Cross-Language Information Access to Multilingual Collections on the Internet

Guo-Wei Bian and Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China. E-mail: gwbian@nlg.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

Language barrier is the major problem that people face in searching for, retrieving, and understanding multilingual collections on the Internet. This paper deals with query translation and document translation in a Chinese-English information retrieval system called *MTIR*. Bilingual dictionary and monolingual corpus-based approaches are adopted to select suitable translated query terms. A machine transliteration algorithm is introduced to resolve proper name searching. We consider several design issues for document translation, including which material is translated, what roles the HTML tags play in translation, what the tradeoff is between the speed performance and the translation performance, and what form the translated result is presented in. About 100,000 Web pages translated in the last four months of 1997 are used for quantitative study of online and real-time Web page translation.

1. Introduction

Internet and digital libraries make available heterogeneous collections in various languages. They provide many useful and powerful information dissemination services. However, about 80 percent of Web sites are in English and about 40 percent of Internet users do not speak English (Euro-Marketing Associates, 1999; Grimes, 1996; Hershman, 1998). Language barrier becomes the major problem in searching, retrieving, and understanding materials in different languages.

Several issues have to be addressed to design a multilingual information processing system. The issues involve these basic operations for multilingual data management:

(1) Data input: input methods;
(2) Data representation: character sets and coding systems;
(3) Data manipulation: language identification, query translation, document translation, information retrieval and extraction, summarization, and so on;
(4) Data output: font mapping.

Issues 1, 2, and 4 have been resolved by operation systems, application programs, or packages that can handle both single-byte and multiple-byte coding systems for different language families. Thus, how to manipulate the multilingual data becomes the major issue. To find solutions, we have to understand the relationship between these operations. A cross-language information-access model is shown in Figure 1.

The first layer is the user-interface. It deals with user requests and system responses. Layer 2 touches on language barriers. On the one hand, it translates the user's information need from the user's familiar language into other languages. On the other hand, the requested material is translated from other languages into the user's familiar language. At layer 3, systems may perform information retrieval, information extraction, information filtering, text classification, text summarization, or other text-processing tasks. These tasks may be operated interactively. For example, the relevant documents retrieved by an information retrieval system may be summarized to users. The fourth layer determines the language categories of the selected resources, and passes them to the other layers. This paper focuses on layer 2. Two kinds of operations—retrieval and browse—are allowed at layer 3. Figure 2 sketches the data flows.

We use Chinese and English as examples. Users work in a Chinese environment to access English and Chinese materials. Both user request and system response are in Chinese no matter what language the materials belong to.

Several papers (Ballesteros & Croft, 1996, 1997; David, 1996; David & Dunning, 1995; Dumais, Littman, & Landauer, 1997; Hull & Grefenstette, 1996; Landauer & Littman, 1990), which may be categorized into dictionary-based, corpus-based, and hybrid-based approaches, have made proposals to deal with query translation. The dictionary-based approach uses a bilingual dictionary to select the target terms for source queries. Coverage is a major problem in this approach. New words are produced very quickly, so that it is hard to put all the words in dictionaries. The corpus-based approach uses parallel or comparable aligned corpora to treat word selection. Domain shifts, term-align accuracy, and scale of corpora are major limitations of the corpus-based approach. This paper employs bilingual dic-

© 2000 John Wiley & Sons, Inc.

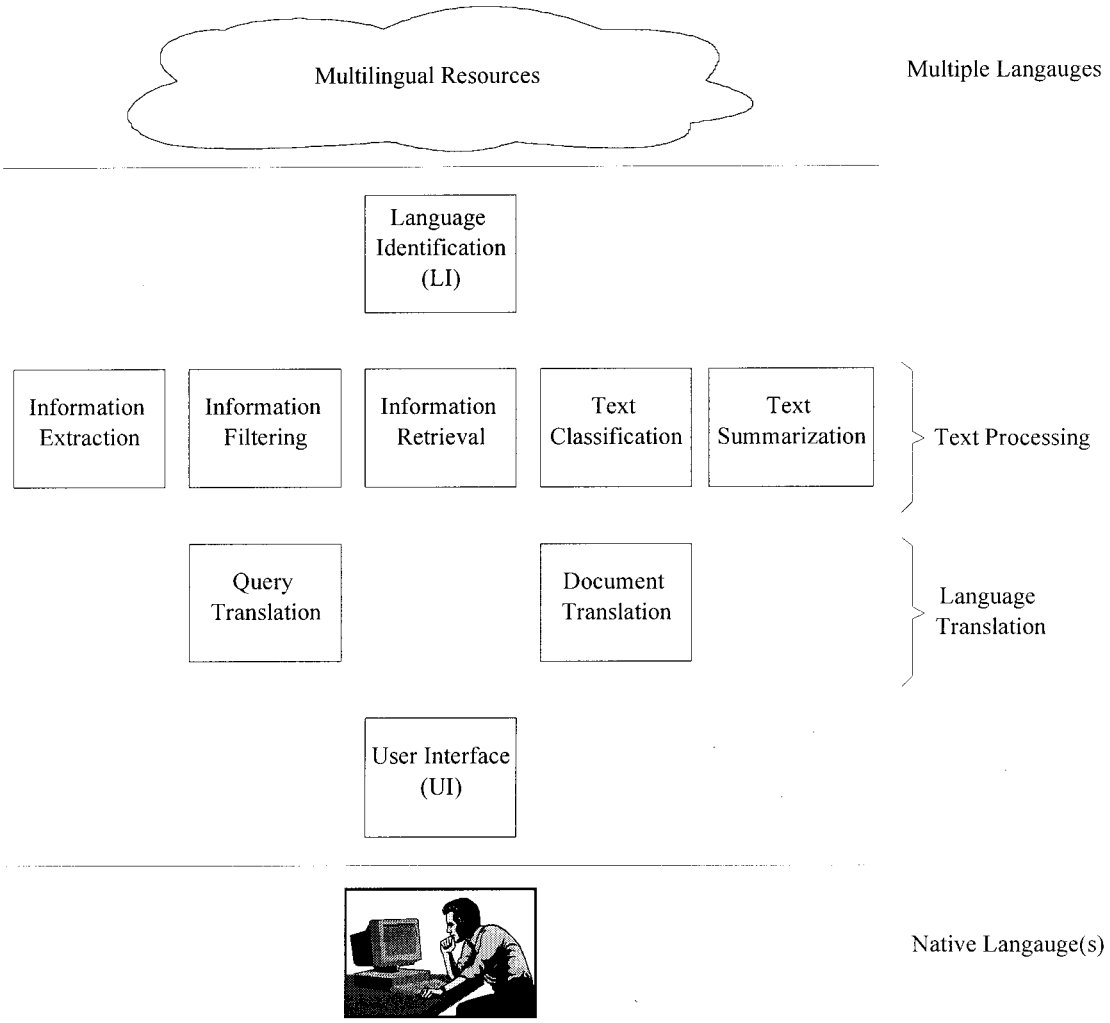

FIG. 1. A four-layer model for multilingual information processing.

tionaries and monolingual corpora to select lexical items and to touch on unknown word translation problems.

Many different approaches (Bennett & Slocum, 1985; Brown et al., 1990; Chen & Chen, 1995; Nagao, 1984; Mitamura et al., 1991; Baker et al., 1994) to machine-translation system design have been proposed in literature. However, traditional machine-translation technologies cannot be applied directly to online real-time Web-page translation. Bian and Chen (1997) pointed out several issues have to be studied, including these: Which material is translated? What roles do HTML tags play in translation? What is the tradeoff between the speed performance and the translation performance? What form is the translated result presented in? Where is the translation capability implemented? This paper integrates the query translation and document translation together in a cross-language information access system. Section 2 proposes a Chinese-English information retrieval system. Sections 3 and 4 present our query translation and document translation strategies. Section 5 concludes our remarks.

2. A Chinese-English Information Retrieval System on the WWW

On the Web, the distinct systems can be easily integrated as a larger distributed system using the HTTP protocol. Each system can be involved using an URL of CGI program. First, the CGI program gets input data from the caller. Then the caller receives the resultant material from the server system. Figure 3 shows the basic architecture of a Chinese-English information retrieval system on WWW called MTIR system (Bian & Chen, 1997). The user interface to access the system is an HTML form, which can be invoked by a standard WWW browser. The form allows users to input the URL of a Web page or a WWW site to navigate. Alternatively, users can type a query in English or Chinese and select the Alta Vista, Excite, Infoseek, Lycos, MetaCrawler, or Yahoo search engine.

Users express their intention by inputting URLs of Web pages or queries in Chinese/English. A Chinese query is translated into the English counterpart using a query-translation mechanism. The translations of query terms are disambiguated using word co-occurrence relationship (Bian & Chen, 1998).

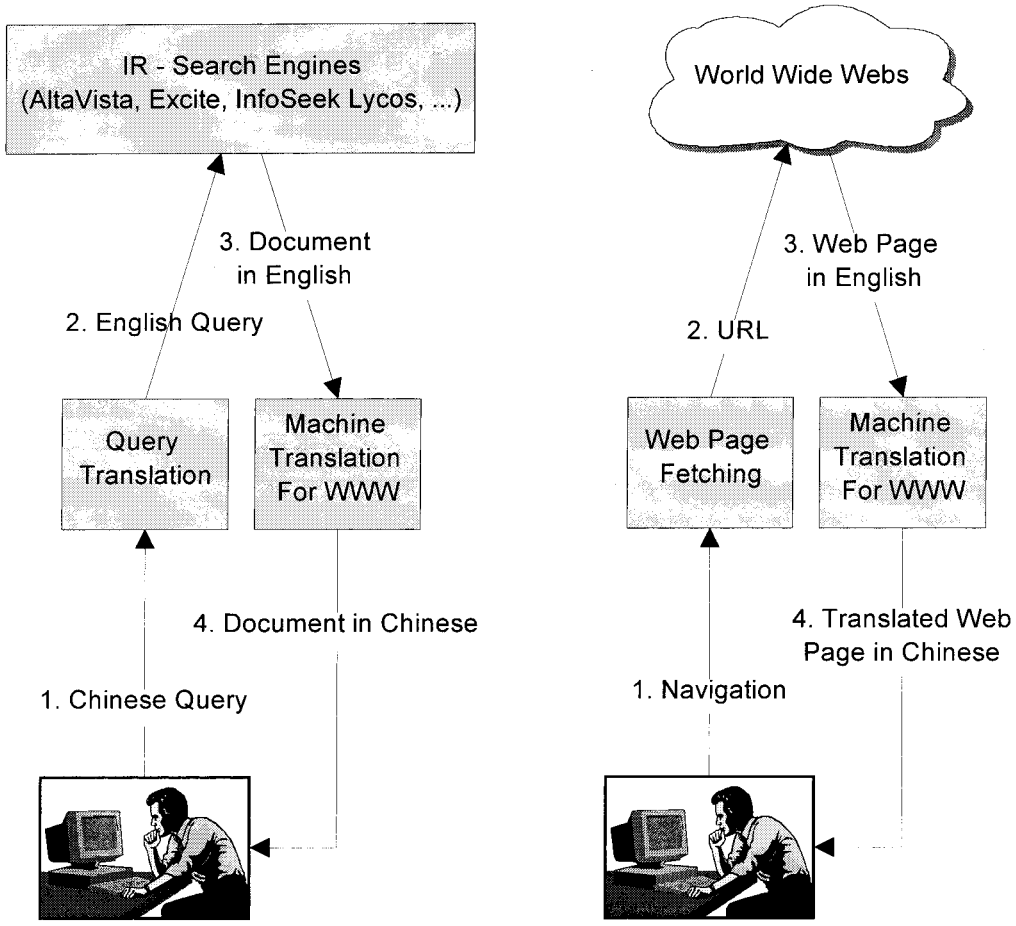

FIG. 2. Data flows in cross-language information access.

Then the system sends the translated query to the searching engine selected by the user in the user interface. The query subsystem takes care of the user interface part.

The subsequent navigation of the WWW is under the control of a communication subsystem. To minimize the Internet traffic, a caching module is presented in this subsystem and some proxy systems are used to process the request. The objects in the cache are checked when a request is received. If the requested object is not found, the communication system fetches the HTML file (.htm or .html file) or text file (.txt or .text file) from the neighboring proxy systems or the original server.

The HTML analyzer examines the retrieved file. In HTML documents, many word strings don't have sentence terminators to indicate the boundary of meaning. HTML tags, such as title, headings, unordered lists, ordered lists, definition lists, forms and tables, play roles similar to that of punctuation marks, such as full stop, question mark, and exclamation mark. These tags divide the whole file into several translation segments for the machine-translation subsystem. We call such segments quasi-sentences. In contrast to the above tags, the font-style elements, such as bold, italic, superscript, and subscript, may produce many unknown words because the whole word is split into several parts. Thus these font-style elements should be hidden from the attributed words during translation processing.

After receiving the first translated document, users may access other information through hyperlinks. We attach our system's URL to those URLs that link to HTML files or text files, which guarantees that successive browses are linked to our system. The other URLs, linking to the inline images and external MIME objects, are changed to their absolute URLs. In other words, the nontextual information is received from the original servers. Our experimental system was made available in July 1997, and became accessible through the URL (http://mtir.csie.ntu.edu.tw). Our system has had 77,000 visitors and 400,000 translations have been made. Figure 4 shows the home page of MTIR. Figures 5 and 6 give a scenario for English-Chinese Web translation after clicking "Library of Congress: Digital Library Collection" and "The LIBRARY TODAY." Figures 7 and 8 demonstrate a scenario to access the Berkeley SunSITE and issue a Chinese query 數位圖書館 (shu4wei4 tu2shu1guan3, digital library).

3. Query Translation

3.1 Query Disambiguation

The query translation procedure for a Chinese-English information retrieval system consists of three major steps:

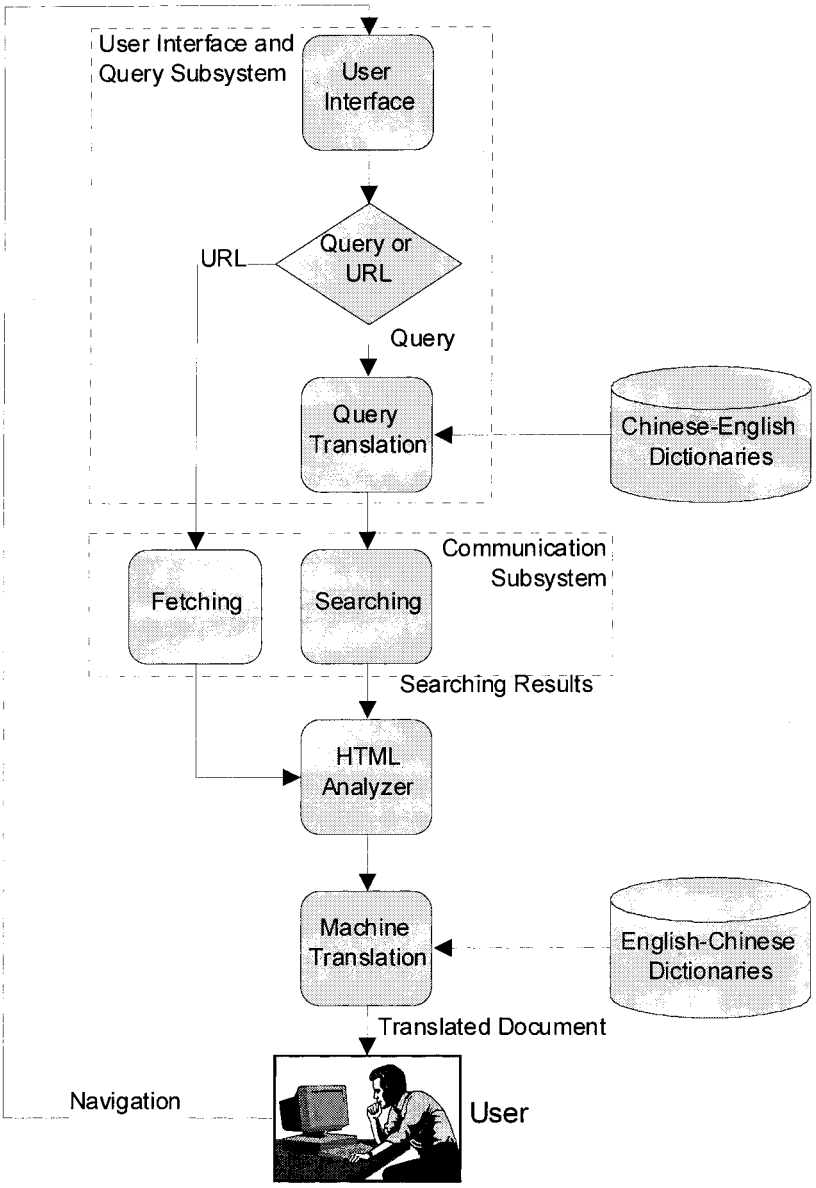

FIG. 3. A Chinese-English information retrieval system MTIR.

(1) Query segmentation: identifying the word boundary of a Chinese query.
(2) Query disambiguation: constructing an English query.
(3) Monolingual IR: selecting the relevant documents using the modified query.

Because Chinese sentences are composed of a sequence of characters without boundaries (Chen & Lee, 1996), segmentation is required. Query segmentation and query disambiguation tasks employ the same bilingual dictionary in this design, which speeds up the dictionary lookup and avoids the inconsistencies resulting from two dictionaries (i.e., segmentation dictionary and transfer dictionary). In the current version, the bilingual dictionary consists of 67,000 English terms, 72,000 Chinese terms, and approximately 125,000 bilingual pairs. The longest-matching method is adopted in Chinese segmentation. The segmentation system searches for a dictionary entry corresponding to the longest sequence of Chinese characters from left to right. After identification of Chinese terms, the system selects the translation equivalents for each query term from the bilingual dictionary. Those terms that are missing from the bilingual dictionary are passed unchanged to the final query, except that proper names are treated in the way shown in Section 3.2.

3.1.1 Selection Strategies

When there is more than one translation equivalent in a dictionary entry, these selection strategies may be explored:

(1) Select-All (SA): The system looks up each term in the bilingual dictionary and constructs a translated query by concatenating of all the senses of the terms.

FIG. 4. Home page of MTIR.

(2) Select-Highest-Frequency (SHF): The system selects the sense with the highest frequency in target language corpus for each term. Because the translation probabilities of senses for each term are unavailable without a large-scale word-aligned bilingual corpus, the translation probabilities are reduced to the probabilities of senses in the target language corpus. In other words, the frequently used sense of a term is used instead of the frequently translated sense.

(3) Select-N-POS-Highest-Frequency (SNHF): This strategy selects the highest-frequent sense of each POS candidate of the term. If the term has N POS candidates, the system will select N translation senses. Compared to strategy (3), strategy (2) always selects only one sense for each term.

(4) Word co-occurrence (WCO): This method considers the content around the translation equivalents to decide the best target equivalent. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of this term and other terms. We adopt mutual information (MI) to measure the strength. MI is defined as:

$$MI(X, Y) = \log_2 \frac{P(X, Y)}{P(X)P(Y)}$$

where X and Y denote two terms,
$P(X)$ and $P(Y)$ are probabilities of X and Y,
$P(X,Y)$ is their co-occurrence probability.

If MI(X,Y) $>>$ 1, then X and Y have strong relationship; if MI(X,Y) $\approx$ 0, then X and Y have no relationship; and if MI(X,Y) $<<$ 0, then X and Y are negatively correlated. The mutual information can be calculated from the retrieval document collection to prevent the domain shift problems in traditional corpus-based approach for query translation.

3.1.2 Experiments

To evaluate the above strategies, we conducted the experiments shown in Figure 9. CACM text collection and Smart information retrieval system (Salton & Buckley, 1988) were employed. The CACM collection contains 3204 texts and 64 queries in English. Each query has the relevant judgments for evaluation. We created the Chinese queries by translating the original English queries into Chinese ones manually. The Chinese queries are regarded as input queries. Table 1 shows the original CACM query 31 and the

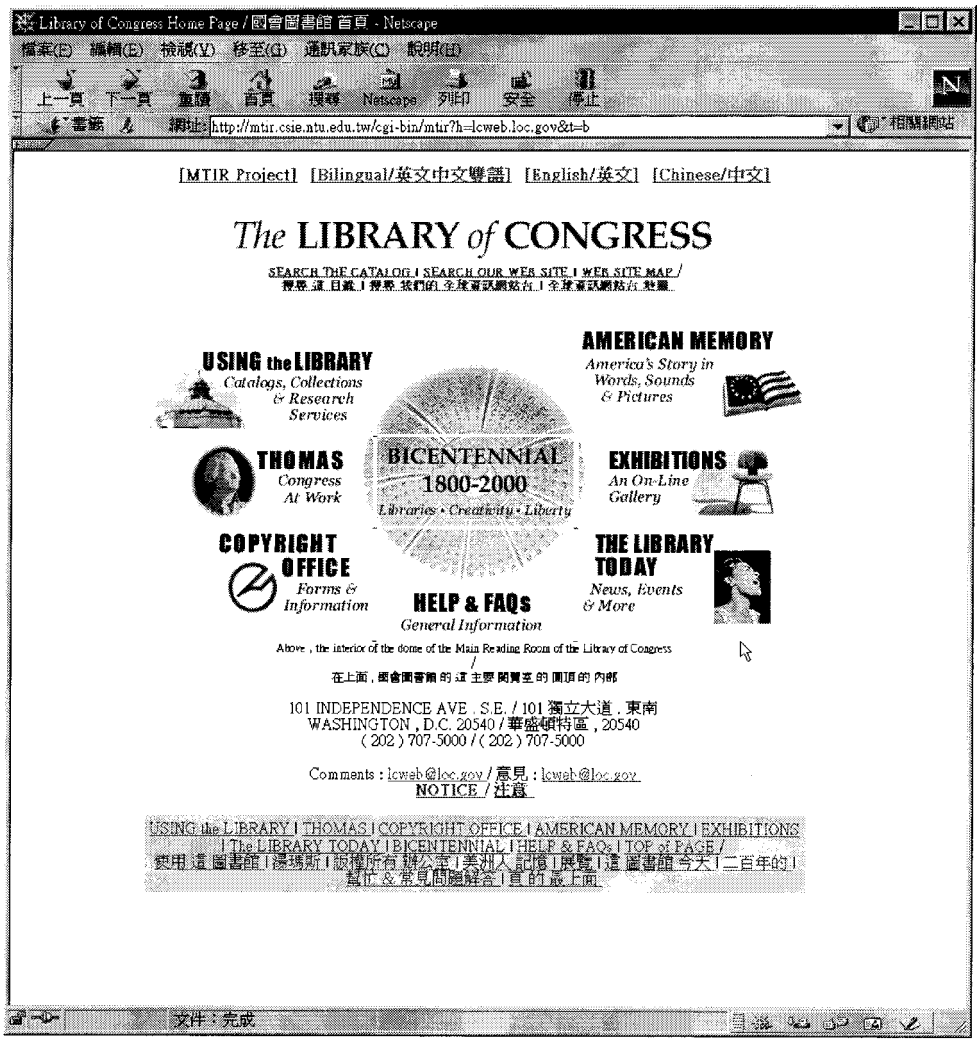

FIG. 5. Bilingual display after clicking "Library of Congress: Digital Library Collections."

human-translated Chinese one. The mutual information is trained using a window size 3 for adjacent words in the text collection. Totally, there are 247,864 distinct word pairs.

The input Chinese query is segmented into several terms, and then translated to four possible representations using various selection methods. Table 2 illustrates an example for the different selection strategies. The Chinese concept 奇異值分解 (ji1yi4 zhi2 fen1jie3) and its phrase-level translation "singular value decomposition" are employed. Column 3 lists the translation equivalents in bilingual dictionary for the query terms at word-level. Four translated representations using different selection strategies on the word-level translation are shown. The MI scores of word pairs "singular value," "singular analysis," "singular decomposition," "value analysis," and "value decomposition" are 6.099, 4.225, 6.669, 1.823, and 4.377, respectively. Other word pairs have no co-occurrence relations in CACM text collection. Considering the example, the translation equivalent "singular" of the term 奇異 (ji1yi4) has the largest MI score with all translation equivalents of the other two words.

Our experiments compare the retrieval performance of the original queries and the four translated versions of Chinese queries. Table 3 shows the performance of various methods. Average 11-point precision is adopted. Rows 3 and 4 show the results compared with the monolingual retrieval and the simple Select-All method. WCO, which is the best method, achieves 65.18 percent of monolingual retrieval. Its performance is 42.28 percent better than that of SA method.

3.1.3 Experiments on Phrase-Level Query Translation

With the dictionary-based approach, three problems result in the major loss in effectiveness of 40 to 60 percent below that of monolingual IR (Hull & Grefenstette, 1996; Ballesteros & Croft, 1997). These factors are (a) missing terminology, (b) translation ambiguity, and (c) the identification and translation of multiterm concepts as phrases.

Among these factors, the correct identification and translation of multiword expressions make the biggest difference

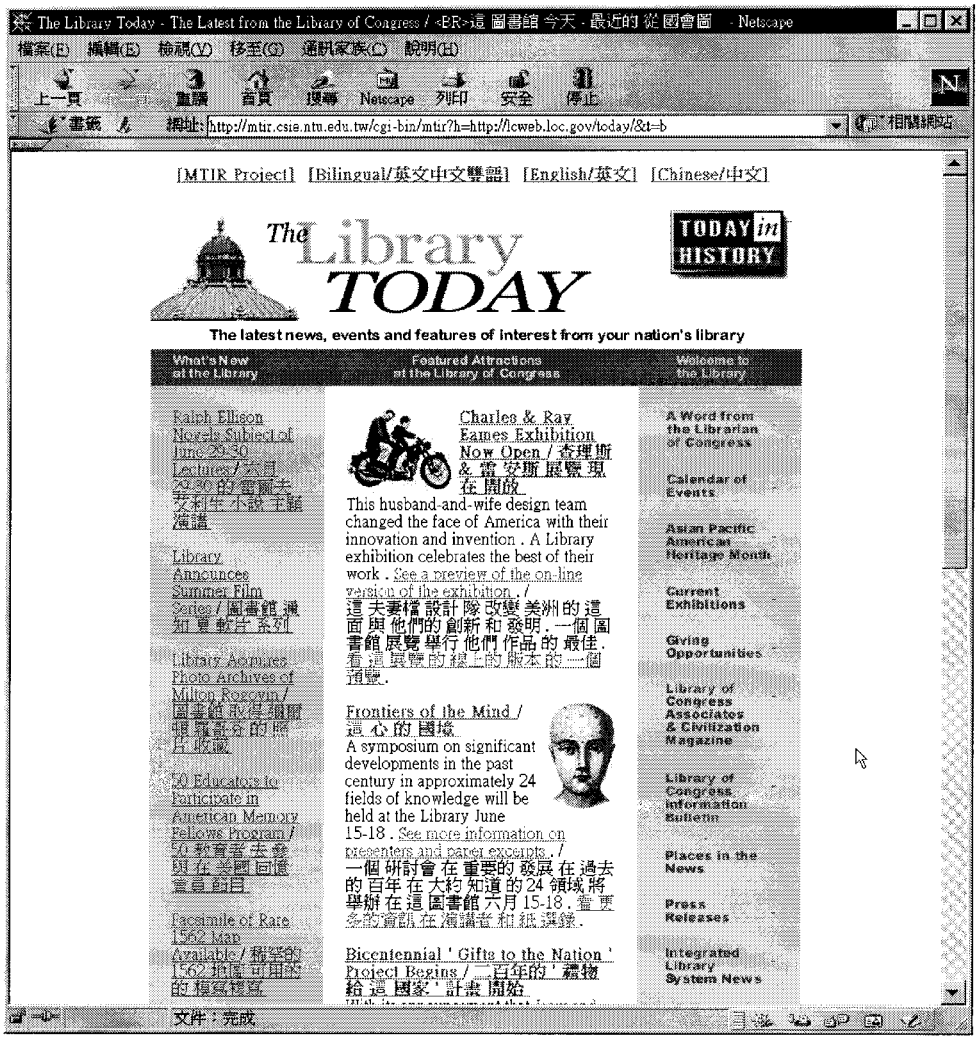

FIG. 6. Bilingual dispaly after clicking "The LIBRARY TODAY."

in average performance (Hull & Grefenstette, 1996). Although dictionaries contain a number of phrasal entries, there are many lexical phrases that are missing. These are typically the technical concepts and the terminology in specific domain. To compare the performances of the word-level translation and phrase-level translation, the CACM English queries are manually checked to find the multiterm concepts that don't appear in our bilingual dictionary. These phrases and their translations are added into the bilingual dictionary for the phrase-level experiments. Altogether, 102 multiword concepts (e.g., singular value decomposition (奇異值分解, ji1yi4 zhi2 fen1jie3), digital image processing (數位影像處理, shu4wei4 ying3xiang4 chu3li3), etc.) are identified in the CACM queries.

By the longest-matching method, the segmentation can handle the identification of these multiword concepts easily within the string of Chinese characters. For example, the string 數位影像處理 (shu4wei4 ying3xiang4 chu3li3, digital image processing) will be segmented into these three words 數位 影像 處理 (shu4wei4 ying3xiang4 chu3li3) if the concept is not stored in the bilingual dictionary. When the concept appears in the bilingual dictionary, it will be considered as the whole word 數位影像處理 (shu4wei4ying3xiang4chu3li3) instead.

Table 4 shows the performance when phrases are added to the bilingual dictionary. The performance of all four methods is enhanced. Compared to the baseline model (i.e., monolingual retrieval), the precision rates of SA, SHF, SNHF, and WCO are increased 11.34 percent, 12.63 percent, 11.99 percent, and 9.53 percent, respectively. On average, the phrase-level translation performance gain is nearly 20 percent from the word-level translation. The WCO method performance gain is less from phrasal translation than from other methods. The major reason is that the WCO method can capture some identification and translation of multiterm concepts in word-level experiments. The word co-occurrence disambiguation can perform good translations even when the multiterm concepts don't appear in the bilingual dictionary or the phrases are not identified in the source language.

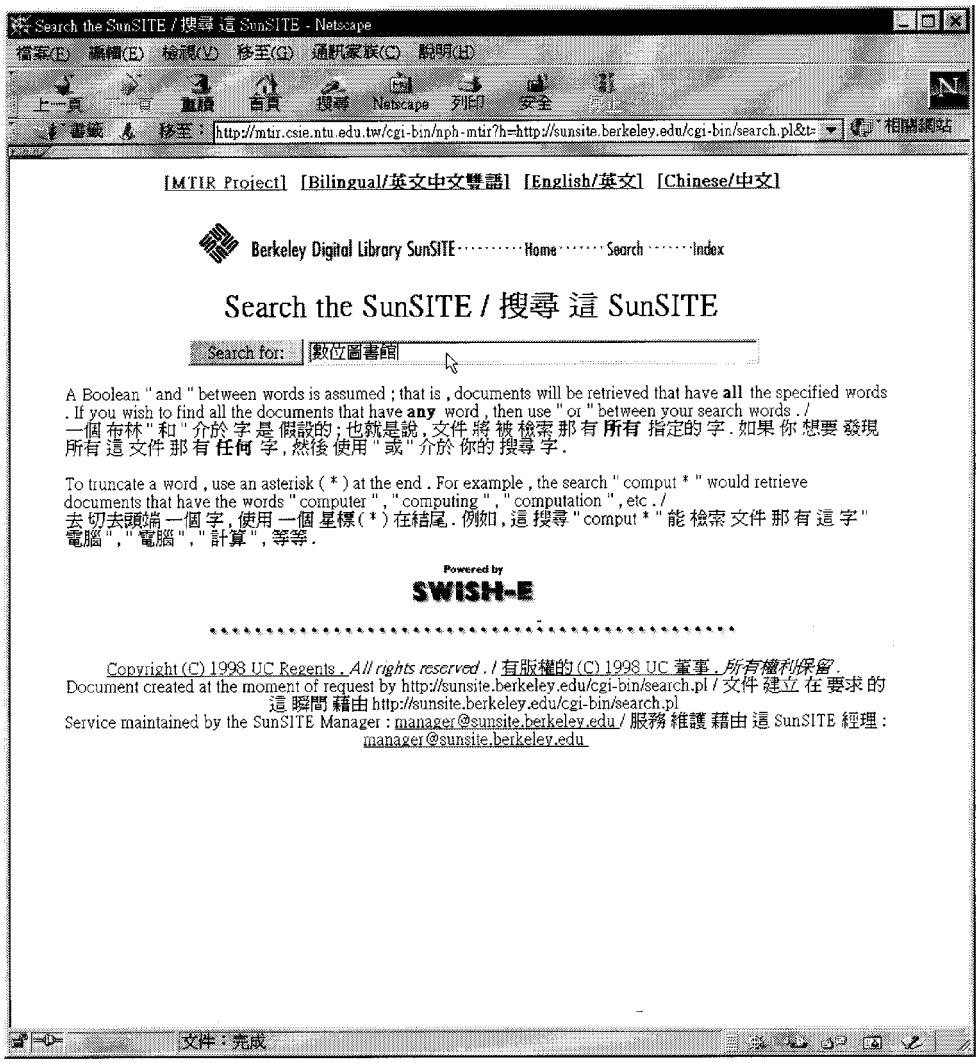

FIG. 7. Access the Berkeley SunSITE and Issue a Query 數位圖書館 (shu4wei4 tu2shu1guan3).

3.1.4 Experiments on Short Queries

In real-world searching, queries are usually short. The average number of terms in user's queries are 1.5–2 words, and rarely more than 4 words (Fitzpatrick & Dent, 1997). To evaluate selection methods under short queries, we conducted other experiments. Three researchers helped us create the English and Chinese versions of short queries from the original English queries of CACM. The queries were shortened to four words. Table 5 shows the short queries for CACM query 31.

Tables 6 and 7 show the performance of word-level translation and phrase-level translation in short queries. The 11-point average precision of the monolingual short English queries is decreased to 29.85 percent. WCO strategy gets 72.96 percent and 87.14 percent performance of the baseline model on word and phrase levels. It is still the best of the four strategies.

3.1.5 Feasibility and Portability

To test the feasibility and portability in other domains, we adopted the same methodologies to a study of other document collections and queries. The TREC-6 text collection and TREC topics 301–350 (Harman, 1997) were used to evaluate the performance. The text collection contains 556,077 documents, and is about 2.2 gigabytes. The collection was also employed to calculate co-occurrence statistics using a context window size 3. Altogether, there are 8,273,633 distinct word pairs. A TREC topic is composed of several fields. The fields of title and description are regarded as queries. Because the goal is to evaluate the performance of Chinese-English information retrieval on different models, we translated these 50 English queries into Chinese by human translation rather than machine translation.

Table 8 shows the retrieval performance of different methods. The 11-point average precision of the monolingual retrieval is 14.49 percent. The performances of SA, SHF, SNHF, and WCO are 6.52 percent, 8.65 percent, 8.57 percent, and 9.78 percent, respectively. The word co-occurrence (WCO) model is the best of the four models and achieves up to 67.49 percent of monolingual performance. It shows a 50 percent greater improvement than the simple

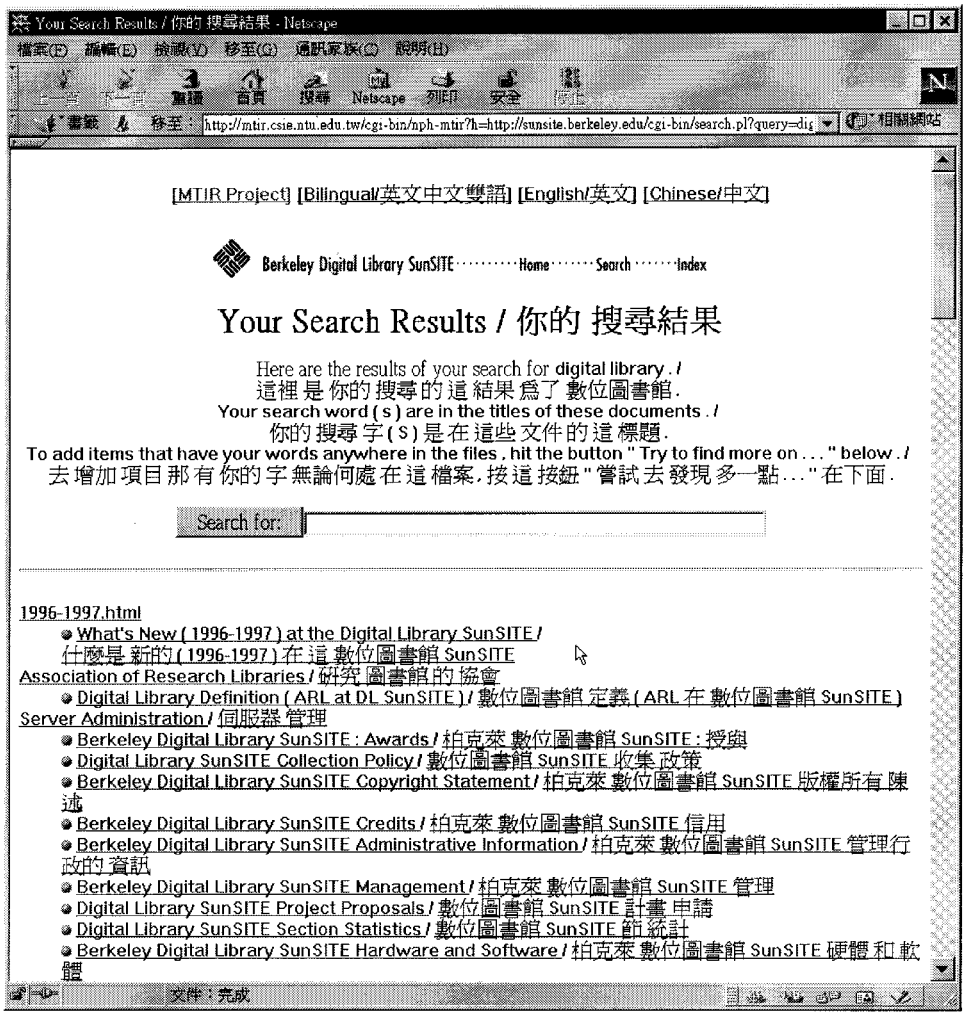

FIG. 8. Search results for the Chinese query "數位圖書館" (shu4wei4 tu2shu1guan3).

select-all (SA) strategy. Thus WCO method is adopted for query disambiguation in Chinese-English information retrieval on WWW.

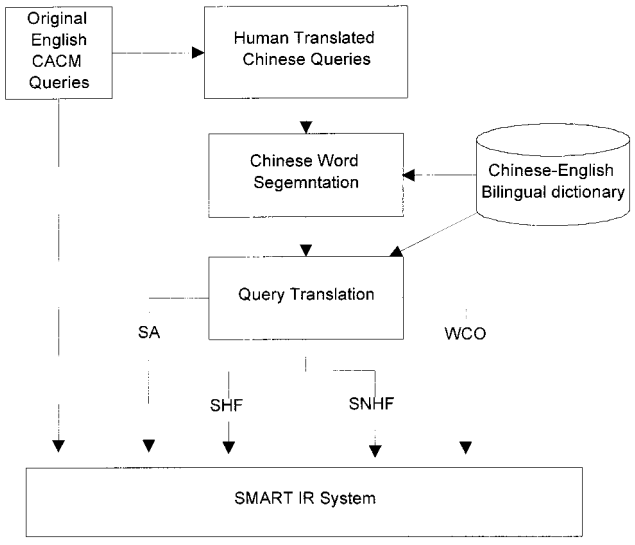

FIG. 9. Experiment outline.

3.2 Proper Name Translation

The percentage of user queries containing proper names is very high. Thompson and Dozier (1997) reported an experiment over a period of several days in 1995. It showed 67.8 percent, 83.4 percent, and 38.8 percent of queries to the *Wall Street Journal*, *Los Angeles Times*, and *Washington Post*, respectively, involve name searching. In cross-language information retrieval, three tasks are needed: name identification, name translation, and name searching. Because proper names are usually unknown words, it is hard to find in a monolingual dictionary not to mention a bilingual dictionary. MTIR incorporates a machine transliteration algorithm (Chen, Huang, Ding, & Tsai, 1998) to deal with this problem.

Chinese and English are the source language and the target language, respectively, in our query translation. The alphabets of these two languages are totally different. Wade-Giles (WG) and Pinyin are two famous systems for romanizing Chinese (Lu, 1995). The proper name translation problem can be formulated as:

- Collect English proper name sets from WWW.

TABLE 1. The original CACM query 31 and the human-translated one.

Type	Query
Original	I'd like to find articles describing the use of singular value decomposition in digital image processing. Applications include finding approximations to the original image and restoring images that are subject to noise. An article on the subject is H. C. Andrews and C. L. Patterson "Outer product expansions and their uses in digital image processing", American Mathematical Monthly, vol. 82.
Human Translated	我想要找敘述用於數位影像處理的奇異值分解的文章. 應用包含尋找對於原來影像及有雜訊的影像修復的近似法. 有關這主題的一篇文章是 H. C. Andrews 和 C. L. Patterson 發表在美國數學月刊第 82 卷上的 "外積的擴展及在數位影像處理上的使用".

- Identify Chinese proper names from queries.
- Romanize the Chinese proper names.
- Select candidates from suitable proper name sets.

In this way, the translation problem is transferred to a phonic string-matching problem. If an English proper name denotes a Chinese entity, for example, Lee Teng-hui denotes 李登輝 (president of ROC), the matching is simple. Otherwise, the matching is not trivial. For example, we issued a Chinese query 阿爾卑斯山 (a1-er3-bei1-si1-shan1) to retrieve information about Alps. The Pinyin romanization of this name is a.er.bei.si.shan. In this notation, the dot is inserted in the romanization of Chinese characters for clear reading. Later, the dot may be dropped when strings are matched. The string "aerbeisishan" is not similar to the string "alps." We developed several language models incrementally to tackle the translation problem. The first issue that we considered was how many common characters there are in a romanized Chinese proper name and an English proper name candidate. Here the order is significant. Consider the Chinese query 埃斯其勒斯 (ai1-si1-ji1-le4-si1). Its WG romanization is "ai.ssu.chi.le.ssu." The corresponding proper name is Aeschylus. Three characters (shown as follows in underline) are matched in order:

```
aeschylus
ais  suchilessu
```

We normalize it by the length of the candidate (i.e., 9), and get a score of 0.33. In an experiment we conducted, there were 1534 pairs of Chinese-English person names. We did a mate matching by using each Chinese proper name as a query and searching for its corresponding English proper name from the 1534 candidates. The performance was evaluated in such a way that how many candidates should be proposed to cover the correct translation. In other words, the average rank of correct translations is reported. The performances of the baseline model under WG and Pinyin systems are 40.06 and 31.05, respectively. The major problem of the baseline model is that if a character is matched incorrectly, those characters that follow this incorrectly matched character will not contribute to the matching. In the above example, chi (其) is useless for translation.

To reduce the error propagation, we considered mate-matching syllables instead of whole words of the candidate names. For example, Aeschylus contains three syllables. The matching is shown as follows:

```
aes    chy  lus
aissu  chi  lessu
```

As a result of syllable matching, the score increases to 0.67 (6/9). In a similar experiment, the performances of the new

TABLE 2. Different word-level translations of Chinese concepts '奇異值分解' (ji1yi4 zhi2 fen1jie3, singular value decomposition).

Term	POS	SA	SHF	SNHF	WCO
奇異 (ji1yi4)	N	oddity singularity		singularity	
	ADJ	singular	singular	singular	singular
值 (zhi2)	N	value worth	value	value	value
分解 (fen1jie3)	N	decomposition analysis dissociation cracking disintegration		decomposition	decomposition
	V	analyze anatomize decompose decompound disassemble dismount resolve	analyze	analyze	
	XV	(split up) (break up)		(split up)	

TABLE 3. Average precision of word-level query translation.

	Original English query (monolingual)	SA	SHF	SNHF	WCO
Average 11-point precision	35.78%	16.39%	21.89%	19.33%	23.32%
% of Monolingual	Baseline	45.81%	61.18%	54.02%	65.18%
% Change		Baseline	+33.56%	+17.94%	+42.28%

TABLE 4. Average precision of phrase-level query translation.

	Original English query	SA	SHF	SNHF	WCO
Average 11-point precision	35.78%	20.45%	26.41%	23.62%	26.73%
% of Monolingual	Baseline	57.15%	73.81%	66.01%	74.71%
% Change		Baseline	+29.14%	+15.50%	+30.71%

language model improved. The average ranks were 35.65 and 27.32 for WG and Pinyin systems, respectively.

Observing the performance differences between WG and Pinyin systems, we found that they use different phones to denote the same sounds. Some examples are:

(1) Consonants

p vs. b; t vs. d; k vs. g; ch vs. j; ch vs. q; hs vs. x; ch vs. zh; j vs. r; ts vs. z; ts vs. c

(2) Vowels

-ien vs. –ian; -ieh vs. –ie; -ou vs. –o; -o vs. –uo; -ung vs. –ong; -ueh vs. –ue; -uei vs. –ui; -iung vs. –iong; -i vs. -yi

A new language model integrates the alternatives. The average rank of the mate-matching algorithm is 25.39. The result is better than that of separate romanization systems.

In the above ranking, each matching character is given an equal weight. We postulate that the first letter of each romanized Chinese character is more important than the remaining characters. For example, *c* in *chi* is more important than *h* and *i*. Thus it should have a higher score. What follows is a new scoring function:

$$\text{score} = \Sigma_i((f_i^*(el_i/(2^* cl_i) + 0.5) + o_i^*0.5)/el$$

where

el: length of English proper name;
el_i: length of syllable *i* in English proper name;
cl_i: number of Chinese characters corresponding to syllable *i*;
f_i: number of matched first-letters in syllable *i*;
o_i: number of matched other letters in syllable *i*.

To test this new model, we used the earlier syllabic example of Aeschylus and capitalized the first letter of each syllable:

```
aes    chy  lus
AiSsu  Chi  LeSsu
```

The corresponding parameters are listed below:

$el_1 = 3$, $cl_1 = 2$, $f_1 = 2$, $o_1 = 0$, $el = 9$,

$el_2 = 3$, $cl_2 = 1$, $f_2 = 1$, $o_2 = 1$,

$el_3 = 3$, $cl_3 = 2$, $f_3 = 2$, $o_3 = 0$.

The new score of this candidate is 0.83. Under the new model, the average rank is 20.64. If the first letter of a romanized Chinese character is not matched, it is given a penalty. The average rank of the enhanced model is 16.78.

We further consider the pronunciation rules in English. For example, *ph* usually has *f* sound. If the similar rules are added to the language model, the average rank is enhanced to 12.11. Table 9 summarizes the distribution of ranks of the correct candidates. The first row shows the range of ranks. The second row shows the number of candidates within the range. About one-third have rank 1. On average, only 0.79 percent of candidates have to be proposed to obtain the correct solution, which makes this method quite effective.

We also performed two additional experiments. Given a query, the best model was adapted to find English locations. There were 1574 candidates in this test. The average rank was 17.40. In other words, 1.11 percent of candidates were proposed. When we merged the person name and the location sets and repeated the experiment, the performance dropped to 27.70. This change emphasizes the importance of classification of proper names. Chen, Ding, Tsai, and Bian (1998) propose a name-entity extraction algorithm to identify and classify Chinese proper names such as person names, organization names, and location names. It is useful to machine transliteration.

4. Document Translation

4.1 Real-Time Web Translator

The requirement for an online real-time machine translation system for users to navigate the WWW is different from traditional off-line batch MT systems. An assisted MT system should help users quickly understand the Web pages and find the interested documents during navigation on a very huge information resource. To fit the requirements, we

TABLE 5. Short queries for CACM query 31.

Type	Query
Short English	singular value decomposition, digital image processing, noise
Short Chinese	奇異值分解 (ji1yi4 zhi2 fen1jie3), 數位影像處理 (shu4wei4 ying3xiang4 chu3li3) 雜訊 (za2xun4)

TABLE 6. Average precision of word-level translation for short queries.

	Short English query	SA	SHF	SNHF	WCO
Average 11-point precision	29.85%	18.28%	19.57%	17.42%	21.78%
% of Monolingual	Baseline	61.24%	65.56%	58.36%	72.96%
% Change		Baseline	+7.06%	−4.70%	+19.15%

propose a real-time Web translator in MTIR, which is outlined here:

(1) Identifying the types of sentences.
(2) Searching the words in various dictionaries with the longest-matching strategy, including idiom and compound-word processing.
(3) Employing morpheme information, heuristic rules, and a HMM model to do part of speech tagging.
(4) Identifying noun phrase chunks by a partial parser.
(5) Doing lexical transfer and structural transfer.
(6) Generating the target sentences.

The following sections discuss each topic in detail.

4.1.1 Analysis Module

The analysis module analyzes the structure of the source sentence for the successive transfer module and synthesis module. At first, we identify the sentence types of source sentences using sentence delimiters. Some structural transfer rules can only be applied to some types of sentences. Then, the system performs the morphological analysis. The words in morphological forms (e.g., +ed, +ing, +ly, +s, etc.) are tagged with the morphological tags, which are useful for part-of-speech (POS) tagging, word sense disambiguation, and the generation of the target sense.

After morphological processing, the words in root form are searched from various dictionaries using the longest-matching strategy. There are about 67,000 word entries in an English-Chinese general dictionary and 5500 idioms in a phrasal dictionary. In addition, some domain-specific dictionaries are required for better translation performance. After dictionary lookup, the idioms and the compound words are treated as complete units for POS tagging and sense translation.

We adopted a hybrid method to deal with part-of-speech tagging. The method treated the certain cases using heuristic rules, and disambiguated the uncertain cases using a statistical model. To reduce the cost of fully parsing in a real-time service, a partial parser, which analyzed the tag sequence and identified the noun phrases, was employed to get the skeletons of the sentences. Then a predicate-argument detector (Chen & Chen, 1995) was used to extract the predicate-argument structures. The determination of PP attachment was based on rule templates (Chen & Chen, 1996).

4.1.2 Transfer Module

Lexical transfer and structural transfer touch on the differences between source and target languages. Idioms and compound words are treated as complete units during lexical selection. For the remaining words, several lexical selection algorithms like select-first, select-highest-frequent-word, and mutual information may be adopted. The select-first method always selects the first translation sense from the candidate list using parts of speech. The select-highest-frequent-word method also employs part of speech information, but chooses the target sense with the highest occurrence probability. The mutual information model considers the content around words to decide the best combination. Different models access various training tables. The

TABLE 7. Average precision of phrase-level translation for short queries.

	Short English query	SA	SHF	SNHF	WCO
Average 11-point precision	29.85%	23.36%	24.93%	22.92%	26.01%
% of Monolingual	Baseline	78.25%	83.52%	76.78%	87.14%
% Change		Baseline	+6.72%	−1.88%	+11.34%

TABLE 8. Average precision of word-level query translation on TREC-6 collection.

	Original English query (monolingual)	SA	SHF	SNHF	WCO
Average 11-point precision	14.49%	6.52%	8.65%	8.57%	9.78%
% of Monolingual	Baseline	45.00%	59.70%	59.14%	67.49%
% Change		Baseline	+32.67%	+31.44%	+50.00%

TABLE 9. The performance of person name translation.

Rank	1	2–5	6–10	11–15	16–20	21–25	25+
Number of items	524	497	107	143	44	22	197

larger the table is, the more time it takes. The next section discusses the time complexity, the table size, and the translation accuracy. In MTIR, the select-highest-frequent-word method is used. For structural transfer, structure-mapping rules are employed on the basis of predicate-argument structures detected in the analysis module.

4.1.3 Synthesis Module

The synthesis module deals with word insertion, deletion, and refinement. Those words carrying the morphological tags are processed according to the target sense generation rules. These rules are used to generate the Chinese translations for the possessives, the present particles, the past particles, the comparative adjectives, the superlatives, and so forth. For example, the target word 國家的 (guo2-jia1-de5) of the input word *nation's* is generated using the translation 國家 (guo2-jia1) of the root word *nation*. If the source word in the morphological form (ADJ+ly) is tagged as an adverb and derived from the adjectival root form, the target sense is generated in the way of deleting the character 的 (de5) and appending 地 (di3). In addition, the character 的 (de5) is inserted into the target words if the present participles and the past participles are tagged as adjectives.

Further, the translation results are presented in a bilingual aligned form. We employ the information of HTML tags. The HTML elements that appear in the document body fall into one of two groups: block-level elements and text-level elements. The former introduces paragraph breaks, and the latter does not. Common block-level elements include H1 to H6 (headers), P (paragraphs), LI (list items), and HR (horizontal rules). The TITLE (document title), TABLE (tables), and the FORM (forms) elements have the same effect. An English-Chinese bilingual document can be generated and aligned using the HTML block-level tags. Users can read both the English and the Chinese blocks simultaneously. The bilingual aligned document is a better representation scheme when both the translation performance and the speed performance are considered.

4.2 Performance Evaluation and Translation Effectiveness

MTIR was opened to Internet users in July 1997. Every translation result is kept in the log, so that we can study the behavior of the MTIR system. About 100,000 Web pages were translated in the last four months of 1997 and were analyzed. In the following sections, we discuss the statistical information, the speed performance, and the translation effectiveness.

4.2.1 Speed Performance

Tables 10 and 11 show the quantitative measurements. Table 10 illustrates the statistical information for the average size of the Web pages, the interactions between HTML and MT modules, the HTML tags, and the different kinds of components in the Web pages. On the average, there are 36.53 translation segments, 127.19 block-level tags, 96.72 font-level tags, and 29.41 anchors in a Web page. A bilingual aligned document is generated using the HTML block-level tags.

The overall speed performance depends on the communication, HTML analyzer, and MT subsystem. For the consideration of the online and real time issues, the highest-frequency-word is adopted for the word selection module. Table 11 shows the average processing time for each subtask of the MT system and the other two modules on a SUN SPARC station 5. On the average, the search engines take approximately 10 to 20 seconds to process a request for the Internet users. In our system, the average communication time to fetch the requested URL is 44.19 seconds and 7.81 percent of requested Web pages time out (exceed 300 seconds). The transmission rate is 200.43 bytes/second.

Because one of the major costs in online Web translation systems is the data retrieval on the Internet, we added a cache module to the communication subsystem to store the translated pages on the same day. This approach reduced the transmission time and the daily redundant work to retrieve and translate the same URLs for different users. The cache hit is near 36.52 percent, so that the average response time is reduced from 50 seconds to 32 seconds. Recently, a faster proxy system was used to fetch the Web pages and the average communication time was reduced to near 20 seconds.

TABLE 10. Quantitative study of web translation for about 100,000 web pages: statistical information for web pages.

		Numbers of HTML tags			Content						
Size (bytes)	Call MT (numbers of quasi-sentences)	Block-level tags	Font-level tags	Anchors	Words	Punctuation marks	Special codes (&code)	E-mails	URLs	Hosts	IPs
7037.80	36.53	127.19	96.72	29.41	308.30	101.80	0.12	0.21	0.37	1.43	0.20

TABLE 11. Quantitative study of web translation for about 100,000 web pages: Speed performance of communication, HTML analyzer, and MT subsystem (in seconds).

	MT module									
	Tagging by				Transferring					
Searching	Morpheme	Rules	HMM	Partial parsing	Structural transfer	Tense refine	Word selection	Word sense generation	HTML + MT modules	Communication
2.03	0.01	0.01	1.31	0.01 3.40	0.00	0.00	0.02	0.01	5.67	44.19

4.2.2 Translation Effectiveness

To determine translation quality, we evaluated lexical selection components for the top 30 Web sites (shown below) accessed by users.

1. www.yahoo.com
2. www.microsoft.com
3. www.geocities.com
4. www.playboy.com
5. www.cnn.com
6. www.nba.com
7. Search.yahoo.com
8. www.hotmail.com
9. www.disney.com
10. www.usnews.com
11. www.netscape.com
12. www.intel.com
13. www.square.co.jp
14. www.pathfinder.com
15. www.ibm.com
16. mac205.sjdccd.cc.ca.us
17. home.netscape.com
18. www.cmu.edu
19. events.yahoo.com
20. www.shareware.com
21. www-nlpir.nist.gov
22. www.real.com
23. www.petersons.com
24. www.windows95.com
25. www.nike.com
26. nssdc.gsfc.nasa.gov
27. www.infoseek.com
28. www.gravis.com
29. www.westwood.com
30. www.apple.com

Three word-selection methods are explored to evaluate the space requirement, speed performance and translation quality. Tables 12 and 13 list the statistical information. A comparison of Tables 10 and 12 discloses that:

(1) The average size of the home pages in the evaluation set is larger because they contain more HTML tags and Java scripts for creative presentations than other Web pages.
(2) In general, the home pages in the evaluation set introduce users to the information and then provide the users with many links. The number of words in these home pages are fewer than the number of words in other Web pages.

The processing time of each task is listed in Table 13. Because the total number of words is smaller, these sites take less translation time. On average, the HTML analyzer and MT subsystem take 4.66 seconds to translate an HTML file. Additionally, these 30 WWW sites are larger sites and have broader bandwidths for users to access. Although these home pages are larger than the other Web pages, the average fetch time (18.12 seconds) is smaller than that (44.19 seconds) in Table 11.

The cost for the lexical selection is discussed based on the factors of time complexity, space requirements, and translation accuracy. Three different statistical models—select first (Model 1), select the highest frequency word (Model 2), and word bigram in target language (Model 3)—are evaluated. For computing the translation accuracy, the translated sense is checked and given a score from 0 to 5 for each word. The highest score (5) indicates that the translated sense is correct. A score of 0 indicates that the correct sense of word doesn't appear in the dictionary. Other scores (1 to 4) indicate the different acceptance levels of translation results. The higher the score is, the more acceptable the translation is. Because the correct senses of 5.48 percent of words are not recorded in the dictionary, we omitted the incomplete cases in the evaluation. The overall

TABLE 12. Further study of the translation of the top 30 web sites: Statistical information for web pages.

		Numbers of HTML tags			Content						
Size (bytes)	Call MT (numbers of quasi-sentences)	Block-level tags	Font-level tags	Anchors	Words	Punctuation marks	Special codes (&code)	E-mails	URLs	Hosts	IPs
8888.76	32.64	163.20	88.48	36.32	188.68	44.12	0.00	0.28	0.12	1.52	0.00

TABLE 13. Further study of the translation of the top 30 web sites: Speed performance of communication, HTML analyzer, and MT subsystem (in seconds).

	MT module									
	Tagging by				Transferring					
Searching	Morpheme	Rules	HMM	Partial parsing	Structural transfer	Tense refine	Word selection	Word sense generation	HTML + MT modules	Communication
1.30	0.01	0.01	1.30	0.01 2.65	0.00	0.00	0.01	0.00	4.66	18.12

results are illustrated in Table 14. In model 2, 85.37 percent of words can be translated correctly. Hence, the translation accuracy of model 2 is 85.37 percent. The correct senses of 5.48 percent of words cannot be found in dictionaries. It forms 37.44 percent of errors. If we neglect the cases that the correct senses are not found in dictionary, the average scores are 3.75, 4.84, and 4.76 for different models.

The different methods employ different training tables to estimate the translation probabilities. Table 15 lists the table size and the time complexity of different word selection methods. Model 1 is the simplest model. It needs less space and translation time, but its translation performance is worse. Model 3, the most complex model, employs the sense association to decide the word meanings of the whole sentence using dynamic programming. It needs to access a bigram table of Chinese words, which has 884,324 records. This method takes about 49 seconds to get the translation sequence with the maximum likelihood, but most of time is spent on I/O and only 15 percent of processing time is used by CPU. Nevertheless, the accuracy of the word selection is slightly lower than the Model 2. Model 2 uses 8.33 megabytes to store a training table. To speed the processing of Model 2, the target senses of words are sorted by their frequencies. In this way, it achieves the same efficiency as Model 1 and has higher translation accuracy than Model 1.

TABLE 14. Evaluation results of the lexical selection.

	Correct (accuracy)	Correct sense is not found	Incorrect	Average score
Model 1	62.12%	5.48%	32.37%	3.75
Model 2	85.37%	5.48%	9.15%	4.84
Model 3	81.53%	5.48%	12.99%	4.76

TABLE 15. Table size and time complexity.

Evaluation	Table size (space)			Speed (time) in seconds
Word selection	Entries	Total frequency	(MB)	
Model 1	none	none	0.00	0.01
Model 2	94,531	2,433,670	8.33	0.01
Model 3	884,324	2,147,571	80.05	48.72

5. Concluding Remarks

This paper considers the retrieval and browsing operations in a cross-language information access system and touches on the corresponding query translation and document translation problems. A Chinese-English dictionary and an English corpus are employed to disambiguate the meaning of words. The monolingual corpus-based approach reduces the availability problem of large-scale and various domain bilingual corpora. Our method also resolves parts of the compound-word translation problem. Machine transliteration technology is introduced to name searching. It achieves some degree of performance but the performance improvement and the integration with named entity extraction have to be studied.

For online and real-time document translation, we evaluated the speed performance and translation effectiveness at the word level. The performance of lexical selection is good. Because of real-time requirements, we adopted rough parsing instead of fully parsing in the current version. The structure information cannot be captured completely, so that the structure transfer part has to be improved. Even so, users gave us a 67.74 percent satisfaction rating when questioned. We are extending the methodologies developed in this paper to study other cross-language information access tasks like multilingual information extraction and summarization.

References

Baker, K., et al. (1994). Coping with ambiguity in a large scale machine translation system. In Proceedings of 15th International Conference on Computational Linguistics, Kyoto, Japan (pp. 90–94).

Ballesteros, L., & Croft, W.C. (1996). Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications (pp. 791–801).

Ballesteros, L., & Croft, W.C. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of ACM SIGIR '97 (pp. 84–91).

Bennett, W., & Slocum, J. (1985). The LRC machine translation system. Computational Linguistics, 11(2-3), 111–119.

Bian, G.W., & Chen, H.H. (1997). An MT meta-server for information retrieval on WWW. In Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web, Palo Alto, CA (pp. 10–16). URL: http://mtir.csie.ntu.edu.tw/gwbian/English_Version/Research/Publications/publications.htm

Bian, G.W., & Chen, H.H. (1998). A new hybrid approach for Chinese-English query translation. In Proceedings of First Asia Digital Library

Workshop, Hong Kong, (pp. 156–167). URL: http://mtir.csie.ntu.edu.tw/gwbian/English_Version/Research/Publications/publications.htm
Brown, P., et al. (1990). A statistical approach to machine translation. Computational Linguistics, 16(2), 79–85.
Chen, K.H., & Chen, H.H. (1995). Machine translation: An integrated approach. In Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven, Belgium (pp. 287–294).
Chen, K.H., & Chen, H.H. (1996). A rule-based and MT-oriented approach to prepositional phrases attachment. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark (pp. 216–221).
Chen, H.H., Ding, Y.W., Tsai, S.C., & Bian, G.W. (1998). Description of NTU system used for MET2. In Proceedings of 7th Message Understanding Conference, Fairfax, VA.
Chen, H.H., Huang, S.J., Ding Y.W., & Tsai, S.C. (1998). Proper name translation in cross-language information retrieval. In Proceedings of 17th International Conference on Computational Linguistics, Montreal, Canada (pp. 232–236).
Chen, H.H., & Lee, J.C. (1996). Identification and classification of proper nouns in Chinese texts. In Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark (pp. 222–229).
David, M.W. (1996). New experiments in cross-language text retrieval at New Mexico State University's computing research laboratory. In Proceedings of the Fifth Text Retrieval Evaluation Conference. Gaithersburg, MD: National Institute of Standards and Technology.
David, M.W., & Dunning, T. (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval. In Proceedings of the Fourth Text Retrieval Evaluation Conference. Gaithersburg, MD: National Institute of Standards and Technology.
Dumais, S.T., Littman, M.L., & Landauer, T.K. (1997). Automatic cross-language retrieval using latent semantic indexing. In Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval (pp. 18–24).
Euro-Marketing Associates (1999). Global Internet statistics (by language). Euro-Marketing Associates. URL: http://www.euromktg.com/globstats/
Fitzpatrick, L., & Dent, M. (1997). Automatic feedback using past queries: Social searching? In Proceedings of ACM SIGIR '97 (pp. 306–313).
Grimes, B.F. (Ed.) (1996). Ethnologue: Languages of the world, 13th edition. Dallas, Texas: Summer Institute of Linguistics. URL: http://www.sil.org/ethnologue/
Harman, D.K. (Ed.) (1997). TREC-6 proceedings. Gaithersburg, MD.
Hershman, T. (1998, June). Real-time web language translator. Byte, 5–10.
Hull, D.A., & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of ACM SIGIR '96 (pp. 49–57).
Landauer, T.K., & Littman, M.L. (1990). Fully automatic cross-language document retrieval. In Proceedings of the Sixth Conference on Electronic Text Research (pp. 31–38).
Lu, S. (1995). A study on the Chinese romanization standard in libraries. Cataloging and Classification Quarterly, 21, 81–97.
Mitamura, T., Nyberg, E., & Carbonell, J. (1991). An efficient interlingua translation system for multilingual document production. Proceedings of Machine Translation Summit III, Washington, DC.
Nagao, M. (1984). A framework of mechanical translation between Japanese and English by analogy principle. In Elithorn, A. (Ed.). Artificial and Human Intelligence (pp. 173–180).
Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 5(24), 513–523.
Thompson, P., & Dozier, C. (1997). Name searching and information retrieval. In Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island.