Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal

Yilu Zhou, Jialun Qin, Hsinchun Chen, Jay F. Nunamaker
Department of Management Information Systems
The University of Arizona
Tucson, AZ 85721
yiluz@eller.arizona.edu, qin@u.arizona.edu, hchen@eller.arizona.edu,
jnunamaker@cmi.arizona.edu

Abstract

The amount of non-English information on the Web has proliferated so rapidly in recent years that it often is difficult for a user to retrieve documents in an unfamiliar language. In this study, we report the design and evaluation of a multilingual Web portal in the business domain in English, Chinese, Japanese, Spanish, and German. Web pages relevant to the domain were collected. Search queries were translated using bilingual dictionaries, while phrasal translation and co-occurrence analysis were used for query translation disambiguation. Pivot translations were also used for language-pairs where bilingual dictionaries were not available. A user evaluation study showed that on average, multilingual performance achieved 72.99% of monolingual performance. In evaluating pivot translation, we found that it achieved 40% performance of monolingual retrieval, which was not as good as direct translation. Overall, our results are encouraging and show promise of successful application of MLIR techniques to Web retrieval.

1. Introduction

The World Wide Web has become a major channel for information service. There are Web pages in almost every popular language including various European, Asian, and Middle East languages. While approximately 70% of Web content is in English, the number of native English speakers constitutes only 36.5% of the world's online population [12]. The broad diversity of the Web presents a substantial research challenge in the field of information retrieval. 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. However, language boundaries prevent information sharing among countries and communities. Multilingual Information Retrieval (MLIR), the study of responding to a query by searching for documents in more than one language [6], is a promising approach to the multilingual problem.

Most MLIR research has used standard TREC collections, predominately news articles, as their training and testing set, but little research has investigated Web-based MLIR systems. Several researchers [14, 20] have suggested that operational applications will be the next step in MLIR research. While MLIR techniques have been shown to be promising, it remains unclear how well these techniques would apply to Web-based content. First, while traditional MLIR only addresses effectiveness, measured by recall and precision, a Web-based MLIR system also considers efficiency and interaction. Second, Web pages are comparatively unstructured and are very diverse in terms of document content and document format (such as HTML, PDF, PHP or ASP). Third, a Web-based MLIR system requires a robust spider algorithm to collect multilingual documents from the Internet. All these aspects add difficulties to multilingual Web retrieval.

The paper is structured as follows. Section 2 reviews related research, including fundamental approaches to MLIR: addressing translation ambiguity and linguistic-resource problems and designing Web-based MLIR system issues. In Section 3, we discuss problems of using existing MLIR techniques in Web applications and present our research questions. In Section 4, we propose our Web-based multilingual retrieval system design. Section 5 discusses the system architecture and implementation details of a prototype Multilingual Web in the business domain. Section 6 reports the setup and results of an experiment designed to evaluate the performance of the prototype. Finally, in Section 7 we conclude our work and suggest some future directions.

2. Related Work: MLIR on the Web

2.1 Query Translation Approaches

In multilingual information retrieval, two fundamental approaches are often considered: query translation or document translation. Query translation translates queries

0-7695-2268-8/05/$20.00 (C) 2005 IEEE

into all target document languages, and monolingual retrieval is carried out separately for each document language. Most reported research in the field has applied query translation approach. There are three main approaches in CLIR and MLIR: using machine translation (MT), a parallel corpus, or a bilingual dictionary.

Machine translation-based (MT-based) approach uses existing machine translation techniques to provide automatic translation of queries. Sakai [24] used MT Avenue, a free web-based Japanese-English translation service, and achieved reasonable effectiveness with the aid of pseudo-relevance feedback. 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. Also, typical search queries lack the contextual information which is necessary for MT to correctly perform word sense disambiguation [24].

A corpus-based approach analyzes large document collections (parallel or comparable corpus) to construct a statistical translation model. Although the approach is promising, the performance relied largely on the quality of the corpus. Davis and Dunning [8] applied evolutionary programming on a parallel Spanish-English collection, and reported 75% of monolingual IR performance. Sheridan and Ballerini [25] applied thesaurus-based query expansion techniques on a comparable Italian-English collection. 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, parallel corpus is 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. This is the most popular approach because of its simplicity and the wide availability of machine-readable dictionaries. Ballesteros and Croft [1, 2] investigated dictionary-based Spanish-English CLIR. Chen et al. [5] focused on short query translation by combining multiple sources in English-Chinese. Bilingual machine-readable dictionary (MRD) is more widely available than parallel corpora. However, there are several challenges to this approach: multiple definitions of a word which could introduce noise into the translated query (a.k.a. ambiguity); failure to translate technical/new terminology; and failure to translate multi-term concepts as phrases [1].

2.2 Reducing Translation Ambiguities and Errors

Phrasal translation techniques are often used to identify multi-word concepts in the query and translate them as phrases. Hull and Grefenstette [13] showed that effectiveness of CLIR is significantly improved when phrases are manually translated. The effectiveness also can be improved by using phrase information in machine-readable dictionaries [3, 9]. The major challenge in using phrasal translation is that many phrases are not covered by dictionaries.

Co-occurrence statistics also has been used in selecting the best translation(s) among the candidates. This technique assumes that the correct translations of query terms tend to co-occur more frequently than the incorrect translations do in documents written in the target language. Co-occurrence analysis has been successfully used in many previous studies to resolve translation ambiguity [3, 11, 17, 23] and some improved co-occurrence analysis methods have been suggested [19]. Previous studies using co-occurrence analysis disambiguation have reported much improvement in MLIR performance. However, the heavy computational and storage requirements of co-occurrence analysis have limited its use in practical retrieval systems where efficiency is a major concern. The corpus used for co-occurrence training needs to be highly relevant to the search domain. Current MLIR training data are mostly news articles from previous TREC collections. Such corpora may not be suitable for a Web-based MLIR system, where new terminologies frequently emerge.

2.3 Scarce Resource Problem in MLIR

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. Of the linguistic resources, bilingual dictionaries between major languages are more prevalent than parallel texts of sufficient in a large domain. However, even relatively widely available bilingual dictionaries, they are available 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 [18]. Several efforts have been made to investigate the scarce resources problem in MLIR.

Obtain 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. Although the Web has become a promising resource for MLIR research, the diversity of Web pages makes significant work necessary for construction of reasonable MILIR resources from Web collections.

Combine Available Resources. Previous research has shown that by combining multiple resources, MLIR achieves higher precision than that of any single resource. Kwok [15] used a machine translation system and a small bilingual wordlist and found the bilingual wordlist to complement the machine translation. Nie et al. [19] combined a parallel corpus mined from the Web with a bilingual wordlist as translation resource. Different MLIR resources often complement each other and could improve MLIR system performance.

Pivot Language Translation. For non-English language pairs, resources are even more difficult to obtain than language pairs involving English. In these cases, MLIR cannot be directly performed across non-English languages. Usually, a language with more available resources is used as an intermediate language [16]. This special form of MLIR is known as pivot language approach (or trans-lingual approach). For example, to perform a Chinese-to-Japanese query translation, Chinese-to-English translation is carried out first and followed by English- to-Japanese translation (Chinese -> English -> Japanese). In this case, English serves as the intermediate or "pivot" language. For language pairs with scant translation resources, pivot translation is a viable approach.

2.4 MLIR for Web applications

As discussed earlier, 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), while 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 1G data), while Web-based MLIR usually deals with larger collections (more than several Giga data). Taking the document collection in TREC 2002 as an example, the collection for the Cross-language Track is 869 Megabytes, and the Web Track in TREC contains 18.1G 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 formats of documents, including HTML, ASP, PDF, PS, Word, and etc.

Efficiency: Traditional MLIR usually focuses on effectiveness of the performance, and efficiency is usually ignored. However, efficiency is important for end users in Web retrieval scenario.

Query length: Traditional MLIR usually uses long query texts, a sentence description or sometimes a narrative paragraph. Queries on Internet are much shorter and have an average length of 2.21 words [26]. Short queries offer less context information for translation disambiguation, and thus are more challenging in MLIR research.

Several commercial 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 78 languages as well as provide machine translation services for certain languages. However, from the user's perspective these search engines are essentially a collection of monolingual search engines. None of the big search engines have incorporated MLIR technology as a service. Although not widely studied, some Web-based MLIR systems have been made available. Some of them are Keizai, TwentyOne and MULINEX. Keizai, developed at the New Mexico State University, is an interactive Web-based MLIR system that accepts English queries and returns Japanese and Korean documents [21]. TwentyOne is developed in Irion Technologies in The Netherlands and supports 6 European languages. MULINEX is a comparatively more mature multilingual Web search and navigation tool, developed in DFKI Language Technology Lab [4] for English, French and German. However, the major problem for most of these systems is that no systematic evaluations are available, leaving their effectiveness uncertain.

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. Based on our review, we 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, we posed the following research questions:

1. How can we develop a generic approach for multilingual Web retrieval system that incorporates European and Asian languages?
2. How can we mine the Web for useful linguistic resources to improve multilingual Web retrieval performance?
3. Does pivot language translation achieve reasonable effectiveness in multilingual Web retrieval?

The remainder of the paper presents our work in studying these questions.

4. Proposed Approach to Multilingual Retrieval on the Web

In this section, we report our experience in implementing a multilingual Web retrieval system using the proposed MLIR approach, a Multilingual Web portal for business intelligence in the information technology

(IT) domain. The prototype initially uses English, Spanish, German, Chinese, and Japanese, but it is designed to be extensible to other languages. The proposed system architecture is shown in Figure 1. Our system architecture consists of three major components: (1) Multilingual Collection Building, (2) Query Translation, and (3) Document Retrieval. In the following, we describe each component in detail.

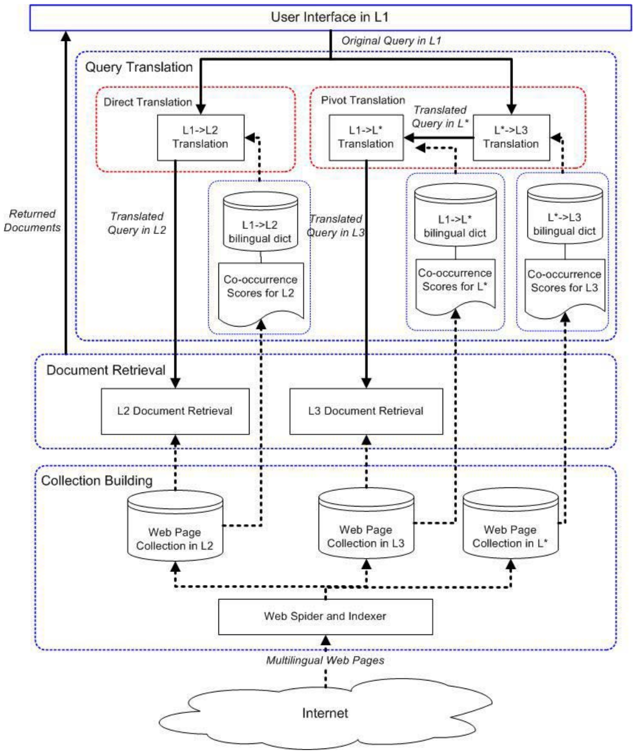


Figure 1: Proposed architecture for multilingual Web retrieval system

4.1 Multilingual Collection Building

Web spiders, or crawlers, are programs that retrieve pages from the Web by recursively following URL links in pages using standard HTTP protocols [7]. The Web Spider component is responsible for building our document collections. Document collections in two or more languages are needed for a multilingual Web portal. These documents are not only the information resources provided to users, but also a comparable corpus that can be used for translation disambiguation and query expansion. To address these needs, we propose a collection building method that combines traditional focused crawling and meta-searching. Similarly to traditional focused crawlers, we start with a set of starting URLs and fetch relevant pages back. At the same time, new starting URLs are identified by querying multiple search engines and combining their top results. This provides both diversity and relevance for our collection. After the Web pages are collected, they can be indexed by Web page indexers to support document retrieval.

4.2 Query Translation

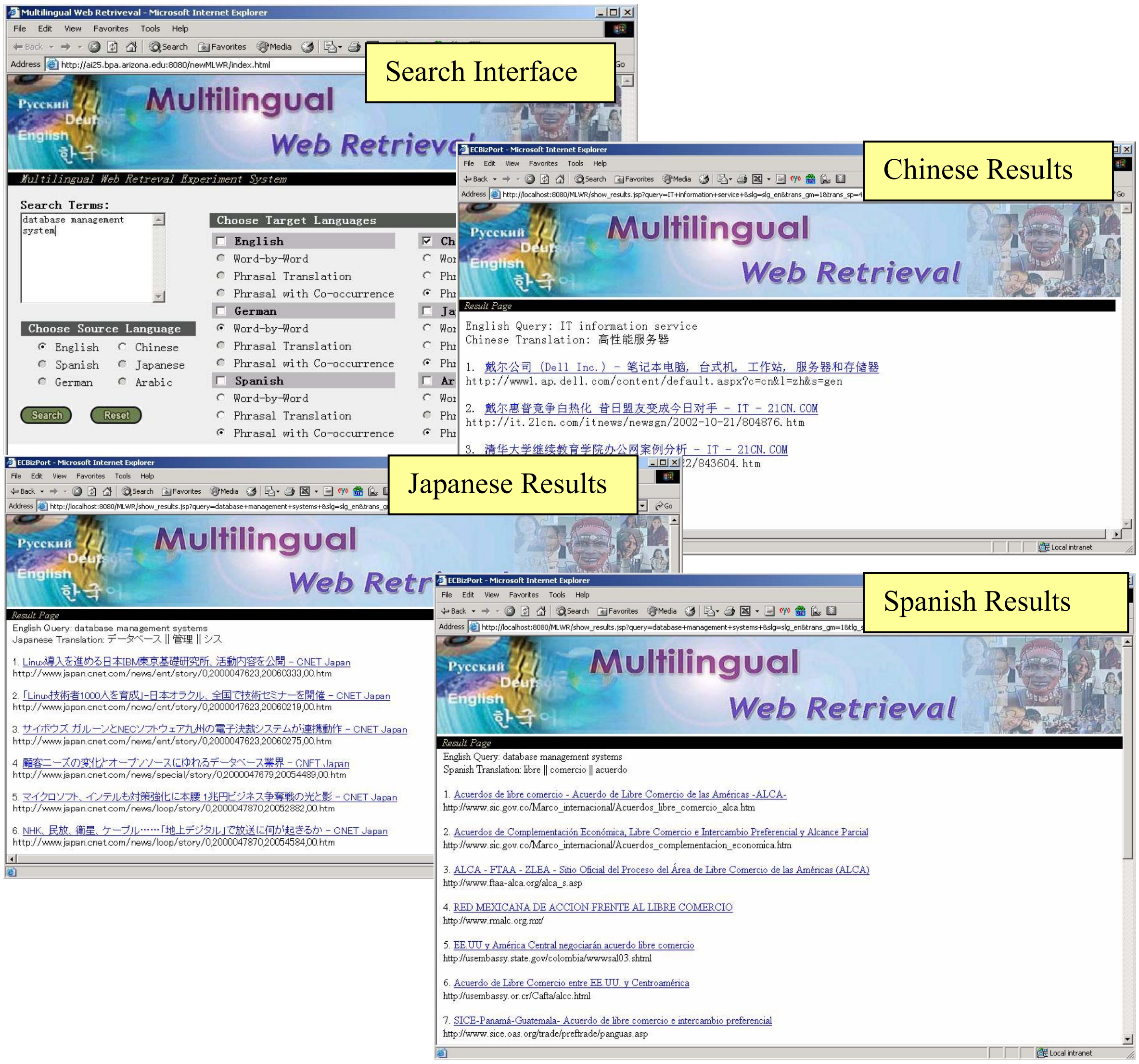
We propose to use a dictionary-based approach combined with phrasal translation and co-occurrence analysis for translation disambiguation. Phrasal translation is used to improve translation accuracy. In the dictionary lookup process, the entry with the smallest number of translation will be preferred to other candidates. In addition, we also propose *maximum phrase matching*. Translations containing more continuous key words will be ranked higher than those containing discontinuous key words.

Co-occurrence analysis also is used to help choose the best translation among candidates. All possible definition pairs in the dictionary are extracted and their co-occurrence scores in our own document collections are calculated. Our method is similar to that of [17] 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 while they calculated the co-occurrence scores "on the fly", we calculated co-occurrence scores in advance, which will not affect run time efficiency and is more suitable for Web applications.

It is not always easy to find suitable bilingual dictionaries between languages. For language pairs L1 and L3, if bilingual dictionaries are not good enough or are not available, direct translation between L1 and L3 may not be possible. However, if there is a dictionary between L1 and L*, and one between L* and L3, it is possible first to translate L1 to L* and then translate from L* to L3. We propose to use pivot language translation where direct translation is not available.

4.3 Document Retrieval

The Document Retrieval component takes the query in the target language and retrieving the relevant documents from the text collection. This component can be designed based on similar retrieval component in traditional retrieval systems.


Multilingual Web Retrieval - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Multilingual
Web Retrieval
Русский
Deutsch
English
한국어
ECBizPort - Microsoft Internet Explorer
Go
Local intranet

Search Interface

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

Multilingual Web Retreval Experiment System

Search Terms:
database management system

Choose Source Language
● English ○ Chinese
○ Spanish ○ Japanese
○ German ○ Arabic

Search Reset

Choose Target Languages
□ English
○ Word-by-Word
○ Phrasal Translation
○ Phrasal with Co-occurrence
□ German
● Word-by-Word
○ Phrasal Translation
○ Phrasal with Co-occurrence
□ Spanish
○ Word-by-Word
○ Phrasal Translation
● Phrasal with Co-occurrence
☑ Ch
○ Wor
○ Phr
● Phr
□ Ja
○ Wor
○ Phr
● Phr
□ Ar
○ Wor
○ Phr
● Phr

Chinese Results

Address: http://localhost:8080/MLWR/show_results.jsp?query=IT+information+service+&slg=slg_en&trans_gm=1&trans_sp=4

Result Page

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

1. 戴尔公司 (Dell Inc.) - 笔记本电脑, 台式机, 工作站, 服务器和存储器
http://www1.ap.dell.com/content/default.aspx?c=cn&l=zh&s=gen

2. 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/newsgn/2002-10-21/804876.htm

3. 清华大学继续教育学院办公网案例分析 - IT - 21CN.COM
22/843604.htm

Japanese Results

Address: http://localhost:8080/MLWR/show_results.jsp?query=database+management+systems+&slg=slg_en&trans_g

Result Page

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

1. Linux導入を進める日本IBM東京基礎研究所、活動内容を公開 - CNET Japan
http://www.japan.cnet.com/news/ent/story/0,2000047623,20060333,00.htm

2. 「Linux技術者1000人を育成」-日本オラクル、全国で技術セミナーを開催 - CNET Japan
http://www.japan.cnet.com/news/ent/story/0,2000047623,20060219,00.htm

3. サイボウズ ガルーンとNECソフトウェア九州の電子決裁システムが連携動作 - CNET Japan
http://www.japan.cnet.com/news/ent/story/0,2000047623,20060275,00.htm

4 顧客ニーズの変化とオープンソースにゆれるデータベース業界 - CNET Japan
http://www.japan.cnet.com/news/special/story/0,2000047679,20054489,00.htm

5. マイクロソフト、インテルも対策強化に本腰 1兆円ビジネス争奪戦の光と影 - CNET Japan
http://www.japan.cnet.com/news/loop/story/0,2000047870,20052882,00.htm

6. NHK、民放、衛星、ケーブル……「地上デジタル」で放送に何が起きるか - CNET Japan
http://www.japan.cnet.com/news/loop/story/0,2000047870,20054584,00.htm

Spanish Results

Address: http://localhost:8080/MLWR/show_results.jsp?query=database+management+systems+&slg=slg_en&trans_gm=1&tlg

Result Page

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

1. Acuerdos de libre comercio - Acuerdo de Libre Comercio de las Américas -ALCA-
http://www.sic.gov.co/Marco_internacional/Acuerdos_libre_comercio_alca.htm

2. Acuerdos de Complementación Económica, Libre Comercio e Intercambio Preferencial y Alcance Parcial
http://www.sic.gov.co/Marco_internacional/Acuerdos_complementacion_economica.htm

3. ALCA - FTAA - ZLEA - Sitio Oficial del Proceso del Área de Libre Comercio de las Américas (ALCA)
http://www.ftaa-alca.org/alca_s.asp

4. RED MEXICANA DE ACCION FRENTE AL LIBRE COMERCIO
http://www.rmalc.org.mx/

5. EE.UU y América Central negociarán acuerdo libre comercio
http://usembassy.state.gov/colombia/wwwsal03.shtml

6. Acuerdo de Libre Comercio entre EE.UU. y Centroamérica
http://usembassy.or.cr/Cafta/alcc.html

7. SICE-Panamá-Guatemala- Acuerdo de libre comercio e intercambio preferencial
http://www.sice.oas.org/trade/preftrade/panguas.asp

Figure 2: User interface of Multilingual Business Intelligence Portal

5. A Multilingual Web Portal for Business Intelligence

In order to demonstrate the feasibility and evaluate the performance of the proposed approach, we implemented a prototype system using the proposed MLIR approach, a Multilingual Web portal for business intelligence in the information technology (IT) domain. The prototype initially uses English, Spanish, German, Chinese, and Japanese, but it is designed to be extensible to other languages. We will also discuss some important issues in multilingual Web retrieval system development.

Figure 2 shows a sample screenshot of the Multilingual Business Intelligence Web Portal prototype. A user can choose a source language to form his/her query, enter a search query in the box provided, and choose among the different target languages and translation methods. The query will be passed to the system for query translation. A set of relevant documents will be retrieved by the system and returned to the user. The translated queries are also displayed to the user so he/she may use the terms to refine the query manually.

5.1 Spidering

The AI Lab SpidersRUs toolkit (http://ai.bpa.arizona.edu/spidersrus/), a digital library development tool developed by our research group, is used to build the multilingual collections for the Web portal. 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.

5.2 Indexing and Stemming

These Web pages need to 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.

Our collections were indexed in two ways: first employing character-based/word-based index, and then using dictionary translations as indexing terms. Using word-based indexing and character-based indexing during our general indexing process avoided information loss. Therefore, we indexed all the pages against their analogous dictionaries. The dictionary word-based indexed terms are potential translations from bilingual dictionaries, and would be used in co-occurrence calculation for translation disambiguation purposes.

Word normalization will lead to much greater improvements in retrieval effectiveness for morphologically rich and lexically complex languages. The indexing procedure uses stemming algorithms for English, Spanish, and German. As a standard, the Porter stemmer is used for the English collection [22]. For Spanish, we 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. According to Chen [6], including both compounds and their composite parts during indexing would improve the performance. We took a completely dictionary-based approach to German word normalization. In a case where a word was not found in the dictionary, we would then search for substrings of the word to see if we could find a match for the word through a matching series of substrings. In Chinese and Japanese, noun phrases do not have morphological variations, so no stemming algorithm was applied to these two languages.

5.3 Query Translation

We use a dictionary-based approach combined with phrasal translation and corpus-based co-occurrence analysis for translation disambiguation. Query term translations were performed using bilingual dictionaries. Table 1 summarizes the dictionaries we used for each language pair.

Language Pair (English-X)	Bilingual Dictionary Used	# of Entries
Chinese	LDC Wordlist	120,000
Japanese	EDICT	106,012
Spanish	EFN Wordlist	25,535
German	TravLang Dictionary	18,554

Table 1: Bilingual dictionaries used in query translation

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.

For each translation pair, all possible definition pairs $\{D_1, D_2\}$ in the bilingual dictionary are extracted such that D_1 is a definition of a term 1 in the source language and D_2 is a definition of a term 2 in the target language. Each pair is used as a query to retrieve documents in the indexed collections. The co-occurrence score between two definitions D_1 and D_2 then can 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.

Besides direct translation, we were interested in investigating the performance of pivot language translation. We experimented with Chinese->Japanese as our initial step in studying this problem. In our pivot language study, Chinese queries were first translated into English using LDC wordlist. The translated English queries were translated into Japanese using EDICT in this use of English as a pivot language between the Chinese-Japanese translation. In both steps, phrasal translation and co-occurrence analysis were performed.

5.4 Document Retrieval

The document retrieval component was performed as in monolingual retrieval. It 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.

6. System Evaluation

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

6.1 Experimental Design and Measures

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 that study.

In general, we followed the standard TREC evaluation process in our experiment design. However, because our study involved Web pages instead of standard collections, there was no established relevance judgment available for precision and recall. Therefore, we recruited human experts to judge the relevance of each document. Since we were particularly interested in how well these techniques would work for Web content in a business intelligence portal, we recruited experts in the business domain. Four bilingual business school students served as domain experts, all fluent in English and one of the target languages (Chinese, Japanese, Spanish, and German). They identified 10 English queries of interest in the business/IT domain and translated these queries into the target language as the base queries. These queries contain 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 get monolingual runs. As discussed, such monolingual retrieval represents the performance of best-case multilingual information retrieval.

The English base queries were used to get multilingual runs based on four settings: word-by-word translation (WBW), phrasal translation, co-occurrence analysis translation, phrasal translation with co-occurrence analysis (Ph-Co).

The experts individually submitted each query to the system under the different settings. The results of the target retrieval 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 picked the first translation in the dictionary and ignored all the other translation candidates. With ten queries and five different settings, each expert performed a total 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 (relevant to the search). The time spent for retrieval was recorded as a measurement of the efficiency of the system.

To compare the effectiveness of the system, we 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 [10]. Precision is calculated as

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

Efficiency is measured by time spent in the system. It was recorded during retrieval.

6.2 Experimental Results and Discussions

Multilingual and Monolingual Comparison. Table 2 compares the best case multilingual performance with that of monolingual performance, measured in precision. On average, multilingual performance achieved 72.99% of monolingual performance, in excess of 2/3 of monolingual performance. This result is encouraging.

Language Pair	Monolingual Precision	Multilingual Precision	% of Monolingual
Chinese	0.68	0.52	76.47%
Japanese	0.54	0.46	85.18%
Spanish	0.58	0.36	62.07%
German	0.76	0.54	70.59%
Average	0.63112	0.4606	72.99%

Table 2: Average precision of monolingual retrieval and multilingual retrieval

Phrasal Translation and Co-occurrence Disambiguation. When comparing improvement from phrasal translation and co-occurrence analysis, we 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 languages pair. English-Chinese and English-Japanese dictionaries are more comprehensive, and contain much 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, there was no doubt that E-C and E-J dictionaries provided more phrase information, which made performance improvement of phrasal translation possible. However, E-G and E-S dictionaries contained very little phrase information and led to little improvement in phrasal translation. We confirmed that having linguistic resources available could significantly improve phrasal translation performance. The performance of phrasal translation is limited by the availability of linguistic resources.

In all cases, co-occurrence analysis quite consistently improved translation performance. We found improvement larger than that in traditional MLIR that could have resulted from the high quality of our 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. Our 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.

Pivot language translation. Our pivot language translation takes Chinese queries and gets Japanese documents through Chinese->English and English->Japanese translations**.** The pivot translation achieved 40% of the performance of monolingual retrieval. Compared with direct translation, it yielded a 52% drop. The performance is encouraging nevertheless, since our pivot language translation is an initial step toward investigating this area and could be used as our benchmark in later research.

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' 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, we conducted a preliminary simulation in which system response times for performing various MLIR 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 MLIR techniques and defined a baseline estimation of their effect on system efficiency. Table 3 summarizes the average time spent under each system setting. Our results showed that phrasal translation with co-occurrence disambiguation took 3.5 times as long as monolingual translation. When pivot translation was involved, the retrieval time increased to 4.7 times that of monolingual retrieval. It should be noticed that our 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 calculation done during indexing time, the efficiency of the prototype is satisfactory.

Method	**Average Time Spent (Sec)**
Monolingual	5.84
WBW	7.25
Co-occurrence	16.47
Phrasal	8.07
Co+Phr	17.48
Pivot	27.35

Table 3: Efficiency of Multilingual Business Intelligence Portal

.

7. 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 paper, we have presented our 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 our 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 our Web portal, following TREC evaluation procedures. Our results showed that our system's phrasal translation and co-occurrence disambiguation led to great improvement in performance. Pivot language translation yielded a 52% drop in performance compared with direct translation, but the approach is still promising. The Web portal was reasonably efficient run on a PC and should achieve better efficiency on a more powerful machine. In sum, our study demonstrated the feasibility of applying MLIR techniques in Web applications and the experimental results are encouraging.

We plan to expand our research in several directions. First, we 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 to the user's familiar language using a commercial machine translation product. We are 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, we 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 we believe that different domains might have different effects on the performance of MLIR techniques, we are interested in testing our approach in other domains, such as medicine.

8. Acknowledgements

This project was supported in part by an NSF Digital Library Initiative-2 grant, PI: H. Chen, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999-March 2002. 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.

9. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Global Reach (2002). "Global Internet Statistics," available at: http://www.glreach.com/globstats/

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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, California, United States, August, 1999, pp. 74-81.

[20] 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.

[21] 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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.