expansion using Local and Global Document Analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 1996, pp. 4-11.

Xu, J. and Weischedel, R. (2000). "TREC-9 Cross-lingual Retrieval at BBN," in *Proceedings of the 9th Text Retrieval Conference*, NIST, 2000.

Yang, C., Li, K. W. (2003). "Automatic construction of English/Chinese parallel corpora," *Journal of the American Society for Information Science and Technology*. Volume 54, Issue 8 , Pages 730 – 742

Zamir, O. and Etzioni, O. (1999). "Grouper: A Dynamic Clustering Interface to Web Search Result," In *Proceedings of the Eighth World Wide Web Conference*, Toronto, 1361–1374.