Multilingual Web Retrieval: An Experiment in English–Chinese Business Intelligence

Jialun Qin and Yilu Zhou
Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721.
E-mail: {qin, yilu}@email.arizona.edu

Michael Chau
School of Business, The University of Hong Kong, Hong Kong, People's Republic of China.
E-mail: mchau@business.hku.hk

Hsinchun Chen
Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721.
E-mail: hchen@eller.arizona.edu

As increasing numbers of non-English resources have become available on the Web, the interesting and important issue of how Web users can retrieve documents in different languages has arisen. Cross-language information retrieval (CLIR), the study of retrieving information in one language by queries expressed in another language, is a promising approach to the problem. Cross-language information retrieval has attracted much attention in recent years. Most research systems have achieved satisfactory performance on standard Text REtrieval Conference (TREC) collections such as news articles, but CLIR techniques have not been widely studied and evaluated for applications such as Web portals. In this article, the authors present their 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 posttranslation query expansion. The portal was evaluated by domain experts, using a set of queries in both English and Chinese. The experimental results showed that co-occurrence-based phrasal translation achieved a 74.6% improvement in precision over simple word-by-word translation. When used together, pre- and posttranslation query expansion improved the performance slightly, achieving a 78.0% improvement over the baseline word-by-word translation approach. In general, applying CLIR techniques in Web applications shows promise.

Introduction

Rapid growth of the Internet has led to a tremendous number of multilingual resources on the Web. There are Web pages in almost every popular non-English language including various European, Asian, and Middle East languages. The number of non-English speakers constitutes 64.2% of the world's online population, which far exceeds the English online population (Global Reach, 2003). 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. Cross-language information retrieval 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 Text REtrieval Conference (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

Accepted March 9, 2005

 ● Published online 1 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20329

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 preprocessing. 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, a query on the Internet has an average length of 2.21 (Spink & Xu, 2000), which is considered a short query in information retrieval. In this study, we posed the following research questions: (a) Can CLIR techniques achieve satisfactory performance for retrieving Web documents that are much "nosier" than traditional text collections?; and (b) Can we combine existing CLIR techniques to build a multilingual Web portal with both satisfactory effectiveness and efficiency?

To address these questions, our 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 article is structured as follows. In the next section, we review related research, including three fundamental approaches to CLIR, translation ambiguity problems, and query expansion techniques. We also discuss problems in using existing CLIR techniques in Web applications and present our research questions. In the following sections, we propose our Web-based multilingual retrieval prototype and discuss the system architecture and implementation details of a prototype English–Chinese Web portal called ECBizPort. We also show an example of how a search query will be translated and expanded by our system, using different CLIR techniques. We report the set-up and results of an experiment designed to evaluate the performance of the prototype. In the last section, we conclude our work and suggest some future directions.

Literature Review

In this section, we review CLIR techniques that are related to our 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 the next sections, we review three major query translation approaches, namely a machine-translation approach, a corpus-based approach, and a dictionary-based approach. In addition, we report on several translation disambiguation techniques that have been used to reduce errors introduced during query translation. We also review applications of CLIR in Web-based systems.

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.

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. Aljlayl, Frieder, and Grossman (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 Asian 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, Sakai, Collier, Kumano, & Sumita, 1999). This also affects the effectiveness of MT-based approach.

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 information retrieval (IR) performance. Sheridan and Ballerini (1996) applied thesaurus-based query expansion techniques on a comparable Italian–English collection. Recently, the BBN Technology 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 Asian language pairs such as e 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, Sionard, Isabelle, and Durand (1999) investigated the possibility of automatically gathering parallel text from the Web. Their Web-mining approach showed the feasibility of using the Web as potential corpus.

Dictionary-based approach. In a dictionary-based approach, queries are translated by looking up terms in a

DOI: 10.1002/asi

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 and Croft (1996, 1997) investigated dictionary-based Spanish–English CLIR and reported that using both pre- and posttranslation 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, Jiang, and Gey (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 a dictionary; there have been many studies of these "out-of-vocabulary" words (Lu, Chien, & Lee, 2002). Researchers have also identified several challenges to this approach: (a) multiple definitions of a word could introduce noise into the translated query (a.k.a. ambiguity); (b) failure to translate technical and new terminology, which is often not found in general dictionaries; and (c) failure to translate multiterm concepts as phrases (Ballesteros & Croft, 1996).

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, we briefly review three of them: phrasal translation, co-occurrence analysis, and query expansion.

Phrasal translation. Phrasal translation techniques are often used to identify multiword 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 (Ballesteros & Croft, 1998; Davis & Ogden, 1997). 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.

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; Gao et al., 2001; Maeda, Sadat, Yoshikawa, & Uemura, 2000; Sadat, Maeda, Yoshikawa, & Uemura, 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.

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 & Fraenkel, 1977; Croft & Harper, 1979). Other query expansion methods also may use both local and global information, such as the local context analysis (LCA) method (Xu & Croft, 1996).

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

Cross-Language Information Retrieval 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.

DOI: 10.1002/asi

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[1] 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[2] 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.

Summary

The Web has become a major information source worldwide for people in any knowledge field. 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 problems in indexing. Different from traditional text documents, a 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 our review, we 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, we posed the following research questions:

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

In the remainder of this article, we present our work in studying these two questions.

Proposed Approach to Multilingual Web Retrieval

Aiming to apply an integrated set of CLIR techniques to the Web environment, we propose an architecture for multilingual Web portal development. The proposed system architecture is shown in Figure 1.

Our system architecture consists of five major components: (a) Web spider and indexer, (b) pretranslation query expansion, (c) query translation, (d) posttranslation query expansion, and (e) document retrieval. In the following sections, we describe each component in detail.

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.

[1]http://www.cs.nmsu.edu/~sliu/cgi-bin/ec_search/index_concord.pl (Accessed September 12, 2004.)

[2]http://www.arabvista.com/ (Accessed September 12, 2004.)

DOI: 10.1002/asi

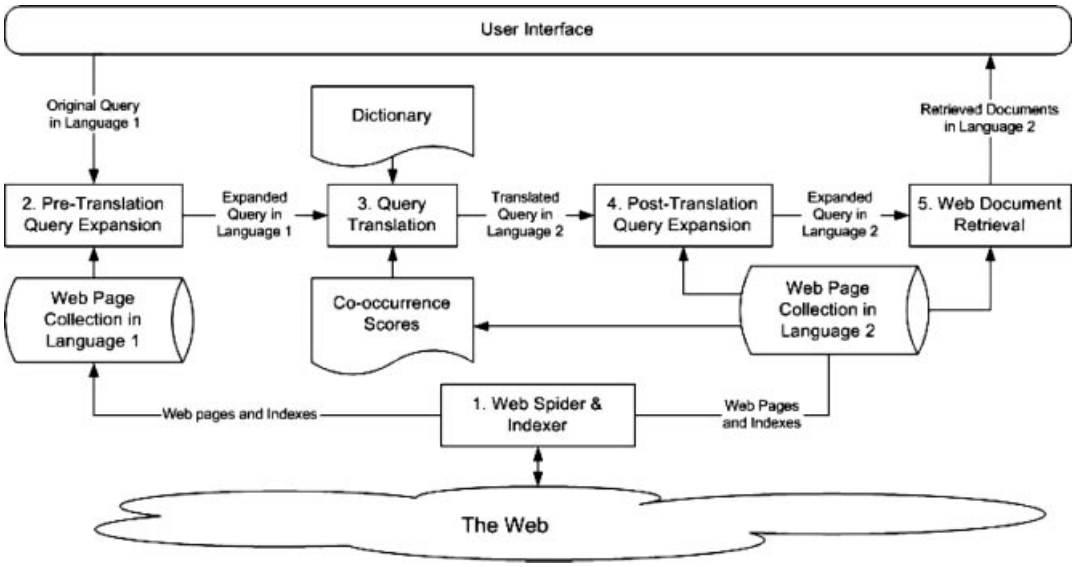

FIG. 1. System architecture.

There are high requirements for the 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, 2003). 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, we proposed a collection-building method that extends the capability of traditional focused crawling by meta-searching multiple large search engines. The process of our method can be described as follows: Similar to traditional focused crawlers, we start our "probing" of the Web with a set of starting URLs and fetch relevant Web pages. At the same time, new 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).

Pretranslation Query Expansion

In our Web portal we undertook pretranslation query expansion to expand users' queries in the original language. As discussed earlier, there are two common ways to perform pretranslation query expansion, namely local feedback and local context analysis. We chose to use the local feedback method because of its higher efficiency, an important factor for Web applications. Our approach followed the method reported by Ballesteros and Croft (1997). The pretranslation 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 $tf \cdot idf$ scores calculated. $tf \cdot idf$ 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).

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 the 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. We proposed to use a

DOI: 10.1002/asi

dictionary-based approach combined with phrasal translation and co-occurrence analysis for translation disambiguation.

In the dictionary lookup process, we first conducted *maximum phrase matching* on English queries. The sequence of English words that matches a dictionary entry was identified as a *phrase*. If a 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 a 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. Our 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 our proposed method from theirs is that they calculated the co-occurrence score "on the fly" which may greatly lower system efficiency; we calculated co-occurrence scores in advance to avoid affecting run-time efficiency, which is extremely important for Web applications.

Posttranslation Query Expansion

The posttranslation query expansion component is responsible for expanding the query in the target language. Similar to pretranslation expansion, we 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 $tf \cdot idf$ scores. The top terms are then combined with the translated query and reweighed to build the final query.

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, Cho, Garcia-Molina, Paepke, & Raghavan, 2001).

ECBizPort: An English–Chinese Web Portal for Business Intelligence

In this section, we report our 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. We 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. We will also discuss some important issues in multilingual Web retrieval system development.

Figure 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 or she may use the terms to refine the query manually.

Domain Selection

We decided to implement an English–Chinese Web portal for IT business intelligence because we 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.

Web Spider and Indexer

The AI Lab SpidersRUs toolkit,[3] a digital library development tool developed by our 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 that use local search algorithms in Web searching, we used a metasearch enhanced focused crawling approach (Qin, Zhou, & Chau, 2004) to build the ECBizPort collections. Similar to traditional focused crawlers, our crawler starts with a set of starting URLs and fetches relevant pages back based on the content- and link-based analysis results.

[3]http://ai.bpa.arizona.edu/spidersrus/ (Accessed September 12, 2004.)

DOI: 10.1002/asi

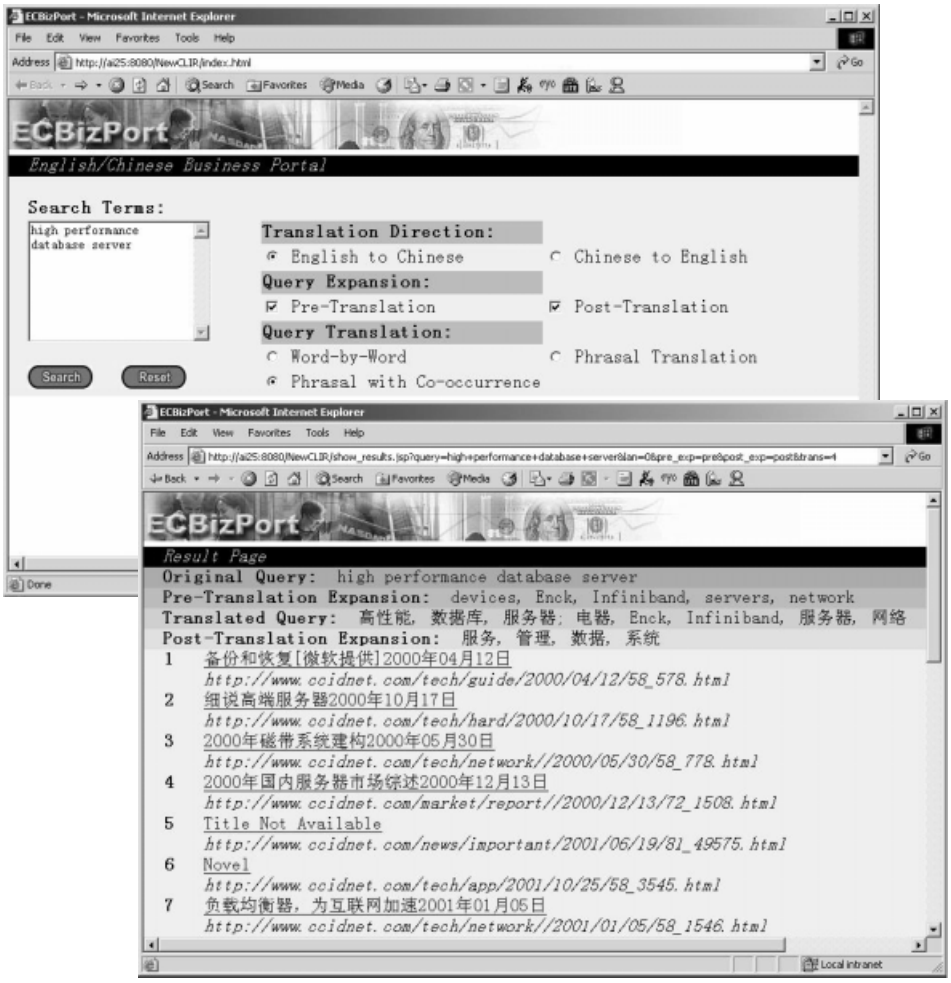

FIG. 2. Sample screenshots of ECBizPort.

Outgoing links in the relevant pages are extracted and put into the URL queue. At the same time, a metasearching 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, we used 20 IT/Business-related starting URLs suggested by our domain expert, such as http://www.zgcsc.com/ (formal name: 第三媒体) and http://www.csdn.net/ (formal name: CSDN). A domain lexicon of 100 typical IT/business-related kewords were created. During the spider process, these keywords identified by the expert were sent in Chinese to five major search engines, Google, Yahoo China, Sina, Sohu, and Baidu.[4] 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 6 hours collecting 100,000 Chinese Web pages. Similar to the Chinese collection, the English collection consisted of 100,000 pages and was built in about 5 hours with 16 expert-identified starting URLs, a 100-keyword lexicon and 5 major general search engines, namely Google, Yahoo, AltaVista, Infoseek and Hotbot.[5] By using this metasearch enhanced spider algorithm, we 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.

Query Translation

English–Chinese dictionary translation. Query term translations were performed using the LDC (Linguistic Data Consortium) English–Chinese bilingual wordlists as dictionaries.[6] The LDC wordlists include two specific

[4]http://www.google.com/; http://cn.yahoo.com/; http://www.sina.com.cn/; http://www.sohu.com/; http://www.baidu.com/ (All accessed December 15, 2005.)

[5]http://www.yahoo.com/; http://www.altavista.com/; http://infoseek.go.com/; http://www.hotbot.com/ (All accessed September 12, 2004.)

[6]http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm (Accessed September 12, 2004.)

DOI: 10.1002/asi

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 our 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	*it* /
情报技术	/ (military) Intelligence Technology / *IT* /
信息技术	/ Information Technology / *IT* /

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" because each was the only translation for the other. Maximum phrase matching was also incorporated in our system by ranking Chinese translations containing more continuous key words higher than those containing discontinuous key words. For example, for the English term "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 separate terms or a single word and a two-word phrase.

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

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 our general indexing process. This led to loss of semantic meaning, because 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 in the IT business domain, which involves a lot of new terminology.

To address that problem, we decided to extract key phrases from our collection to build our own lexicon. Arizona Noun Phraser (AZNP), developed by our research group, was used to extract phrases from the English collection (Tolle & Chen, 2000). The 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 our 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 posttranslation query expansion.

The local feedback method was implemented for both pre- and posttranslation query expansion in our system. For pretranslation expansion, the top-10 English Web pages were retrieved by the original English query, and the five English terms/phrases with the highest $tf \cdot idf$ scores were added to the original query. In posttranslation expansion, the top-10 Chinese Web pages were retrieved, and the top-5 Chinese terms/phrases were added to the translated Chinese query. In both pre- and posttranslation expansion, the terms in the expanded query were reweighed using the Rocchio formula (Xu & Croft, 1996).

Document Retrieval

The document retrieval component was supported by the AI Lab SpidersRUs toolkit and the design was relatively

DOI: 10.1002/asi

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.

An Example of Query Translation and Expansion

In this section, we 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." To this end, the user 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 fit well into the context of the original search query. To further improve the retrieval performance, the user also chose to perform pre- and posttranslation query expansion. The results are shown below:

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

The pretranslation 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 posttranslation 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

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

System Evaluation

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

Cross-Language Information Retrieval Evaluation Methodologies

Cross-language information retrieval 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.

DOI: 10.1002/asi

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 prejudged by human experts to be relevant to the original query. Because it would have been unrealistic to expect every document in the collection to be judged, precision was often of more interest than recall, and was usually reported at low-recall levels. In other words, precision rate was 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.

Cross-language information retrieval always yields precision loss compared to traditional monolingual information retrieval; therefore, 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).

Experiment Design and Measure

In general, we followed the TREC evaluation process in our experiment design (Voorhees, 1998). However, because our study used Web pages instead of standard collections, no established relevance judgment was available for precision and recall. Therefore, we decided to create relevance judgment by recruiting human experts and adopting the precision at top-n retrieved documents ($n = 10$ in our experiments) as our primary performance measure. We were particularly interested in how well these techniques would work for Web content in a business intelligence portal; hence, we 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 seven 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 cooccurrence analysis (Ph-Co), Ph-Co with pretranslation expansion, Ph-Co with posttranslation expansion, and Ph-Co with both pre- and posttranslation expansion.

The experts individually submitted each query to the system under the different settings. It was not practical for the experts to read all the 100,000 Web pages in the collection; therefore, we emphasized precision only for the top-10 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: (a) monolingual information retrieval (the best-case scenario), and (b) 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.

Experimental Results and Discussion

Table 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 1. Precision and time.

Method	Average top-10 precision	Performance compared with Monolingual	Improvement compared with WBW	Time used (s)
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

Note. 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 pretranslation query expansion; Phr-Co-Post = phrasal translation with co-occurrence disambiguation and posttranslation query expansion; Phr-Co-Pre-Post = phrasal translation with co-occurrence disambiguation and both pre- and posttranslation query expansion.

DOI: 10.1002/asi

TABLE 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

Note. 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 pretranslation query expansion; Phr-Co-Post = phrasal translation with co-occurrence disambiguation and posttranslation query expansion; Phr-Co-Pre-Post = phrasal translation with co-occurrence disambiguation and both pre- and posttranslation query expansion.

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

Precision. In Table 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 60–90%. The results demonstrated that CLIR techniques assisted users in searching for documents in a different language in a noisy Web portal setting.

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 2.

As shown in Table 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 we 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, pretranslation expansion did not further improve performance. Posttranslation expansion, used alone or combined with pretranslation expansion, slightly improved performance, but the improvement was not as significant as we had expected. We suspect the noisy factor of Web pages might have 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. We 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 untranslated and did not improve performance.

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, we 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, we analyzed the processes of different CLIR techniques and made a baseline estimation of their effect on system efficiency. Our results showed that phrasal translation with co-occurrence disambiguation took 3.5 times longer than monolingual translation. When both pre- and posttranslation expansions were used, the retrieval time increased to 10 times that of monolingual retrieval, which reached 70s. It should be noted that our 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 74–78% 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 25–71 s. This result was not as good as had been expected, but we believe it could be greatly improved if the system were to be run on a high-performance server.

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 article, we have presented our 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 our Web portal following TREC evaluation procedures (Voorhees, 1998). Our results showed that our 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

DOI: 10.1002/asi

was reasonably efficient on a PC and should achieve better efficiency on a more powerful machine. In sum, our study demonstrated the feasibility of applying CLIR techniques to Web applications and the experimental results are encouraging.

We plan to expand our research in several directions. First, we plan to integrate more CLIR techniques into the Web portal to make it more robust. Our techniques in ECBizPort are our first attempts to study Web-based CLIR. We are also investigating how to improve the speed of the system to achieve faster response time, which is necessary for a Web portal. In addition, we 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 we believe that different domains might have different effects on the performance of CLIR techniques, we are interested in testing our approach in other domains, such as medicine. Finally, we plan to extend our system evaluation and user study. On the system performance aspect, we are undertaking a more systematic comparison on the performance of Web-based CLIR systems versus a traditional CLIR system. A user study on the usability and information accessibility of interactive Web-based systems will be conducted in the future.

Acknowledgments

This project was supported, in part, by an NSF Digital Library Initiative-2 grant, to H. Chen (IIS-9817473). We would also like to thank the AI Lab team members who developed the AI Lab SpidersRUs toolkit, the Mutual Information software, and the AZ Noun Phraser. Finally, we also want to thank the domain experts who took part in the evaluation study.

References

Aljlayl, M., Frieder, O., & Grossman, D. (2002). On bidirectional English-Arabic search. Journal of the American Society for Information Science and Technology, 53(13), 1139–1151.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. ACM Transactions on Internet Technology, 1(1), 2–43.

Attar, R., & Fraenkel, A.S. (1977). Local feedback in full-text retrieval systems. Journal of the Association for Computing Machinery, 24(3), 397–417.

Ballesteros, L., & Croft, B. (1996, September). Dictionary methods for cross-lingual information retrieval. In Paper presented at the 7th DEXA Conference on Database and Expert Systems Applications, Zurich, Switzerland.

Ballesteros, L., & Croft, B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In N. Belkin, D. Nara Simhalu, & P. Willett (Eds.), Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 84–91). New York: ACM.

Ballesteros, L., & Croft, B. (1998). Resolving ambiguity for cross-language retrieval. In W.B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 64–71). New York: ACM.

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002, September). Focused crawls, tunneling, and digital libraries. Paper presented at the European Conference on Digital Libraries, Rome, Italy.

Capstick, J., Diagne, A.K., Erbach, G., Uszkoreit, H., Cagno, F., Gadaleta, G., et al. (1998, August). MULINEX: Multilingual Web Search and Navigation. Paper presented at the Conference on Natural Language Processing and Industrial Applications, Moncton, Canada.

Chau, M., & Chen, H. (2003). Comparison of three vertical search spiders. IEEE Computer, 36(5), 56–62.

Chen, A., Jiang, H., & Gey, F. (2000, September). Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. Paper presented at the Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, China.

Chen, K.-H., Chen, H.-H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S.H., et al. (2002, October). Overview of CLIR Task at the Third NTCIR Workshop. Paper presented at the Third NTCIR Workshop, Tokyo, Japan.

Cheong, F.C. (1996). Internet agents: Spiders, wanderers, brokers, and bots. IN: New Riders Publishing.

Croft, W.B., & Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285–295.

Davis, M., & Dunning, T. (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval. In D.K. Harman (Ed.), Proceedings of the Fourth Text Retrieval Evaluation Conference. Gaithersburg, MD: National Institute of Standards and Technology.

Davis, M.W., & Ogden, W.C. (1997). Free resources and advanced alignment for cross-language text retrieval. In Proceedings of the Sixth Text Retrieval Conference.

Eguchi, K., Oyama, K., et al. (2002, May). Evaluation design of Web retrieval task in the Third NTCIR Workshop. Paper presented at the 11th International World Wide Web Conference, Honolulu, Hawaii. Retrieved September 12, 2004, from http://www.2002.org/CDROM/poster/22/

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 96–104). New York: ACM.

Gey, F., & Chen. A. (2000). TREC-9 cross-language information retrieval (English–Chinese) overview. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the Ninth Text Retrieval Conference (pp. 15–24). Gaithersburg, MD: National Institute of Standards and Technology.

Global Reach. (2003). Global internet statistics. Retrieved October 24, 2003, from http://global-reach.biz/globstats/index.php3

Hull, D.A., & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of 19th ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 49–57). New York: ACM.

Jones, G., Sakai, T., Collier, N., Kumano, A., & Sumita, K. (1999, September). Exploring the use of machine translation resources for English–Japanese cross-language information retrieval. Paper presented at the Post-Conference Workshop on Machine Translation for Cross Language Information Retrieval at AAMT Machine Translation Summit (pp. 181–188).

Kando, N. (2002). Evaluation—the way ahead: A case of the NTCIR. In Proceedings of the 25th ACM SIGIR Workshop on Cross-Language Information Retrieval: A Research Roadmap (pp. 72–77). New York: ACM.

Kwok, K.L. (2000, September). Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. Paper presented at the Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, China.

Landauer, T.K., & Littman, M.L. (1991, May). A statistical method for language-independent representation of the topical content of text segments. Paper presented at the 11th International Conference on Expert Systems and Their Applications, Avignon, France.

DOI: 10.1002/asi

Lawrence, S., & Giles, C.L. (1998). Searching the world wide web. Science, 280, 98–100.

Liu, S. (2001). ECIRS: an English-Chinese Cross-language Information-retrieval System. In A. El Kamel, K. Mellouli, & P. Borne (Eds.), Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, (Vol. 2, pp. 954–959). Piscataway, NJ: IEEE.

Lu, W.-H., Chien, L.-F., & Lee, H.-J. (2002). Translation of Web queries using anchor text mining. ACM Transactions on Asian Language Information Processing, 2(1), 159–172.

Lu, W.-H., Chien, L.-F., Lee, H.-J. (2004). Anchor text mining for translation of web queries: A transitive translation approach. ACM Transactions on Information Systems, 22, 1–28.

Maeda, A., Sadat, F., Yoshikawa, M., & Uemura, S. (2000). Query term disambiguation for web cross-language information retrieval using a search engine. Paper presented at the Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, China.

McNamee, P., & Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. In R. Baeza-Yates, N. Fuhr, & Y. Maarek (Eds.), Proceedings of the 25th ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 159–166). New York: ACM.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In M.A. Hearst, F. Gey, & R. Tong (Eds.), Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 74–81). New York: ACM.

Oard, D. (1997, March). Cross-language text retrieval research in the USA. Paper presented at the 3rd ERCIM DELOS Workshop, Zurich, Switzerland.

Oard, D. (2002). When you come to a fork in the road, take it: Multiple futures for CLIR research. Cross-language information retrieval: A research roadmap. In R. Baeza-Yates, N. Fuhr, & Y. Maarek (Eds.), Proceedings of the 25th ACM SIGIR International Conference on Research and Development in Information Retrieval (pp. 5–7). New York: ACM.

Oard, D., & Wang, J. (2001). NTCIR-2 ECIR Experiment at Maryland: Comparing structured queries and balanced translation. In J. Adachi & N. Kando (Eds.), Proceedings of the Second National Institute of Informatics (NII) Test Collection Information Retrieval (NTCIR) Workshop. Tokyo, Japan: NII.

Ogden, W.C., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H., et al. (1999): Keizai: An interactive cross-language text retrieval system. In S. Ananiadou, Y. Hayashi, C. Jacquemin, M.K. Leong, & H.-H. Chen (Eds.), Proceedings of Workshop on Machine Translation for Cross Language Information Retrieval. Retrieved May 6, 2003, from http:// crl.nmsu. edu/Research/Projects/tipster/ursa/Papers/MTsummit.pdf

Ong, T.-H. and Chen, H. (1999). Updateable PAT-tree approach to Chinese key phrase extraction using mutual information: A linguistic foundation for knowledge management. Paper presented at the 2nd Asian Digital Library Conference, Taipei, Taiwan.

Peters, C. (2002). The contribution of evaluation. In R. Baeza-Yates, N. Fuhr, & Y. Maarek (Eds.), Proceedings of the ACM SIGIR Workshop on Cross-language Information Retrieval: A Research Roadmap (pp. 16–19). New York: ACM.

Qin, J., Zhou, Y., & Chau, M. (2004). Building domain-specific Web collections for scientific digital libraries: A meta-search enhanced focused crawling method. In H. Chen & M. Christel et al. (Eds.), Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'04) (pp. 135–141).

Sadat, F., Maeda, A., Yoshikawa, M., & Uemura, S. (2002, September). A combined statistical query term disambiguation in cross-language information retrieval. In Paper presented at the 13th International Workshop on Database and Expert Systems Applications (DEXA'02), Aix-en-Provence, France.

Sakai, T. (2000, October). MT-based Japanese-English cross-language IR experiments using the TREC test Collections. Paper presented at the Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, China.

Salton, G. (1972). Experiments in multi-lingual information retrieval. (Technical Report TR 72-154). Ithaca, NY: Computer Science Department, Cornell University.

Sheridan, P., & Ballerini, J.P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 58–65). New York: ACM.

Spink, A., & Xu, J. (2000). Selected results from a large study of web searching: The excite study. Information Research, 6(1). Retrieved October 24, 2003, from http://InformationR.net/ir/6-1/paper90.html

Tolle, K.M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, 51(4), 352–370.

Voorhees, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 315–323).

Wang, J.H., Teng, J.W., Cheng, P.J., Lu, W.H., & Chien, L.F. (2004). Translating unknown cross-lingual queries in digital libraries using a web-based approach. In H. Chen, H.D. Wactlar, C.-C. Chen, E.-P. Lim, & M.G. Christel (Eds.), Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries (pp. 4–10). New York: ACM.

Xu, J., & Croft, B. (1996). Querying expansion using local and global document analysis. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 4–11). New York: ACM.

Xu, J., & Weischedel, R. (2000). TREC-9 Cross-lingual retrieval at BBN. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 9th Text Retrieval Conference (pp. 106–116). Gaithersburg, MD: National Institutes of Standards and Technology.

DOI: 10.1002/asi