Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach

Jenq-Haur Wang[1], Jei-Wen Teng[1], Pu-Jen Cheng[1], Wen-Hsiang Lu[2], and Lee-Feng Chien[1,3]

[1] Institute of Information Science, Academia Sinica, Taiwan

[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

[3] Department of Information Management, National Taiwan University, Taiwan

{jhwang, jackteng, pjcheng, whlu, lfchien}@iis.sinica.edu.tw

ABSTRACT

Users' cross-lingual queries to a digital library system might be short and not included in a common translation dictionary (unknown terms). In this paper, we investigate the feasibility of exploiting the Web as the corpus source to translate unknown query terms for cross-language information retrieval (CLIR) in digital libraries. We propose a Web-based term translation approach to determine effective translations for unknown query terms by mining bilingual search-result pages obtained from a real Web search engine. This approach can enhance the construction of a domain-specific bilingual lexicon and benefit CLIR services in a digital library that only has monolingual document collections. Very promising results have been obtained in generating effective translation equivalents for many unknown terms, including proper nouns, technical terms and Web query terms.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries.

General Terms

Algorithms, Experimentation, Performance

Keywords

Cross-Language Information Retrieval, Digital Library, Term Extraction, Term Translation, Web Mining

1. INTRODUCTION

With the development of digital library technologies, large amounts of library content and cultural heritage material are being digitized all over the world. As digital library systems become commonly constructed and digitized content becomes widely accessible on the Web, digital libraries that cross language and regional boundaries will be in increasingly high demand globally. Unfortunately, most of existing digital library systems only provide monolingual content and search support in certain target languages. To facilitate a cross-language information retrieval (CLIR) service in digital library systems, it is important to develop a powerful query translation engine. This must be able to automatically translate users' queries from multiple source languages to the target languages that the systems accept.

Conventional approaches to CLIR incorporate parallel texts [16] as the corpus. These texts contain bilingual sentences, from which word or phrase translations can be extracted with appropriate sentence alignment methods [7]. The basic assumption of such an approach is that queries may be long so query expansion methods can be used to enrich query terms not covered in parallel texts. However, this approach presents some fundamental difficulties for digital libraries that wish to support practical CLIR services. First, since most existing digital libraries contain only monolingual text collections, there is no bilingual corpus for cross-lingual training. Second, real queries are often short, diverse and dynamic so that only a subset of translations can be extracted through the corpora in limited domains. How to efficiently construct a domain-specific translation dictionary for each text collection has become a major challenge for practical CLIR services in digital libraries. In this paper, we propose a Web-based approach to deal with this problem. We intend to exploit the Web as the corpus to find effective translations automatically for query terms not included in a dictionary (unknown terms). Besides, to speedup online translation process of unknown terms, we extract possible key terms from the document set in digital libraries and try to obtain their translations in advance.

For some language pairs, such as Chinese and English, as well as Japanese and English, the Web offers rich texts in a mixture of languages. Many of them contain bilingual translations of proper nouns, such as company names and personal names. We want to realize if this positive characteristic makes it possible to automatically extract bilingual translations of a large number of query terms. Real search engines, such as Google[1] and AltaVista[2], allow us to search English terms for pages in a certain language, e.g. Chinese or Japanese. This has motivated us to develop the proposed approach for mining bilingual search-result pages, which are normally returned in a long, ordered list of *snippets* of summaries to help users locate interesting documents. The proposed approach uses the bilingual search-result pages of unknown queries as the corpus for extracting translations by utilizing the following useful techniques: (1) Term extraction

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.

[1] http://www.google.com/

[2] http://www.altavista.com/

methods that extract translation candidates with correct lexical boundaries. (2) Term translation methods that determine correct translations based on co-occurrence and context similarity analysis.

Several preliminary experiments have been conducted to test the performance of the proposed approach. For example, very promising translation accuracy has been obtained in generating effective translation equivalents for many unknown terms, including proper nouns, technical terms and Web query terms. Also, it has been shown that the approach can enhance bilingual lexicon construction in a very efficient manner and thereby benefit CLIR services in digital libraries that only have monolingual document collections. In Section 2 of this paper, we examine the possibility of using search-result pages for term translation. The technical details of the proposed approach, including the term extraction and term translation methods, are presented with some experiments in Sections 3 and 4 respectively. An application of the proposed approach to bilingual lexicon construction is described in Section 5. Finally, in Section 6 we list our conclusions.

2. OBSERVATIONS AND THE PROPOSED APPROACH

A large number of Web pages contain a mixture of multiple languages. For example, Chinese pages on the Web consist of rich texts in a mixture of Chinese (main language) and English (auxiliary language), many of which contain translations of proper nouns and foreign terms. In fact, in the Chinese writing style, the first time a foreign term appears in the text, we might also write its original word, e.g., "雅虎" (Yahoo). In our research, we are seeking to determine if the percentage of correct translations for real queries is high enough in the top search-result pages. If this is the case, search-result-based methods can be useful in alleviating the difficulty of term translation. According to our observations, many query terms are very likely to appear simultaneously with their translations in search-result pages. Figure 1 illustrates the search-result page of the English query "National Palace Museum", which was submitted to Google to search Chinese pages. Many relevant results were obtained, including both the query itself and its Chinese aliases, such as "國立故宮博物院" (National Palace Museum), "故宮" (an abbreviation of National Palace Museum) and "故宮博物院" (Palace Museum), which might not be covered in general-purpose translation dictionaries.

Figure 1. An illustration showing translation equivalents, such as National Palace Museum/"國立故宮博物院" ("故宮"), which co-occur in search results returned from Google.

Although search-result pages might contain translations, the difficulties in developing a high-performance search-result-based term translation approach still remain. For example, it is not straightforward to extract translation candidates with correct lexical boundaries and minimum noisy terms from a text. It is also challenging to find correct translations for each unknown term within an acceptable number of search-result pages and an acceptable amount of network access time. To deal with these problems, the proposed approach contains three major modules: search-result collection, term extraction and term translation, as shown in Figure 2 (a). In the search-result collection module, a given source query (unknown term) is submitted to a real-world search engine to collect top search-result pages. In the term extraction module, translation candidates are extracted from the collected search-result pages using the term extraction method. Finally, the term translation module is used to determine the most promising translations based on the similarity estimation between source queries and target translations.

In fact there are two scenarios to which the proposed approach can be applied. Except online translation of unknown queries, another application is offline translation of key terms as in Figure 2 (b). To reduce unnecessary online translation processes, the proposed approach can be used to augment the bilingual lexicon via translating key terms extracted from the document set in a digital library. These extracted key terms are likely to be similar to terms that users may use in real user queries. The proposed approach can be applied to those unknown key terms to obtain their translations with an offline batch process (the extracted translations might be edited by indexers). Furthermore, the constructed bilingual lexicon can be incrementally updated with the input of unknown queries from users and the performing of online translation processes. To facilitate the above scenarios the proposed term extraction and term translation techniques are required, which will be further described in the following sections.

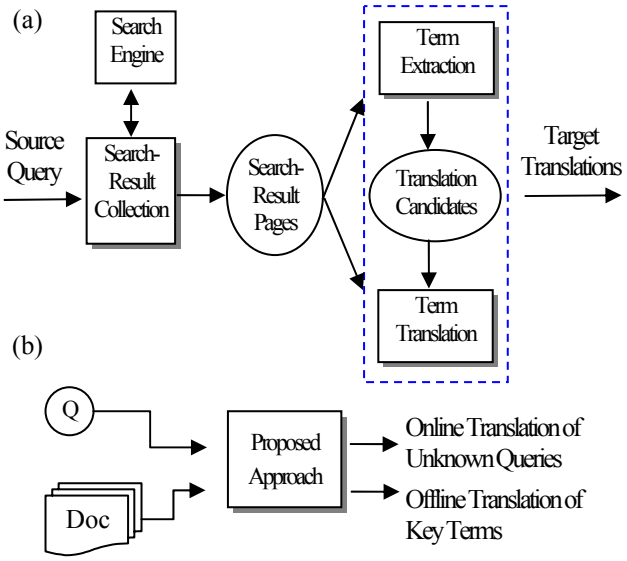

Figure 2. (a) An abstract diagram showing the concept of the proposed approach for translating an unknown query. (b) Two application scenarios of the proposed Web-based term translation approach: online translation of unknown queries and offline translation of key terms extracted from the document set.

3. TERM EXTRACTION

The first challenge of the proposed approach is: how to efficiently and effectively extract translation candidates for an unknown source term from a set of search-result pages. Other challenging issues include: whether all possible translations can be extracted and whether their lexical boundaries can be correctly segmented. Conventionally, there are two types of term extraction methods that can be employed. The first is the language-dependent linguistics-based method that relies on lexical analysis, word segmentation and syntactic analysis to extract named entities from documents. The second type is the language-independent statistics-based method that extracts significant lexical patterns without length limitation, such as the local maxima method [19] and the PAT-tree-based method [3]. Considering the diverse applications in digital library and Web environments, we have adopted the second approach. Our proposed term extraction method, i.e., the PAT-tree-based local maxima method, is a hybrid of the local maxima method [19] and the PAT-tree-based method [3], which has been found more efficient and effective. First, we construct a PAT tree data structure for the corpus, in this case, a set of search-result pages retrieved using the source term as query. (The same term extraction method will be applied to extract key terms from digital libraries in Section 5 where the corpus is the documents in digital libraries). By utilizing the PAT tree, we can efficiently calculate the association measurement of every character or word *n*-gram in the corpus and apply the local maxima algorithm to extract the terms. The association measurement is determined not only by the symmetric conditional probability [19] but also by the context independency ratio [3] of the *n*-gram. We detail the proposed method in the following subsections.

3.1 Association Measurement

The proposed association measurement, called *SCPCD*, combines the symmetric conditional probability (*SCP*) [19] with the concept of context dependency (*CD*) [3]. *SCP* is the association estimation of the correlation between its composed sub *n*-grams, which is as defined below:

$$SCP(w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \dots w_i)\, p(w_{i+1} \dots w_n)} = \frac{freq(w_1 \dots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1 \dots w_i)\, freq(w_{i+1} \dots w_n)} \quad (1)$$

where $w_1 \dots w_n$ is the *n*-gram to be estimated, $p(w_1 \dots w_n)$ is the probability of the occurrence of the *n*-gram $w_1 \dots w_n$, and $freq(w_1 \dots w_n)$ is the frequency of the *n*-gram.

To a certain degree, *SCP* can measure the cohesion holding the words together within a word *n*-gram, but it cannot determine the lexical boundaries of the *n*-gram. An *n*-gram with complete lexical boundaries implies that it tends to have free association with other *n*-grams appearing in the same context. Therefore, to further ensure that an *n*-gram has complete lexical boundaries, the concept of context dependency is introduced. Moreover, we consolidate the concept with *SCP* to form one association measurement. In order to achieve this goal, a refined measure, the context independency ratio - which is a ratio value between 0 and 1 - is extended from [3]. It is defined as follows:

$$CD(w_1 \dots w_n) = \frac{LC(w_1 \dots w_n) RC(w_1 \dots w_n)}{freq(w_1 \dots w_n)^2} \quad (2)$$

where $LC(w_1 \dots w_n)$ is the number of unique left adjacent words in western languages, or characters in oriental languages, for the *n*-gram in the corpus, or is equal to the frequency of the *n*-gram if there is no left adjacent word/character. Similarly, $RC(w_1 \dots w_n)$ is the number of unique right adjacent words/characters for the *n*-gram, or is equal to the frequency of the *n*-gram if there is no right adjacent word/character. Using this ratio we are able to judge whether the appearance of an *n*-gram is dependent on a certain string containing it. For example, if $w_1 \dots w_n$ is always a substring of string $xw_1 \dots w_n y$ in the corpus, then $CD(w_1 \dots w_n)$ is close to 0.

Combining formulae (1) and (2), the proposed association measure *SCPCD* is as follows

$$SCPCD(w_1 \dots w_n) = SCP(w_1 \dots w_n) * CD(w_1 \dots w_n) = \frac{LC(w_1 \dots w_n) RC(w_1 \dots w_n)}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1 \dots w_i)\, freq(w_{i+1} \dots w_n)} \quad (3)$$

Note that the difference between the formulae of *SCPCD* and *SCP* is in their numerator items. For *SCP*, those *n*-grams with low frequency tend to be discarded, which is prevented in the case of *SCPCD*. The proposed new measure determines a highly cohesive term because of the frequencies of its substrings and the number of its unique left and right adjacent words/characters.

3.2 Local Maxima Algorithm

The local maxima algorithm, called LocalMaxs in [18], is based on the idea that each *n*-gram has a kind of cohesion that holds the words together within the *n*-gram. This is a heuristic algorithm used to combine with the previous association measurements to extract *n*-grams, which are supposed to be key terms from the text. We know different n-grams usually have different cohesion values. Given that:

- An *antecedent* (in size) of the *n*-gram $w_1 w_2 \dots w_n$, $ant(w_1 \dots w_n)$, is a sub-*n*-gram of the *n*-gram $w_1 \dots w_n$, having size *n - 1*. i.e., the *(n-1)*-gram $w_1 \dots w_{n-1}$ or $w_2 \dots w_n$.
- A *successor* (in size) of the *n*-gram $w_1 w_2 \dots w_n$, $succ(w_1 \dots w_n)$, is a *(n+1)*-gram *N* such that the *n*-gram $w_1 \dots w_n$ is an *ant(N)*. i.e., $succ(w_1 \dots w_n)$ contains the *n*-gram $w_1 \dots w_n$ and an additional word before (on the left) or after (on the right) it.

The local maxima algorithm extracts each term whose cohesion, i.e. association measure, is local maxima. That is, the term whose association measure is greater than, or equal to, the association measures of its antecedents and is greater than the association measures of its successors.

3.3 The PAT-Tree Based Local Maxima Algorithm

Despite the usefulness of the local maxima algorithm, without a suitable data structure the time complexity of the algorithm is high. The main time complexity problems occur in two areas. One is calculating the context independency ratio (CD) for each unique *n*-gram in the corpus and the other is to find the successor of an *n*-gram. The two problems can be treated as one, i.e. finding the successors of an *n*-gram. An intuitive way to do this is to find out all *(n+1)*-grams and then compare the *n*-gram with them sequentially to see if they are the successors of it. As this is time-consuming, we introduce PAT tree as the data structure.

The above method is time consuming, however, so we use the PAT tree, which is a more efficient data structure. It was developed by Gonnet [8] from Morrison's PATRICIA algorithm (Practical Algorithm to Retrieve Information Coded in Alphanumeric) [15] for indexing a continuous data stream and locating every possible position of a prefix in the stream. The PAT tree structure is conceptually equivalent to a compressed digital search tree, but smaller. The superior feature of this structure mostly resulted from its use of semi-infinite strings [14] to store the substream values in the nodes of the tree. This also makes it easier and more efficient to find the successors of an *n*-gram. More details on the PAT tree can be found in [3].

By utilizing the constructed PAT tree as the corpus, we can efficiently retrieve all *n*-grams from the corpus, obtain their frequencies and context dependency values, and then calculate the association measures, *SCPCD*, of all of them.

3.4 Experiments on Term Extraction

To determine the effectiveness of the proposed association measure *SCPCD* and the efficiency of the PAT-tree data structure, we conducted several experiments on Web search-result pages using the proposed PAT-tree-based local maxima algorithm.

First, to test whether *SCPCD* can perform better than *SCP* and *CD,* we randomly selected 50 real queries in English from a Chinese search engine called Openfind[3]. We then submitted each of them to Google to search Chinese result pages. Most of these query terms such as proper nouns and technical terms were not covered in the common translation dictionary. After using the term extraction method, the top 30 extracted Chinese translation candidates were examined and the extraction accuracy of each candidate to the source query was manually determined. We applied this test mainly to determine whether the *SCPCD* measurement can extract more relevant translation candidates and segment them with correct lexical boundaries. A translation candidate was taken as correctly extracted only if it was correctly segmented and contained meanings relevant to the source term. A relevant translation candidate was not necessarily a correct translation. The whole relevant set was determined by examining the terms extracted by all of the test methods, e.g., *CD*, *SCP*, and *SCPCD*. Table 1 clearly shows that the method based on the *SCPCD* measurement achieves the best performance.

Table 1. The obtained extraction accuracy including precision, recall, and average recall-precision of auto-extracted translation candidates using different methods.

Association Measure	Precision	Recall	Avg. R-P
CD	68.1 %	5.9 %	37.0 %
SCP	62.6 %	63.3 %	63.0 %
SCPCD	79.3 %	78.2 %	78.7 %

In order to determine the efficiency of the PAT-tree data structure, we compared the speed performance of the local maxima method and the PAT-tree-based local maxima method. As Table 2 shows, the PAT-tree data structure is more efficient in term extraction. Although the PAT-tree construction phase took a little more time in a small corpus, in a real-world case for a large corpus - where 1,367 and 5,357 scientific documents were tested (refer to Section 5.2 for the details) - the PAT-tree-based local maxima method performed much better than the local maxima method.

Table 2. The obtained average speed performance of different term extraction methods.

Term Extraction Method	Time for Preprocessing	Time for Extraction
LocalMaxs (Web Queries)	0.87 s	0.99 s
PATtree+LocalMaxs (Web Queries)	2.30 s	0.61 s
LocalMaxs (1,367 docs)	63.47 s	4,851.67 s
PATtree+LocalMaxs (1,367 docs)	840.90 s	71.24 s
LocalMaxs (5,357 docs)	47,247.55 s	350,495.65 s
PATtree+LocalMaxs (5,357 docs)	11,086.67 s	759.32 s

4. TERM TRANSLATION

In the term translation module, we utilize the co-occurrence relation and the context information between source queries and target translations to estimate their semantic similarity and determine the most promising translations. Several similarity estimation methods were investigated based on co-occurrence analysis. These included mutual information, DICE coefficient, and statistical tests including the chi-square test and the log-likelihood ratio test [17, 20], where the chi-square test and the context vector analysis achieved the best performance. These will be introduced below.

4.1 The Chi-Square Test

The chi-square test (χ^2) was adopted as the major method of co-occurrence analysis in our study. One major reason is that the required parameters for the chi-square test can be effectively computed using the search-result pages, which alleviates the data sparseness problem. It also makes good use of all relations of co-occurrence between the source and target terms, especially the information that they do not co-occur. For source term *s* and target term *t*, the conventional chi-square test can be transformed as the similarity measure defined below [6]:

$$S_{\chi 2}(s,t) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \quad (4)$$

where

a: the number of pages containing both terms *s* and *t*;

b: the number of pages containing term *s* but not *t*;

c: the number of pages containing term *t* but not *s*;

d: the number of pages containing neither term *s* nor *t*;

N: the total number of pages, i.e., $N = a+b+c+d$.

Since most search engines accept Boolean queries and can report the number of pages matched, the required parameters for the chi-square test can be obtained by submitting Boolean queries such as '$s \cap t$', '$\sim s \cap t$', '$s \cap \sim t$' to search engines and utilizing the returned page counts. On the other hand, it is easy to get number *N* using some search engines (e.g., Google), which indicates the

[3] http://www.openfind.com/

total number of their collected Web pages. The number *d* may not be directly available from the search engine, but it can be calculated using the formula $N= a+b+c+d$, i.e., $d = N-a-b-c$.

4.2 Context Vector Analysis

Co-occurrence analysis is applicable to higher frequency terms since they are more likely to appear with their translation candidates. On the other hand, lower frequency terms have little chance of appearing with candidates on the same pages. The context vector method (CV) is therefore adopted to deal with this problem. As translation equivalents may share similar terms, for each query term, we take the co-occurring feature terms as the feature vector. The similarity between query terms and translation candidates can be computed based on their feature vectors. Thus, lower frequency query terms still have a chance to extract correct translations.

The context vector-based method has been used to extract translations from comparable corpora, such as the use of Fung et al.'s seed word [5]. In our method, real users' popular query terms are used as the feature set, which should help to avoid many inappropriate feature terms. Like Fung et al.'s vector space model, we also use the TF-IDF weighting scheme to estimate the significance of context features. This is defined as follows:

$$w_{t_i} = \frac{f(t_i, d)}{\max_j f(t_j, d)} \times \log(\frac{N}{\mathrm{n}}) \quad (5)$$

where $f(t_i,d)$ is the frequency of term t_i in search-result page d, N is the total number of Web pages in the collection of search engines, and n is the number of pages containing t_i. Given the context vectors of a source query term and each target translation candidate, their similarity is estimated with cosine measure as follows:

$$S_{cv}(s,t) = \frac{\sum_{i=1}^{m} w_{s_i} \times w_{t_i}}{\sqrt{\sum_{i=1}^{m} (w_{s_i})^2 \times \sum_{i=1}^{m} (w_{t_i})^2}} \quad (6)$$

It is not difficult to construct context vectors for source query terms and their translation candidates. For a source query term, we can use a fixed number of the top search results to extract translation candidates. The co-occurring feature terms of each query can also be extracted, and their weights calculated, which together form the context vector of the query. The same procedure is used to construct a context vector for each translation candidate.

4.3 The Combined Method

Benefiting from real-world search engines, the search-result-based method using the chi-square test can reduce the work of corpus collection, but has difficulty in dealing with low-frequency query terms. Although context vector analysis can deal with difficulties encountered by the chi-square test, it is not difficult to see that the feature selection issue needs to be carefully handled. Intuitively, a more complete solution is to integrate the above two methods. Considering the various ranges of similarity values in the two methods, we use a linear combination weighting scheme to compute the similarity measure as follows:

$$S_{all}(s,t) = \sum_m \frac{\alpha_m}{R_m(s,t)} \quad (7)$$

where α_m is an assigned weight for each similarity measure S_m, and $R_m(s,t)$ - which represents the similarity ranking of each target candidate t with respect to source term s - is assigned to be from *1* to k (the number of candidates) in decreasing order of similarity measure $S_m(s,t)$.

4.4 Experiments on Term Translation

4.4.1 The Test Bed

To determine the effectiveness of the proposed approach, we conducted several experiments to extract translation pairs for Chinese and English terms in different domains.

Web Queries: We collected query terms and the logs from two real-world Chinese search engines in Taiwan, i.e., Dreamer and GAIS. The Dreamer log contained 228,566 unique query terms for a period of over 3 months in 1998, while the GAIS log contained 114,182 unique query terms for a period of two weeks in 1999. We prepared two different test query sets based on these logs. The first, called the popular-query set, contained a set of 430 frequent Chinese queries in the logs. These queries were obtained from the Chinese translations of 1,230 English terms out of the most popular 9,709 query terms (with frequencies above 10 in both logs), which co-occurred with their English counterparts in the logs. The popular-query set was further divided into two types: type Dic (the terms covered in the dictionary), consisting of about 36% (156/430) of the test queries and type OOV (out of vocabulary; the terms not in the dictionary), consisting of about 64% (274/430) of the test queries.

The second set, called the random-query set, contained 200 Chinese query terms, which were randomly selected from the top 20,000 queries in the Dreamer log, where 165 (about 82.5%) were not included in general-purpose translation dictionaries.

Proper Names and Technical Terms: To further investigate the translation effectiveness for proper names and technical terms, we prepared two other query sets containing 50 scientists' names and 50 disease names in English. These were randomly selected from the 256 scientists (Science/People) and 664 diseases (Health/Diseases and Conditions) in the Yahoo! Directory. It should be noted that 76% (38/50) of the scientists' names and 72% (36/50) of the disease names were not included in the general-purpose translation dictionary, which contained 202,974 entries collected from the Internet.

To evaluate the search-result-based methods, we obtained search-result pages of the source query terms by submitting them to real-world Chinese search engines, such as Google Chinese and Openfind. Basically, we used only the first 100 retrieved results (*snippets*) to extract translation candidates. The context vector of each source query and the required parameters (page counts) for the chi-square test were also extracted from the retrieved search-result pages.

To evaluate the performance of translation extraction, we used the average top-n inclusion rate as a metric. For a set of test queries, the top-n inclusion rate was defined as the percentage of queries whose translations could be found in the first n extracted translations. Also, we wished to know if the coverage rate of translations, i.e. the percentage of queries whose translations could be found in the whole extracted candidate set, was high enough in the top search-result pages for real queries.

Table 3. Coverage and inclusion rates for popular Chinese queries using different methods.

Method	Query Type	Top-1	Top-3	Top-5	Coverage
CV	Dic	56.4%	70.5%	74.4%	80.1%
	OOV	56.2%	66.1%	69.3%	85.0%
	All	56.3%	67.7%	71.2%	83.3%
χ^2	Dic	40.4%	61.5%	67.9%	80.1%
	OOV	54.7%	65.0%	68.2%	85.0%
	All	49.5%	63.7%	68.1%	83.3%
Combined	Dic	57.7%	71.2%	75.0%	80.1%
	OOV	56.6%	67.9%	70.9%	85.0%
	All	57.2%	68.6%	72.8%	83.3%

Table 4. Coverage and inclusion rates for popular English queries using different methods.

Method	Top-1	Top-3	Top-5	Coverage
CV	50.9%	60.1%	60.8%	80.9%
χ^2	44.6%	56.1%	59.2%	80.9%
Combined	51.8 %	60.7%	62.2%	80.9%

Table 5. Coverage and inclusion rates for random queries using the different methods.

Method	Top-1	Top-3	Top-5	Coverage
CV	25.5%	45.5%	50.5%	60.5%
χ2	26.0%	44.5%	50.5%	60.5%
Combined	29.5%	49.5%	56.5%	60.5%

Table 6. Inclusion rates for proper names and technical terms using the combined method.

Query Type	Top-1	Top-3	Top-5
Scientist Name	40.0%	52.0%	60.0%
Disease Name	44.0%	60.0%	70.0%

4.4.2 Performance

Web Queries

We carried out experiments to determine the performance of the proposed approach by extracting translations for the popular-query set. Tables 3 and 4 show the results in terms of top 1-5 inclusion rates and coverage rates for Chinese and English queries respectively. In this table, "CV", "χ^2" and "Combined" represent the context-vector analysis, the chi-square test, and the combined method, respectively. In addition, "Dic", "OOV" and "All" represent the terms covered in a dictionary, the terms not in a dictionary, and the total test query set, respectively. The coverage rates we obtained were promising, which shows that the Web contains rich mixed texts in both languages. The performance of the English query set was not as good as the Chinese query set. The reason for this was that the English queries suffered from more noise in Chinese translation candidates since the search-result pages in the Chinese Web generally contain much more Chinese than English content. We also conducted an experiment for random queries. As Table 5 shows, the coverage rates were encouraging.

Proper Names, Technical Terms and Common Terms

To further determine the effectiveness of the proposed approach in dealing with the translation of proper names and technical terms, we conducted an experiment on the test sets of scientists' names and medical terms using the combined method. As the results in Table 6 show, the top-1 inclusion rates for the scientists' and disease names were 40% and 44% respectively. Some examples of the extracted correct translations are shown in Table 7.

Although the achieved performance for real queries looked promising, we wished to know if it was equally effective for common terms. We randomly selected 100 common nouns and 100 common verbs from a general-purpose Chinese dictionary. Table 8 shows the results obtained using the combined method. It is easy to see that the proposed approach is less reliable in

Table 7. Some examples of the test English proper names and technical terms, and their extracted Chinese translations.

Query Type	English Query	Extracted Translations (in Traditional Chinese)
Scientist Name	Galilei, Galileo (Astronomer) Crick, Francis (Biologists) Kepler, Johannes (Mathematician) Dalton, John (Physicist) Feynman, Richard (Physicist)	伽利略/伽里略/加利略 克立克/克里克 克卜勒/開普勒/刻卜勒 道爾頓/道耳吞/道耳頓 費曼
Disease Name	Hypoplastic Left Heart Syndrome Legionnaires' Disease Shingles Stockholm Syndrome Sudden Infant Death Syndrome (SIDS)	左心發育不全症候群 退伍軍人症 帶狀皰疹/帶狀庖疹 斯德哥爾摩症候群 嬰兒猝死症

extracting translations of such common terms. One possible reason is that the usages of common terms are diverse on the Web and the retrieved search results are not highly relevant. It is fortunate that many of these common words can be found in general-purpose translation dictionaries.

Table 8. Top 1, 3, 5 inclusion rates obtained using the combined method for extracting translations of common nouns and verbs.

Query Type	Top-1	Top-3	Top-5
100 Common Nouns	23.0%	33.0%	43.0%
100 Common Verbs	6.0%	8.0%	10.0%

5. BILINGUAL LEXICON CONSTRUCTION

5.1. The Approach

To enhance CLIR services in a digital library that only has monolingual document collections, the proposed approach can be used to construct a domain-specific bilingual lexicon. We take the document set in digital libraries into consideration. The document set in the target language is first analyzed and possible key terms that are representative of the document set are extracted, using the proposed term extraction method. These extracted key terms are likely to be similar to terms that users may use in real user queries, since they are relatively more significant than other terms in the documents. The proposed term translation method can then be applied to those key terms not included in common translation dictionaries to obtain the translation of key terms in the source language. Therefore, a bilingual lexicon can then be constructed where the mappings between key terms and relevant terms in the source and target languages are maintained.

As we have already indicated, the constructed bilingual lexicon can benefit CLIR services. For a given source query, the similarity with candidate source relevant terms can be calculated using the context vector method presented in Section 4. Also, and the top-ranked relevant terms can be extracted using the constructed bilingual lexicon. After the corresponding translations of relevant terms are obtained, relevant documents in the target language can be retrieved, using these relevant translations. The source query can then be expanded with the relevant translations and conventional CLIR methods can be used to retrieve documents in the target language.

5.2. An Application

We tested the STICNET Database[4], which is a government-supported Web-accessible digital library system providing a search service for scientific documents collected in Taiwan. The system contained documents in either English or Chinese, but no cross-language search was provided. To test the performance of bilingual lexicon construction, we selected 1,367 Information Engineering documents and 5,357 Medical documents respectively from the STICNET Database for the period 1983 to 1997 as the test bed. Using the PAT-tree-based term extraction method, key terms were automatically extracted from each document collection and their relevant translations were extracted by the proposed term translation approach.

In the collection of Information Engineering documents, 1,330 key terms (with a threshold of 2 to 6-gram character strings, a term frequency>10, and an association value>0.1) were automatically extracted. Meanwhile, 5,708 key terms (with a threshold of 2 to 6-gram character strings and a term frequency>40) were automatically extracted from the Medical document collection. Among the 1,330 auto-extracted key terms from the Information Engineering documents, 32% were not included in KUH Chinese Dictionary[5] (unknown terms) - one of the largest Chinese dictionaries with 158,239 term entries - where 75% of these unknown terms were found useful. In the case of Medical documents, 71% of the 5,708 auto-extracted key terms were not included in KUH Chinese Dictionary where 36.6% of these unknown terms were found useful. Table 9 shows the accuracy of the extracted translations for these useful unknown terms. The promising result shows the potential of the proposed approach to assist bilingual lexicon construction.

[4] http://sticnet.stic.gov.tw/

[5] http://www.edu.tw/mandr/clc/dict/

Table 9. The top-*n* inclusion rates of translations for auto-extracted useful unknown terms.

Query Type	Top-1	Top-3	Top-5
Auto-extracted useful terms in Information Engineering	33.3%	37.5%	50.0%
Auto-extracted useful terms in Medicine	34.6%	46.2%	50.0%

6. RELATED WORK

Many effective retrieval models have been developed for CLIR. For example, the Latent Semantic Indexing (LSI) method [4] has been utilized to model inter-term relationships, instead of exact term matching. Other methods include the cross-lingual relevance model [11], which integrates popular techniques of disambiguation and query expansion. However, translation of queries not covered in a bilingual dictionary remains one of the major challenges in practical CLIR services [9].

To deal with the translation of out-of-dictionary terms, conventional research on machine translation has generally used statistical techniques to automatically extract translations from domain-specific, sentence-aligned parallel bilingual corpora [20]. However, a large parallel corpus is difficult to obtain. Some work has been done on term translation extraction from comparable texts, such as bilingual newspapers [5], which are easier to obtain. Using a non-parallel corpus is more difficult than a parallel one, due to the lack of alignment correspondence for sentence pairs. On the other hand, research on digital libraries has made the same endeavor. Larson et al. [10] proposed a method for translingual vocabulary mapping using multilingual subject headings of book titles in online library catalogs - a kind of parallel corpus. However, book titles are still limited in coverage, compared to the rich resources on the Web.

A new potential research direction is to perform query translation directly, through mining the Web's multilingual and wide-range resources [16]. Web mining is a new research area that focuses on finding useful information from large amounts of semi-structured hypertexts and unstructured texts [1]. Chen et al. [2] proposed a dictionary-based approach in which the search results returned from Yahoo China search engine were utilized to extract translations for terms not covered in the dictionary. In their work only an English term appearing (maybe in parenthesis) immediately or closely after a Chinese term was considered a possible translation. In our previous research, we proposed an approach for extracting translations of Web queries through the mining of anchor texts and link structures and obtained very promising results [12, 13]. Previous experiments showed that the anchor-text-based approach can achieve a good precision rate for popular queries. Its major drawback is the very high cost of the hardware and software required to collect sufficient anchor texts from Web pages. Collecting anchor texts requires a powerful Web spider and takes cost of network bandwidth and storage. Because of the practical needs of digital libraries, search-result pages, which are easier to obtain are, therefore, investigated in this paper.

7. CONCLUSION

In this paper, we have introduced a Web-based approach for dealing with the translation of unknown query terms for cross-language information retrieval in digital libraries. With the proposed term extraction and translation methods, it is feasible to translate unknown terms and construct a bilingual lexicon for key terms extracted from documents in a digital library. With the help of such bilingual lexicons, it would be convenient for users to formulate cross-lingual queries. The simplicity of the approach not only makes it very suitable for digital library systems, but would also facilitate the implementation of CLIR services.

8. REFERENCES

[1] Chakrabarti, S. *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kaufmann, 2002.

[2] Chen, A., Jiang, H., and Gey, F. Combining Multiple Sources for Short Query Translation in Chinese-English Cross-Language Information Retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL 2000)*, 2000, 17-23.

[3] Chien, L.F. PAT-Tree-based Keyword Extraction for Chinese Information Retrieval. In *Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1997)*, 1997, 50-58.

[4] Dumais, S. T., Landauer, T. K., and Littman, M. L. Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing. In *Proceedings of ACM-SIGIR Workshop on Cross-Linguistic Information Retrieval (SIGIR 1996)*, 1996, 16-24.

[5] Fung, P. and Yee, L. Y. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics (ACL 1998)*, 1998, 414-420.

[6] Gale, W. A. and Church, K. W. Identifying Word Correspondences in Parallel Texts. In *Proceedings of DARPA Speech and Natural Language Workshop*, 1991, 152-157.

[7] Gale, W.A. and Church, K.W. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics, 19*, 1 (1993), 75-102.

[8] Gonnet, G.H., Baeza-yates, R.A. and Snider, T. New Indices for Text: Pat Trees and Pat Arrays. *Information Retrieval Data Structures & Algorithms*, Prentice Hall, 1992, 66-82.

[9] Kwok, K. L. NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS. In *Proceedings of NTCIR workshop meeting*, 2001, 111-118.

[10] Larson, R. R., Gey, F., and Chen, A. Harvesting Translingual Vocabulary Mappings for Multilingual Digital Libraries. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, 2002, 185-190.

[11] Lavrenko, V., Choquette, M., and Croft, W. B. Cross-Lingual Relevance Models. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR 2002)*, 2002, 175-182.

[12] Lu, W. H., Chien, L. F., and Lee, H. J. Translation of Web Queries using Anchor Text Mining. *ACM Transactions on Asian Language Information Processing*, 1 (2002), 159-172.

[13] Lu, W. H., Chien, L. F., and Lee, H. J. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems*, 22 (2004), 1–28.

[14] Manber, U. and Baeza-yates, R. An Algorithm for String Matching with a Sequence of Don't Cares. *Information Processing Letters*, 37 (1991), 133-136.

[15] Morrison, D. PATRICIA: Practical Algorithm to Retrieve Information Coded in Alphanumeric. *JACM*, 1968, 514-534.

[16] Nie, J. Y., Isabelle, P., Simard, M., and Durand, R. Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR 1999)*, 1999, 74-81.

[17] Rapp, R. Automatic Identification of Word Translations from Unrelated English and German Corpora, In *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics (ACL 1999)*, 1999, 519-526.

[18] Silva, J. F., Dias, G., Guillore, S., and Lopes, G. P. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. *Lecture Notes in Artificial Intelligence, 1695*, Springer-Verlag, 1999, 113-132.

[19] Silva, J. F. and Lopes, G. P. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In *Proceedings of the 6th Meeting on the Mathematics of Language*, 1999, 369-381.

[20] Smadja, F., McKeown, K., and Hatzivassiloglou, V. Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics, 22, 1* (1996), 1-38.