Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System

Guo-Wei Bian and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
Email: gwbian@nlg.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw
http://nlg3.csie.ntu.edu.tw

Abstract. Due to the explosive growth of the WWW, very large multilingual textual resources have motivated the researches in Cross-Language Information Retrieval and online Web Machine Translation. In this paper, the integration of language translation and text processing system is proposed to build a multilingual information system. A distributed English-Chinese system on WWW is introduced to illustrate how to integrate query translation, search engines, and web translation system. Since July 1997, more than 46,000 users have accessed our system and about 250,000 English web pages have been translated to pages in Chinese or bilingual English-Chinese versions. And the average satisfaction degree of users at document level is 67.47%.

1 Introduction
In the past few years, the World Wide Web (WWW) grows explosively and has become the most useful and powerful information retrieval and accessing system on the Internet. The WWW breaks the boundaries of countries and provides very large online documents (more than 10 million documents) in multiple languages. These multilingual textual resources have motivated the researches in Cross-Language Information Retrieval (CLIR) and online Machine Translation (MT) to build the multilingual information accessing system. Although a number of searching engines and information discovery systems have been introduced on the Internet for users to locate interesting and relevant information, the language barrier becomes the major problem for people to search, retrieve, and understand WWW documents in different languages. That decreases the dissemination power of the WWW to some extent.

To alleviate this barrier, some information providers and WWW servers keep multiple copies of their information in different languages for multilingual service. Due to the dynamic nature of the WWW environment, the provided information is updated frequently. This approach is involved with the data inconsistency problem and the management problem of multilingual documents. How to incorporate the capability of language translation into WWW becomes indispensable for multilingual service. Recently, several online machine translation systems [1-4] have been

D. Farwell et al. (Eds.): AMTA ’98, LNAI 1529, pp. 250–265, 1998.
© Springer–Verlag Berlin Heidelberg 1998
presented. Traditional MT systems cannot be employed to the WWW directly because they are usually used to translate the written documents in the off-line batch mode. Translation quality is the most important criterion. In on-line and real-time applications, speed performance is also an important factor.

In this paper, we will focus on the following problems to alleviate the language barrier on WWW:

1. Language translation techniques
2. The integration of language translation system and text processing system

The language translation system is proposed to incorporate with the different kinds of text processing systems (e.g., searching engines, text summarization systems, etc.). A system integrated MT and IR technologies for WWW (abbreviated as MTIR) is introduced to illustrate our solutions for the mentioned problems. Section 2 describes a general model of the multilingual information system and introduces the architecture of our bilingual English-Chinese system for WWW. We discuss the Chinese-English query translation in section 3. Section 4 specifies how to integrate the query translation of CLIR with several searching engines on WWW. Section 5 describes the online and real-time web translation. Section 6 makes evaluations for such a multilingual information system from different users’ viewpoints. Section 7 concludes the remarks.

2 Multilingual Information System

Some of multilingual requirements for computer systems are shown as follows:

1. Data Representation: character sets and coding systems
2. Data Input: input methods and transliterated input
3. Data Display and Output: font mapping
4. Data Manipulation: the application must be able to handle the different coding characters
5. Query Translation: to translate the information need of users
6. Document Translation using Machine Translation (MT): to translate documents

The first three requirements have been resolved by system applications in several computer operating systems. Some of applications and packages can also handle both single-byte and multiple-byte coding systems for Indo-European and Eastern-Asian languages. However, the language barrier becomes the major problem for people to access the multilingual documents. How to incorporate the capability of language translation to meet the requirements 5 and 6 becomes indispensable for multilingual systems.

2.1 Four-Layer Multilingual Information System (MLIS)

Fig. 1 shows a four-layer multilingual information system. We put the different types of processing systems on the four layers:

- Layer 1: Language Identification (LI)
- Layer 2: Text Processing Systems
- Level 3: Language Translation Systems
- Level 4: User Interface (UI)
Fig. 1. A Four-Layer Model of Multilingual Information System (MLIS)

Fig. 2. The Overall Architecture of MTIR System
Because most of natural language processing techniques (e.g., lexical analysis, parsing, etc.) are dependent on the language of processed document, the layer 1 resolves language identification problem before text processing. The language identification system employs cues from the different character sets and coding systems of languages. At layer 2, the systems may perform information extraction, information filtering, information retrieval, text classification, text summarization, or other text processing tasks. Some of the text processing systems may have interaction with another one. For example, the relevant documents retrieved by IR system can be summarized to users. Additionally, a multilingual text processing system should be able to handle the different coding characters to match the requirement 4 (data manipulation). Several searching engines (e.g., AltaVista, Infoseek, etc.) have the ability to index the documents of multiple languages. The language translation systems at layer 3 are used to translate the information need of users for text processing systems and translate the resultant documents from text processing systems to users in their native languages. The user interface is the closest layer to users. It gets the user’s information need (included parameters, query and user profile) and displays the resultant document to user.

2.2 Bilingual English-Chinese Information System for WWW

On the WWW, 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 gets the resultant document from the server system. Fig. 2 shows the basic architecture of MTIR system. 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 query translation mechanism. The translations of query terms are disambiguated using word co-occurrence relationship. Then the system sends the translated query to the searching engine that selected by user in the user interface. The query subsystem takes care of the user interface part.

The subsequent navigation on the WWW is under the control of the communication subsystem. To minimize the traffic of Internet, 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. It divides the whole file into several translation segments for the machine translation subsystem. The HTML tags such as title, headings, unordered lists, ordered lists, definition lists, forms and tables play the similar roles of punctuation marks like full stop, question mark and exclamation mark. In contrast to the above tags, the font style elements, e.g., bold, italic, superscripts, subscripts, and font styles, 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 the hyperlinks. We attach our system’s URL to those URLs that link to HTML files or text files. Such a way guarantees the successive browses are linked with our system. The other URLs, including inline images and external MIME objects, are changed into their absolute URLs. In other words, the non-textual information is received from the original servers. Our experimental system is accessible with the following URL:
http://mtir.csie.ntu.edu.tw

3 Query Translation
Several approaches have been proposed for CLIR recently. There are four main approaches for query translation:
1. Dictionary-based approach [5-8]
2. Corpus-based approach [9-10]
3. Hybrid approach (combined dictionary-based and corpus-based) [6]
4. Machine Translation based approach (MT-based) [11]

Because the large parallel Chinese-English corpora are not available, the dictionary-based approach is adopted in our system. The query translation for Chinese-English CLIR consists of three major steps:
1. Word segmentation: To identify the word boundary of the input stream of Chinese characters.
2. Query translation: To construct the translated English query using the bilingual dictionary. The translation disambiguation is done using the monolingual corpus.
3. Monolingual IR: To search the relevant documents using the translated queries.

The segmentation and the query translation use the same bilingual dictionary in this design. That speeds up the dictionary lookup and avoids the inconsistencies resulting from two dictionaries (i.e., segmentation dictionary and transfer dictionary). This bilingual dictionary has approximately 90,000 terms. The longest-matching method is adopted in Chinese segmentation. The segmentation processing 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 some of the translation equivalents for each query term from the bilingual dictionary. The terms of query can be translated in two different levels of dictionary translations: word-level (word-by-word) and phrase-level translations. Those terms, missing from the transfer dictionary, are passed unchanged to the final query.

3.1 Selection Strategies
When there is more than one translation equivalent in a dictionary entry, the following selection strategies are 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.

(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 sense in the target language corpus. So, the frequently-used transferring 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 this strategy, the strategy (2) always selects only one sense for each term.

4. Word co-occurrence (WCO): This method classifies words on the basis of their co-occurrence with other words. The translation of a query term can be disambiguated with the co-occurrence of its translation equivalents and other words’ equivalents. The mutual information (MI) of word pairs reflects the word association norms in one language. If two words x and y have probabilities \( P(x) \) and \( P(y) \), their mutual information [12] is defined to be

\[
I(x, y) = \log_2 \frac{P(x,y)}{P(x)P(y)}
\]

This method considers the content around the translation equivalents within the text collection to decide the best target equivalent. The mutual information of word pairs is trained using a window size 3 in the CACM text collection [13]. Totally, there are 247,864 word pairs.

Table 1 illustrates an example for different translation. The Chinese concept ‘奇异值分解’ (jiyi zhi fenjie) and its phrase-level translation ‘singular value decomposition’ are employed. Four translated representations using different selection strategies on the word-level translation is shown in Table 1 (a). Column 3 shows the translation equivalents in transfer dictionary for the query terms at word-level. Table 1 (b) lists the mutual information of some word pairs of translation equivalents. Most of word pairs have no co-occurrence relations. Considering the example, the equivalent ‘singular’ of the term ‘奇异’ (jiyi) has the largest MI score with all translation equivalents of other two words.

### 3.2 Experiments and Evaluations

In the following experiments, the word-level and the phrase-level translations are touched to demonstrate the problems from missing terminology and multi-term concepts. In addition, we will evaluate these selection strategies with the long and the short versions of queries. The short queries are used to simulate the behavior of our methods for WWW. The SMART information retrieval system [14] is utilized to measure the similarity of the query and each document using the vector space model. The query weights are multiplied by the traditional IDF factor. The test collection CACM is used to evaluate the performance of different selection strategies. This collection contains 3204 texts and 64 queries in English. Each query has relevance judgements. The average number of words in the query is approximately 20.

In order to test the effectiveness of query translation, we create the Chinese queries by manually translating the original English queries to Chinese ones. The Chinese queries are regarded as the input queries later. Each Chinese query is translated to four target queries using different selection strategies. The following experiments compare the retrieval performances of the four translated versions of
Chinese queries to the results of the original English queries. One example of the original English query, human translated Chinese version, and translated queries are shown in Table 2. It gives the segmented Chinese string and four automatically translated representations for the CACM Q1. Parentheses surround the English multi-term concepts and the brackets surround the translation equivalents of each term.

To compare the performances of the word-level translation and phrase-level translation, the CACM English queries are manually checked to find the multi-term concepts that are not contained in our bilingual dictionary. These concepts and their translations are added into the bilingual dictionary for the phrase-level experiments. Totally, 102 multi-word concepts (e.g., remote procedure call (遠端程序呼叫), singular value decomposition (奇异值分解), etc.) are identified in the CACM queries.

Table 1. Different translations of Chinese concept ‘奇异值分解’ (singular value decomposition)

| Term | POS | SA          | SHF | SNHF | WCO          |
|------|-----|-------------|-----|------|--------------|
| 奇异 (jiyi) | N   | 奇异性 | 奇异性 | 奇异性 | 奇异性 |
| 值 (zhi)     | N   | 值 | 值 | 值 | 值 |
| 分解 (fenjie) | N   | 分解 | 分解 | 分解 | 分解 |
| V          | Decomposition analysis | dissociation cracking | disintegration | analyze | analyze | decomposition |
| XV        | (split up) (break up) | analyze | decompose | disassemble | dismount | resolve | (split up) |

Table 1(b). The mutual information for some word pairs

| word | Equivalents | 奇异 (jiyi) | 值 (zhi) | 分解 (fenjie) |
|------|-------------|----------|--------|-------------|
| 奇异 (jiyi) | oddity | w11 | w12 | w13 | w21 | w22 | w31 | w32 | w33 | w34 | w35 | w36 |
| 值 (zhi) | value | w21 | 6.099 | 4.115 | 6.669 | 4.377 | 1.823 | 4.377 |
| 分解 (fenjie) | analysis | w31 | 4.115 | 1.823 | 6.669 | 4.377 | 6.669 | 4.377 |
Table 2. The Chinese query and four translated representations for CACM Q1

| Original Query                                                                 | Chinese Query                                                                 |
|-------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| What articles exist which deal with TSS ‘Time Sharing System’, an operating system for IBM computers? | 那些文章是有关 TSS ‘分時系統’, 一種 IBM 電腦的作業系統 |
| 1 Segmentation                                                                | 那些文章是有关 TSS ‘分時系統’, 一種 IBM 電腦的作業系統 |
| 2.1 SA those article [be yes yah yep] about TTS ‘[minute cent apportion deal disserver sharing] time [formation lineage succession system]’, [a ace mono] [class seed] IBM [computer computing] of [(operating system) (operation system) OS] | 2.1 SA those article [be yes yah yep] about TTS ‘[minute cent apportion deal disserver sharing] time [formation lineage succession system]’, [a ace mono] [class seed] IBM [computer computing] of [(operating system) (operation system) OS] |
| 2.2 SHF those article be about TTS ‘deal time system’, a class IBM computer of (operating system) | 2.2 SHF those article be about TTS ‘deal time system’, a class IBM computer of (operating system) |
| 2.3 SNHF those article [be yes] about TTS ‘[minute deal] time system’, [a mono] class IBM computer of [(operating system) OS] | 2.3 SNHF those article [be yes] about TTS ‘[minute deal] time system’, [a mono] class IBM computer of [(operating system) OS] |
| 2.4 WCO those article be about TTS ‘sharing time system’, a class IBM computer of (operating system) | 2.4 WCO those article be about TTS ‘sharing time system’, a class IBM computer of (operating system) |

Over a wide range of operational environments, the average terms of user-supplied queries are 1.5 ~ 2 words and rarely more than 4 words. Hull and Grefenstette [7] work with the short versions of queries (average length of seven words) from French to English in TREC experiments. But no comparison of the short and long queries is available. To evaluate the behavior of user’s short queries, we make additional experiments to compare with the results of the original long queries. Three researchers help us to create the English and Chinese versions of short queries from the original English queries of CACM. For example, the short version of CACM Q1 is “TSS Timing Sharing System”. On the average, the short query has near 4 words, including single-word terms and multi-term concepts. The short version of English queries is regarded as the baseline to compare the results of translated queries of the short Chinese queries.

The overall results are shown in Fig. 3. The 11-point average precision [15] of the monolingual short English queries is 29.85%. It achieves the 83.42% performance of the original English queries. In word-level experiments, the best WCO (word co-occurrence) strategy gets the 72.96% performance of the monolingual English short version and 65.18% of the monolingual original English version. In phrase-level, the WCO achieves 87.14% and 74.71% respectively. The SHF, SNHF, and WCO selection strategies perform better in the long queries than that in short ones. However, the simple SA strategy has opposite result. Because users give more specific terms in short queries, the SA strategy introduces less extraneous terms to the query. Alternatively, the phrase-level translation improves the performance up to 14~31% over the word-level translation for Chinese-English CLIR. Combining the phrase dictionary and co-occurrence disambiguation can bring the performance of CLIR up to 87% of monolingual retrieval in short queries. Recall that the multi-word concepts and their translations are added to the dictionary in our experiments after the domain experts have examined the queries. Hence the coverage of bilingual phrasal dictionary will affect the performance of CLIR. Even though the bilingual dictionary does not contain these multi-word concepts, the WCO method still achieves near 70% monolingual effectiveness for different length of query at word-level translation.
Search Engines
Six popular search engines are integrated with language translation in our MTIR system. User inputs query and selects one of the search engines in the user interface. The Chinese query terms will be translated to English ones. After the processing of query translation, our system will send an HTTP request composed of the translated query to the chosen search engine. The retrieved results from the search engine will be translated to the user’s native language (Chinese). In general, the CGI program of searching engine processes the HTTP request of query. For instance, assuming “machine translation” is the translated query of the Chinese query “機器翻譯” (jīqì fānyì). The HTTP requests for the CGI programs of several search engines are listed in Table 3. The query words should be separated with the symbol ‘+’ for the standard URL encoding.

Hull and Grefenstette [7] give five different definitions for multilingual information retrieval. The type 4 is “IR on a multilingual document collection, where queries can retrieve documents in multiple languages”. How to merge and rank the retrieved documents in different languages is a problem in CLIR. Among these systems, the AltaVista and Infoseek have indexed both the English and Chinese web pages. If a bilingual query (“機器翻譯+machine+translation”) is invoked, the two systems will list the relevant documents of both languages. However, the ranking for documents in different languages seems not good. It’s still a problem for multilingual IR.

![Figure 3](image-url)
Table 3. HTTP Requests for the CGI Programs of Searching Engines

| Search Engine  | HTTP Requests for the CGI Programs of Searching Engines | Chinese Indexing |
|----------------|--------------------------------------------------------|------------------|
| AltaVista      | http://www.altavista.digital.com/cgi-bin/query?pg=q&what=web &kl=XX&q=machine+translation&search.x=35&search.y=9 | Yes              |
| Excite         | http://search.excite.com/search.gw?search=machine+translation | No               |
| Infoseek       | http://www.infoseek.com/Titles?qt=machine+translation&col=WW &sv=IS&lk=noframes&nh=10 | Yes              |
| Lycos          | http://www.lycos.com/cgi-in/pursuit?matchmode=and&cat=lycos& query=machine+translation&x=30&y=4 | No               |
| MetaCrawler    | http://www.metacrawler.com/crawler?general=machine+translation &method=0&target=&region=0&rpp=20&timeout=5&hpe=10 | No               |
| Yahoo          | http://search.yahoo.com/bin/search?p=machine+translation | No               |

5 Document Translation

The requirement for an online machine translation system for users to navigate on 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 resources. That is, different users’ behaviors affect the requirements of machine translation systems.

From users’ viewpoint, a high-quality and high-speed online machine translation is required. However, several steps should be performed after a query is issued. It takes time for the transfer of the query, the query translation, the retrieval of the document satisfying the query, the transfer of the retrieved document and the document translation. How to find the tradeoff between the speed performance and the translation performance on the WWW is an important issue. Besides this issue, our previous work [1] addressed four other issues, including which material is translated, what roles the HTML tags play in translation, what form the translated result is presented in, and where the translation capability is implemented, to design online machine translation systems for the WWW.

Many different approaches to machine translation design have been proposed [16-21]. These include rule-based, example-based, statistics-based, knowledge-based, and glossary-based approaches. A hybrid approach [22] integrates the advantages of these approaches and tries to get rid of their disadvantages. A rule-based partial parsing method is adopted and the translation process is performed chunk by chunk. We follow this design strategy and consider the characters of web translation. The following sections depict the details of analysis, transfer and synthesis modules.

5.1 Analysis 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, we take a 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 tagging, word sense disambiguation, and the generation of the target sense using the sense of the root word.

After morpheme processing, the words in root forms are searched from various
dictionaries using the longest-matching strategy. There are about 67,000 word entries in an English-Chinese general dictionary and 5,500 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.

For consideration of the speed and robustness issues, a three-stage hybrid method is adopted to deal with part-of-speech tagging. It treats the certain cases using heuristic rules, and disambiguates the uncertain cases using a statistical model. At stage 1, the words with specific morphological tags can be tagged without ambiguities. For example, the word of the pattern ADJ+ly is tagged with RB. The tagging of some morphological words depends on the morphological tag and the POS of its root form. For example, if the dictionary tag of the root of a word (root+er) is JJ, then this word is an adjective. Otherwise, it is a noun. Besides, if a word does not have any morphological tags and has only one POS candidate in the dictionary, then the unique POS is assigned to this word. At stage 2, a pattern matching method that considers the morphological tags of the current and the next words, as well as the POS of the next word, is employed to do the POS tagging. Stage 3 deals with the remaining words, which have not been tagged up to now. A statistical bigram HMM model is followed to solve the uncertain cases.

To reduce the cost of fully parsing in a real-time service, we adopt a partial parser to get the skeletons of sentences. A NP/ADJP finite state machine (FSM) is used to segment the source sentence into a sequence of chunks. This FSM analyzes the tag sequence, and recognizes the fundamental noun phrases and adjective phrases in linear time. Then a predicate-argument detector is followed to analyze the skeleton of sentence [23]. The determination of PP attachment is based on the rule templates [24].

5.2 Transfer Module
The structural transfer, the tense transfer, and the lexical selections touch on the differences of source and target languages. The major structural transfers occur in the comparative clauses, the question sentences, and the modifications of noun phrases. The structure of noun phrases is left-recursion in English, but is right-recursion in Chinese. Due to the recursion in the noun phrases, the transferred target structure is treated as a whole chunk for the subsequent processing. For different tenses, the words “have” and “be” have different senses in Chinese.

Phrases and idioms are treated as complete units during lexical selection. A bilingual phrase dictionary is employed to produce phrase-by-phrase translation. For those remaining words, several word selection algorithms like select-first, select the highest-frequency word and mutual information method may be adopted to select the target sense. The select-first method always selects the first translation sense from the candidates with the matched POSes. The second method chooses the target sense with the highest occurrence probability, trained from a large-scale corpus of the target language. The mutual information model considers the content around the words to decide the best combination of target words. Different models access various training tables. The larger the table is, the more time it takes. Section 6 will discuss the time complexity, the table space and the translation accuracy.
5.3 Synthesis Module

The synthesis module deals with word insertion, deletion and word order refinement. For example, if the source word with morpheme tag YJB, is tagged as adverb (RB) and derived from the adjective (JJ) word form, the target sense will be generated in the way of deleting the character “θθ” (de) and appending “θθ” (di). The character “θθ” (de) always appears at the end of Chinese adjectives, and the character “θθ” (di) at the end of adverbs. In addition, if the present participle and the past participle are tagged as adjective. The character “θθ” (de) is inserted into the target sense.

Our previous work [1] introduced the generation of bilingual aligned document for web translation system. A bilingual document can be generated and aligned using the HTML block-level tags. Users can read both the English and the Chinese blocks simultaneously. Bilingual aligned document is a better representation scheme when both the translation performance and the speed performance are considered.

6 Evaluation for Multilingual Information System

The implemented system has been opened to Internet users. We analyze each subsystem of MTIR and measure the quantitative evaluation results in 100,000 translated web page during the last four months of 1997. In such a large experimental resource, a web page has 308.30 words and 101.80 punctuation marks on the average. Total 14.08% of words are in morphological forms, and their root words can be derived from the morphological analyzer. Most of morphological processing (82.21%) is done by the morphological rules and the else is done using the morphological dictionary. Excepting numbers and punctuation marks, the words make up 78.66% of a web page’s content. Table 4(a) illustrates the statistical information for the average size of the web pages, the interactions between HTML and MT modules, the HTML tags and the content. Importantly, 11.87% of a web page are unknown words. For example, GeoCities, ICQ, NT, CERN, ASEAN, WinZip, DNS, Pentium, RealAudio, Newswire, QuickTime, NSSDC, AltaVista, Cybertown, MicroSim, HomeBanking, W3C, Hotmail, Website, CNET, ZDNet, RealVideo, Perl, BIOS, AOL, GeoPlus, Win95, and CGI are the terms extracted from the web pages. Most of these unknown words are production names, proper names, technological terms, and web sites.

The overall speed performance depends on the communication, HTML analyzer, and MT subsystem. For the consideration of online and real-time issue, the highest-frequency-word method is adopted for the word selection module. Table 4(b) shows the average processing time for the each subtask of the MT system, and the other two modules on the SUN SPARC station 5. In our system, the average communication time to fetch the requested URL (document) is 44.19 seconds. And 7.81% of requested web pages is time out (exceeding 300 seconds). Recently, a faster proxy system is used to fetch the web pages and the average communication time is reduced to near 20 seconds. After the web pages are fetched, the HTML analyzer parses the HTML structures and calls the MT system to translate the content. On average, these two subsystems take 5.67 seconds to translate an HTML file. In the following, we will discuss the time complexity and the translation quality for the major tasks in MT subsystem.
(a). Statistical Information for Web Pages and HTML Tags

| Size (Bytes) | Call MT (numbers of quasi-sentences) | Numbers of HTML Tags | Content |
|--------------|--------------------------------------|----------------------|---------|
|              | 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 |

(b). Speed Performance of Communication, HTML Analyzer, and MT Subsystem (in seconds).

| MT Module | HTML + MT modules | Communication |
|-----------|-------------------|---------------|
| Dictionary Accessing | Tagging by morpheme rules | Partial Parsing HMM | Transfering Structural | Tense | Word Refine | Selection | Synthesis | 5.67 | 44.19 |
| 2.03 | 0.01 | 0.01 | 1.31 | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 | 3.40 |

Table 5. Time Complexity and Translation Quality using Different Word Selection Methods

| Evaluation | Table Size (Space) | Speed (Time) in seconds | Translation Quality (Accuracy) |
|------------|---------------------|--------------------------|-------------------------------|
|            | Entries | Total Frequency | (MB) |                   |                          |
| Model 1    | none | none | 0.00 | 0.01 | 62.12% |
| Model 2    | 94,531 | 2,433,670 | 8.33 | 0.01 | 85.37% |
| Model 3    | 884,324 | 2,147,571 | 80.05 | 48.72 | 81.53% |

To be an online and real-time service of web translation on Internet, how much time users can endure is an important issue. However, most tasks of natural language processing have the problem of large time complexity. Some tradeoff between speed and quality must be done for the real-time NLP applications. The following shows some discussions. Because the translation quality depends on tagging and word selection in MT system, we evaluate these two major components for the top 30 WWW sites accessed by users of our system. And two additional word selection methods are explored to compare the performances of speed and quality. We assign four graduate and five undergraduate students to evaluate the translation results of these web sites.

With our 3-stage hybrid tagging method, the words of the web page are tagged within 2 seconds. Further, the accuracy of the 3-stage hybrid tagging method is 97.36%. Before the HMM stage, 86.21% of words can be assigned unique POS tags according to the morphological information and tagging rules. For a pure statistical HMM tagger, it spent 12.76 seconds and has 95.02% of accuracy on the average.
Besides, our partial parsing takes about 1 second to get the skeleton of the sentences of the web page. Comparatively, a full parser takes 25 seconds to analyze the sentences.

The cost for the lexical selection is discussed from the factors of time complexity, space requirement, and translation accuracy. Three statistical models, i.e., select-first, the highest-frequency word (word unigram in target language) and word bigram in target language, are evaluated. However, different methods need different training tables to estimate the probabilities. Table 5 lists the table size, the time complexity and the translation quality for different word selection methods. To speed up the processing of the model 2, the target senses of words are sorted by their frequencies. Then, the method 2 has the same efficiency as the method 1. The translation accuracy of model 2 is higher than the model 1 by using the frequently-used words. Model 3 (the more complex selection method) employs the sense association to decide the word meanings for the whole sentence using dynamic programming. It needs to access a bigram table of Chinese words, which is huge with 2,147,571 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% of processing time is used by CPU. Nevertheless, the accuracy of the word selection is lower than the model 2. For one web page, we have to lookup the table 1235.52 times on the average using the bigram model. With MI (mutual information) model, the average number of table accessing increases to 9026.04 for each web page. In other words, the MI model spends about 7 times more than the model 3. The MI model is not suitable for the real-time document translation on web.

Besides the evaluation on the word-level, we provide a questionnaire for Internet users to evaluate the effect of this web translation system on document level. The users can fill in a form containing a sequence of questions to describe their interests, suggestions, and satisfaction degrees about MTIR system. The degrees are varied from 0% to 100% with step 10%. Total 372 users answer the questions. The average satisfaction degree is 67.47%. The satisfaction has shown the importance of language translation in the multilingual information system.

7 Conclusion

The explosive growth of the WWW has brought very large multilingual textual resources to users. How to incorporate the technologies of natural language processing and text processing has shown very important in the information age. In this paper, we have proposed a general model of multilingual information system to integrate the text processing systems and language translation systems. A system integrated MT and IR technologies for WWW (MTIR system) has illustrated our solutions for multilingual services. This system can help users to access and retrieve documents on WWW in their native language(s). Additionally, the online and real-time web translation system can assist the users to understand the web pages during their navigation in the huge information resources. This multilingual system has been developed and evaluated. Several experiments for the query translations of CLIR have simulated and shown the applicability for short queries on WWW. A quantitative study of 100,000 web pages and the 30 top requested WWW sites have
reflected the importance of the tradeoff between speed and translation quality for document translation. Additionally, many contemporary terms and proper names can be extracted for dictionary refinement and other NLP researches [25] during users’ navigation on WWW using this system.

References
1. Bian, G.W. and Chen, H.H.: “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, California, USA, March, 1997, pp.10-16.
2. David, M.W. and Ogden, W.C.: “QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System.” In Proceedings of ACM SIGIR’97, 1997, pp.92-98.
3. Gachot, D.A.; Lange, E. and Yang, J.: “The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Multilingual Information Retrieval.” In Proceedings of Workshop on Cross-Linguistic Information Retrieval, 1996, pp. 44-54.
4. Hayashi, Y.; Kikui, G. and Susaki, S.: “TITAN: A Cross-linguistic Search Engine for the WWW.” In Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval, 1997, pp. 58-65.
5. Ballesteros, L. and Croft, W.C.: “Phrasal Translation and Query Expansion Techniques for Cross-Language Information retrieval.” In Proceedings of ACM SIGIR’97, pp.84-91, 1997.
6. David, M.W.: “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 (TREC-5), Gaithersburg, MD, National Institute of Standards and Technology, 1996.
7. Hull, D.A. and Grefenstette, G.: “Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval.” In Proceedings of ACM SIGIR’96, pp.49-57, 1996.
8. Kwok, K.L.: “Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment.” In Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval, 1997, pp. 110-114.
9. David, M.W. and Dunning, T.: “A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval.” In Proceedings of the Fourth Text Retrieval Evaluation Conference (TREC-4), Gaithersburg, MD, National Institute of Standards and Technology, 1995.
10. Landauer, T.K. and Littman, M.L.: “Fully Automatic Cross-Language Document Retrieval.” In Proceedings of the Sixth Conference on Electronic Text Research, pp. 31-38, 1990.
11. Radwan, K.: Vers l’Acces Multilingue en Langage Naturel aux Bases de Donnees Textuelles. PhD Thesis, Universite de Paris-Sud, Centre d’Orsay. 1994.
12. Church, K. and Hanks, P.: “Word Association Norms, Mutual Information and Lexicography,” Computational Linguistics, 16(1), 1990, pp. 22-29.
13. Fox, E., ed.: Virginia Disk One, Blacksburg: Virginia Polytechnic Institute and State University, 1990.
14. Salton, G. and Buckley, C.: “Term Weighting Approaches in Automatic Text
15. Rijsbergen, C.J. van: Information Retrieval, 2nd Edition, London, Butterworths, 1979.

16. Baker, K., et al.: “Coping with Ambiguity in a Large-scale Machine Translation System.” In Proceedings of COLING-94, 1994, pp. 90-94.

17. Bennett, W. and Slocum, J.: “The LRC Machine Translation System.” Computational Linguistics, 11(2-3), 1985, pp. 111-119.

18. Brown, P.; et al.: “A Statistical Approach to Machine Translation.” Computational Linguistics, 16(2), 1990, pp. 79-85.

19. Mitamura, T.; Nyberg, E. and Carbonell, J.: “An Efficient Interlingua Translation System for Multilingual Document Production.” In Proceedings of Machine Translation Summit III, 1991.

20. Nagao, M.: “A Framework of Mechanical Translation between Japanese and English by Analogy Principle.” Artificial and Human Intelligence, 1984, pp. 173-180.

21. Nirenburg, S., et al.: “Multi-purpose Development and Operations Environments for Natural Language Applications.” In Proceedings of Applied Language Processing, Trento, Italy, 1993.

22. Chen, K.H. and Chen, H.H.: “Machine Translation: An Integrated Approach.” In Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation, 1995, pp. 287-294.

23. Chen, K.H. and Chen, H.H.: “Acquisition of Verb Subcategorization Frames from Large Scale Texts.” In Proceedings of KONVENS94, 1994, pp. 407-410.

24. Chen, K.H. and Chen, H.H.: “A Rule-Based and MT-Oriented Approach to Prepositional Phrases Attachment.” In Proceedings of the 16th COLING, 1996, pp. 216-221.

25. Chen, H.H. and Bian, G.W.: “Proper Name Extraction from Web Pages for Finding People in Internet.” In Proceedings of the 10th Research on Computational Linguistics (ROCLING X) International Conference, 1997, pp. 143-158.