Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval

Pu-Jen Cheng[*], Jei-Wen Teng[*], Ruei-Cheng Chen[*], Jenq-Haur Wang[*], Wen-Hsiang Lu[+], and Lee-Feng Chien[*,†]

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

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

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

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

ABSTRACT

It is crucial for cross-language information retrieval (CLIR) systems to deal with the translation of *unknown queries*[1] due to that real queries might be short. The purpose of this paper is to investigate the feasibility of exploiting the Web as the corpus source to translate unknown queries for CLIR. We propose an online translation approach to determine effective translations for unknown query terms via mining of bilingual search-result pages obtained from Web search engines. This approach can alleviate the problem of the lack of large bilingual corpora, translate many unknown query terms, provide flexible query specifications, and extract semantically-close translations to benefit CLIR tasks – especially for cross-language Web search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms

Algorithms, Experimentation, Performance.

Keywords

Query Translation, Cross-Language Information Retrieval, Cross-Language Web Search, Web Mining.

1. INTRODUCTION

Conventionally CLIR approaches [4,7,8,12,21] have focused mainly on incorporating dictionaries and domain-specific bilingual corpora for query translation [6,10,18]. The general assumption of such approaches is that the incorrect translation of a few query terms in a query is tolerable and can be remedied via query expansion in the process of document retrieval. For longer queries, such as TREC queries (the average length of a TREC topic description was 15 words [19]), it is still possible to retrieve relevant documents in target languages even if there exist a few unknown query terms. However, real queries are often short. For example, previous Web search engine log analyses revealed that the average query length for a Web search was about 2.3 words in English [17] and 3.18 characters in Chinese [13].

Conventional CLIR approaches [8,21] that are based on domain-specific corpora might not be applicable to dealing with the translation of short queries with unknown terms. First, sufficiently large bilingual corpora for certain subject domains and language pairs are not always available. Second, using small corpora may provide a low coverage rate for translation. In our analysis of a 3-month log with 228,566 unique queries from a real-world Chinese search engine log[2], nearly 82.9% of the top 19,124 high frequent query terms (with 80% coverage rate) were not included in the LDC[3] (*Linguistic Data Consortium*) English-to-Chinese lexicon. Furthermore, 14.9% of the unknown query terms were in English (with 1.19 words on average). These English terms were potential queries in the Chinese log that needed correct cross-language translations. How to efficiently translate unknown terms in short queries has, therefore, become a major challenge for real CLIR systems [4,7].

As more data is being put on the Web every day, there is a great potential to exploit the Web as the corpus to automatically find effective translations for unknown queries. Mining of bilingual corpora from the Web has attracted a lot of attention [5,11,15,22], but some of the proposed methods might not be general to common applications in which queries are short and diverse. For this reason, this paper presents a search-result-based approach to fully exploiting Web resources for query translation.

For some language pairs, such as Chinese and English, as well as Japanese and English, the Web consists of rich texts in a mixture of multiple languages. Many of them contain bilingual translations of proper nouns, such as company names and personal names. We are interested in realizing: whether this nice characteristic makes it possible for the bilingual translations of a large number of unknown query terms to be automatically extracted; and whether the extracted bilingual translations (if any) can effectively improve CLIR performance. Real search engines, such as *Google*[4] and *AltaVista*[5], allow us to search English terms only for pages in a certain language, e.g., Chinese or Japanese, which are normally returned in a long ordered list of *snippets* of summaries (including titles and

[1] Unknown queries refer to those queries containing terms not covered by general translation dictionaries.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

[2] Dreamer (http://www.dreamer.com.tw), which was a popular Chinese search engine.

[3] An English-to-Chinese lexicon (http://www.ldc.upenn.edu/Projects/Chinese) with about 120K entries.

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

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

page descriptions) to help users locate interesting documents. This motivates us to investigate how to mine query translations from these dynamically-retrieved bilingual search-result pages.

Two major difficulties in mining query translations from search-result pages will be encountered: one is *term extraction* - how to extract terms with correct lexical boundaries from the noisy bilingual search-result pages as translation candidates; and the other is *translation selection* - how to estimate term similarity for determining correct or relevant translations from the extracted candidates. The purpose of this paper is to deal with the two problems. Moreover, in our previous research [9] anchor texts of Web pages have been utilized as an aligned bilingual corpus for query translation. We are also interested in whether the proposed search-result-based approach can be combined with it to improve CLIR performance.

Taking the advantage of search engines' quick responses, the proposed approach requires little cost to collect an effective and dynamic corpus for a query, and the translation process of it can be finished online and in acceptable access/computing time. To determine the performance of the proposed approach when applied to CLIR, we have conducted extensive experiments including the experiments with the NTCIR-2 English-Chinese IR task. The proposed approach was found to be effective in extracting correct translations of unknown query terms contained in the NTCIR-2 title queries and real-world Web queries. Moreover, the proposed approach provides flexible query specifications and extracts semantically-close translations to benefit CLIR tasks – especially for cross-language Web search.

In the rest of this paper, we first make a brief review on some translation techniques based on Web corpora and the previously-developed anchor-text-based approach in Section 2, and describe the proposed search-result-based approach in Section 3. The experiments and discussions are presented in Sections 4 and 5, respectively. Finally, in Section 6, we present our conclusions.

2. REVIEW ON WEB-BASED APPROACHES

The parallel-corpus-based approaches: Collecting parallel texts of different language versions from the Web has recently received much attention [5]. Nie et al. [11] tried to automatically discover parallel Web documents written in English and French. They assumed a Web page's parents might contain the links to different versions of it and Web pages with the same content might have similar structures and lengths. Resnik [15] addressed the issue of language identification for finding Web pages in the languages of interest. Yang et al. [22] presented an alignment method to identify one-to-one Chinese and English title pairs based on dynamic programming. Mining of parallel texts is feasible, but some of the proposed methods might not be general to common applications in which queries are short and diverse. Moreover, these methods often require powerful crawlers to gather sufficient Web data as well as more network bandwidth and storage.

The comparable-corpus-based approaches: Less attention has been devoted to mining of comparable texts from the Web. Fung et al. [2] used a vector-space model and took a bilingual lexicon (called *seed words*) as feature sets to estimate the similarity between a word and its translation candidates. Comparable corpora are far easier to obtain; however, how to automatically gather appropriate comparable corpora from the Web is still a challenging task.

The anchor-text-based approach: In addition to mining parallel texts on the Web, anchor texts have been utilized as an aligned bilingual comparable corpus for query translation in our previous research [9]. An anchor text is the descriptive part of an out-link of a Web page used to provide a brief description of the linked Web page. There are a variety of anchor texts in multiple languages that might link to the same pages from all over the world. For an unknown term appearing in an anchor text of a Web page, it is likely that its corresponding target translations may appear together in other anchor texts linking to the same page. Such a bundle of anchor texts pointing together to the same page is called as an *anchor-text set*. To determine the most probable target translation for a query term, a probabilistic model was presented. This model was used to estimate the probability value between the query term and all the translation candidates that co-occur in the same anchor text sets. A translation candidate had a higher chance of being an effective translation only if it was written in the target language and frequently co-occurred with the query term in the same anchor text sets. The model further assumed that the translation candidates in the anchor texts of the pages with higher authority may be more reliable. Hence, the similarity between a source query s and a translation candidate t was defined as:

$$S_{AT}(s,t) = \frac{P(s \cap t)}{P(s \cup t)} = \frac{\sum_{i=1}^{n} P(s \cap t \mid u_i) P(u_i)}{\sum_{i=1}^{n} P(s \cup t \mid u_i) P(u_i)}$$

$$= \frac{\sum_{i=1}^{n} P(s \mid u_i) P(t \mid u_i) P(u_i)}{\sum_{i=1}^{n} [P(s \mid u_i) + P(t \mid u_i) - P(s \mid u_i) P(t \mid u_i)] P(u_i)}.$$

The measure was estimated based on the co-occurrence of s and t in the anchor texts of the concerned Web pages $U = \{u_1, u_2, \ldots u_n\}$, in which u_i is a page of concern and $P(u_i)$ is the probability value used to measure the authority of page u_i. By considering the link structures and concept space of Web pages, $P(u_i)$ was estimated along with the probability of u_i being linked, and its estimation was defined by $P(u_i) = L(u_i)/\Sigma_{j=1,n} L(u_j)$, where $L(u_j)$ indicates the number of in-links of page u_j. In addition, we assumed that s and t are independent given u_i; then, the joint probability $P(s \cap t|u_i)$ was equal to the product of $P(s|u_i)$ and $P(t|u_i)$. The values of $P(s|u_i)$ and $P(t|u_i)$ were estimated by calculating the fractions of the numbers of u_i's in-links containing s and t over $L(u_i)$, respectively.

Although the anchor-text-based approach has been proven effective in extracting the translations of proper nouns in multiple languages, like most of the Web-based approaches, it has a drawback: it requires crawling the Web to gather sufficient training data as well as more network bandwidth and storage. For this reason, this paper presents the search-result-based approach to fully exploiting Web resources but reduce such costs.

3. THE PROPOSED SEARCH-RESULT-BASED APPROACH

3.1 Observation

The Web contains rich texts in a mixture of multiple languages. For example, Chinese pages on the Web may be written in Chinese as a main language and in English as an auxiliary language. According to our observations, translated or semantically-close terms frequently occur together with a source query term in mixed-

language texts. For example, Figure 1 illustrates the search-result page of the English query "*Yahoo*," which was submitted to Google to search for traditional Chinese pages. Many relevant terms could be obtained including both the query itself and its Chinese aliases, such as correct translations *"雅虎"* (*Yahoo*) and *"奇摩"* (*Kimo*, a Yahoo's alias in traditional Chinese) and relevant terms *"雅虎中國"*(*Yahoo China*), *"雅虎香港"* (*Yahoo HongKong*), and *"雅虎台灣"* (*Yahoo Taiwan*). Note that despite the fact that the translated terms are not exact translation equivalents, if they are strongly related to the original query term, the translations may be still helpful for CLIR. This characteristic of bilingual search-result pages might not be applied to all language pairs, but is useful for some Asian languages mixed with English such as Japanese and Korean.

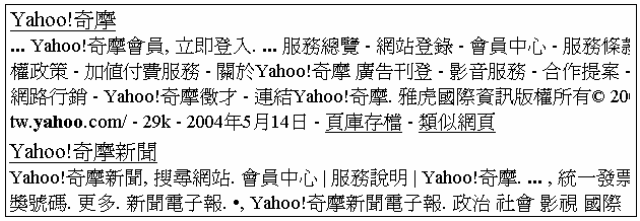
Yahoo!奇摩
... Yahoo!奇摩會員, 立即登入. ... 服務總覽 - 網站登錄 - 會員中心 - 服務條款 - 隱私權政策 - 加值付費服務 - 關於Yahoo!奇摩 廣告刊登 - 影音服務 - 合作提案 - 網路行銷 - Yahoo!奇摩徵才 - 連結Yahoo!奇摩. 雅虎國際資訊版權所有© 20
tw.yahoo.com/ - 29k - 2004年5月14日 - 頁庫存檔 - 類似網頁
Yahoo!奇摩新聞
Yahoo!奇摩新聞, 搜尋網站. 會員中心 | 服務說明 | Yahoo!奇摩. ... , 統一發票獎號碼. 更多. 新聞電子報. •, Yahoo!奇摩新聞電子報. 政治 社會 影視 國際

Figure 1: An example of the search-result page in traditional Chinese of the English query "*Yahoo*" obtained from Google.

To examine the feasibility of mining search-result pages for query translation in a large scale, we conducted some experiments. First, we selected 430 popular English query terms (*PE-430*) from a real search engine log (referred to as Section 4.2) and translated them into a Chinese query set (*PC-430*). Also, we randomly selected 100 English query terms (*RE-100*) from the top 19,124 query terms in the log and translated them into a Chinese query set (*RC-100*). The coverage rates of the test queries' correct translations in different numbers of the retrieved snippets were then observed. The experiment result, as shown in Figure 2, reveals that more than 95% of the popular queries' translations appeared in top 30~40 snippets of summaries from Google, and about 70% of the random queries' translations were covered as well. Moreover, many relevant translations were also found in the top snippets. The above nice characteristic makes it possible for the English-Chinese bilingual translations of a large number of unknown terms to be automatically extracted.

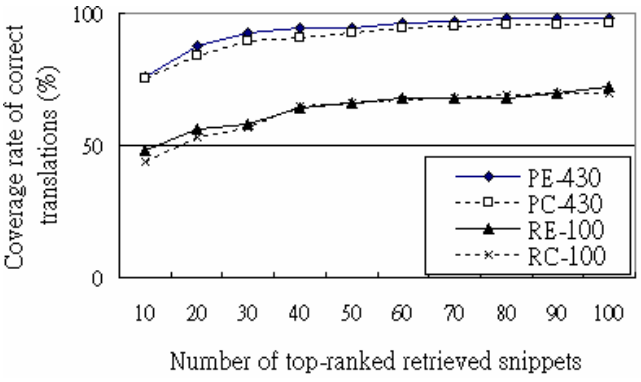

Figure 2: Coverage rates of correct translations in top-ranked snippets retrieved from Google for the four test query sets.

3.2 Considered Problem and Challenge

For each query term expressed in one (source) language, the problem is whether it is possible to extract translations in another (target) language from bilingual search-result pages containing the query term. To deal with the problem, we need to extract terms from the search-result pages as translation candidates first. All extracted translation candidates are ranked according to their similarity to the source query terms. The top-ranked K candidates will be selected as possible translations of the given query term. However, the process is not straightforward. It is challenging to (1) extract terms with correct lexical boundaries and minimum noisy terms from the search-result pages, and (2) find correct or semantically-close translations for each unknown query term within an acceptable amount of search-result pages and network access time. In the following sections, the proposed methods of term extraction and translation selection will be described in detail.

3.3 Term Extraction

The issues of term extraction include whether all possible terms in the target language can be extracted from the search-result pages and their lexical boundaries can be correctly segmented. To extract a term, we present a new association measure, called *SCPCD*, of every character or word n-gram in the target language. *SCPCD* combines the symmetric conditional probability (*SCP*) [16] with the concept of context dependency (*CD*) [1], and is defined as:

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

where *SCP* is the association estimation of its composed sub n-grams and defined as:

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

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. *CD* is a refined measure varying from 0 to 1, and is defined as:

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

where $LC(w_1 \dots w_n)$ (or $RC(w_1 \dots w_n)$) is the number of unique left (or right) adjacent words/characters for the n-gram in the corpus, or equal to the frequency of the n-gram if there is no left (or right) adjacent word/character.

In *SCPCD*, *SCP* is to measure the cohesion sticking the words together within a word n-gram to a certain degree; *CD* is to judge whether the appearance of an n-gram is dependent on a certain string containing it.

To extract key terms from the obtained n-grams, *SCPCD* can be further combined with a local maxima algorithm [16] as well as the assistance of the data structure of PAT-tree [1]. Further information about *SCPCD* and its performance could be referred to our earlier work which emphasized on translating unknown queries for digital library applications [20].

3.4 Translation Extraction

The challenge of translation extraction lies in how to estimate the similarity between a query term and each extracted translation

candidate solely based on the search-result pages. We present two methods for estimating term similarity. The first method is simpler and depends on the co-occurrences of a query term and its translation candidates on the Web. The second method extracts a so-called *context vector* as a feature from the search-result pages for each term. The similarity between the query term and a candidate is then determined by the distance between their features. The two methods differ in the complexity of computation and robustness under various conditions of distribution.

The Chi-square Method: A number of statistical measures have been proposed for estimating term association based on co-occurrence analysis, including mutual information (MI), DICE coefficient, chi-square test, and log-likelihood ratio [14]. Chi-square test (χ^2) is adopted in our study because the required parameters for it can be obtained by submitting Boolean queries to search engines and utilizing the returned page counts (number of pages). Given a source query s and a translation candidate t, suppose

the total number of Web pages is N,
the number of Web pages containing both s and t, $n(s,t)$, is a,
the number of Web pages containing s but not t, $n(s,\neg t)$, is b,
the number of Web pages containing t but not s, $n(\neg s,t)$, is c, and
the number of Web pages containing neither s nor t, $n(\neg s, \neg t)$, is d.

(Although d is not provided by search engines, it can be computed by $d=N-a-b-c$.)

Assume s and t are independent. Then

the expected frequency of (s,t), $E(s,t)$, is $(a+c)(a+b)/N$,
the expected frequency of $(s,\neg t)$, $E(s,\neg t)$, is $(b+d)(a+b)/N$,
the expected frequency of $(\neg s,t)$, $E(\neg s,t)$, is $(a+c)(c+d)/N$, and
the expected frequency of $(\neg s,\neg t)$, $E(\neg s,\neg t)$, is $(b+d)(c+d)/N$.

Hence, the conventional chi-square test can be computed as:

$$S_{\chi 2}(s,t) = \sum_{\forall X\in\{s,\neg s\},\forall Y\in\{t,\neg t\}} \frac{[n(X,Y)-E(X,Y)]^2}{E(X,Y)} = \frac{N\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)}.$$

The Context-Vector Method: The basic idea of the second method is that a query term's translation equivalents may share common contextual terms with the query term in their search-result pages. Similar work includes [14]. For both of the query term and its candidates, we take their contextual terms constituting the search-result pages as their features. The similarity between the query term and its candidates will be computed based on their feature vectors in the vector-space model.

Herein, we adopt the conventional *tf-idf* weighting scheme to estimate the significance of features, which is defined as:

$$w_{t_i} = \frac{f(t_i,p)}{\max_j f(t_j,p)} \times \log(\frac{N}{n}),$$

where $f(t_i,p)$ is the frequency of term t_i in search-result page p, N is the total number of Web pages, and n is the number of the pages containing t_i. Finally, the similarity between a query term s and its translation candidate t can be estimated with the cosine measure, i.e., $S_{CV}(s,t)=cos(cv_s, cv_t)$, where cv_s and cv_t are the context vectors of s and t, respectively.

Analysis: Although the chi-square method is simple to compute, it may have some limitations in query translation. First, the method is more applicable to high-frequency query terms than low-frequency query terms since high-frequency query terms are more likely to appear with their candidate terms. In the context-vector method, low-frequency query terms still have a chance of extracting correct or relevant translations if the query term and its translations share common contexts in the search-result pages. Second, certain candidates that frequently co-occur with a query term may not imply that they are appropriate translations. Although the context-vector method provides an effective way to overcome this problem, its performance strongly depends on the quality of the retrieved search-result pages such as the sizes and amounts of snippets. Feature selection needs to be carefully handled in some cases.

Benefiting from mining of search-result pages, both of the methods do not need to collect large corpora in advance, compared with the anchor-text-based approach (as mentioned in Section 2). Their execution time is determined by the processes of Web search and term/feature extraction. Suppose n_t translation candidates are extracted for each query term. The chi-square method requires $1+3$ n_t Web searches and the context-vector method requires $1+n_t$ ones. The chi-square method is approximately 3 times the number of Web searches needed by the context-vector method when n_t is large. However, the context-vector method needs to do extra $1+n_t$ feature extraction tasks. In general, feature extraction takes much more time than Web search needs.

3.5 The Combined Approaches

The context-vector and chi-square methods are basically complementary. The chi-square method has difficulties in dealing with infrequent terms and the context-vector method needs to carefully handle the issue of feature selection. Intuitively, a more complete solution is to integrate the two different methods. The proposed search-result-based approach is actually a combination of them (χ^2+CV). However, to effectively exploit the two kinds of Web resources: anchor texts and search-result pages, we further develop a combined approach. In this approach, we may combine the probabilistic inference model with the context-vector and chi-square methods. Considering the various ranges of similarity values between the methods, we use a linear combination weighting scheme to compute the similarity between a query term s and its translation candidate t as follows:

$$S_{\{m_i\}}(s,t) = \sum_{m_i} \frac{\alpha_{m_i}}{R_{m_i}(s,t)},$$

where $m_i \in \{\chi^2, CV, AT\}$, α_{mi} is an assigned weight for each similarity measure S_{mi}, and $R_{mi}(s,t)$, which represents the similarity ranking of each translation candidate t with respect to s, is assigned to be from 1 to k (number of candidates) in decreasing order of similarity measure $S_{mi}(s,t)$. The combined approach may have different combinations of these similarity estimation methods. The performance achieved by the different combinations will be described in the next section.

4. PERFORMANCE EVALUATION

In this section, we conducted extensive experiments on English-Chinese CLIR and cross-language Web search to examine the performance of the proposed approach. For comparison with conventional parallel-corpus-based approaches, we used the Hong Kong Law parallel text collection, called *HK parallel corpus*, which

contained 238,236 English-Chinese text paragraphs and was available from LDC. The corpus was selected because it has been employed in related works [7]. We adopted χ^2, a χ^2 –like statistic, to measure the association between terms, and extracted word/phrase translation pairs [3]. To compare with the anchor-text-based approach, we had collected 1,980,816 traditional Chinese Web pages in Taiwan, and then extracted 109,416 pages (URLs), whose anchor-text sets contained both traditional Chinese and English terms, as the anchor-text-set corpus for testing the anchor-text-based approach. We obtained the search-result pages of our test queries by submitting them to the real-world search engines, including *Google* and *Openfind*[6]. Basically, we used only the first 100 retrieved snippets to extract terms and features. The average top-*n* inclusion rate was adopted as a metric. For a set of test query terms, its top-*n* inclusion rate was defined as the percentage of the queries whose translations could be found in the first *n* extracted translations.

4.1 Experiments on NTCIR-2

In the NTCIR-2 English-Chinese task, we carried out experiments to retrieve Chinese documents using English queries.

4.1.1 Query Translation

Several experiments have been made for the translation of NTCIR-2 "Title Queries," whose length was close to that of real Web queries. There were a total of 178 unique query terms in the 50 test English title queries, and 22 of them were not included in the LDC English-Chinese lexicon. The average length of the title queries was 3.8 English words (after removing stop words), which was close to the length of real Web queries as mentioned in Section 1. Table 1 lists several examples of the title queries. Table 2 shows the results in terms of the top 1-5 inclusion rates for the English queries. In this table, "*Dic*", "*OOV*" and "*All*" (i.e. *Dic* and *OOV*) represent the terms existing in the dictionary, the terms not in the dictionary, and the total test query set, respectively. "χ^2", "*CV*", "*AT*", "χ^2+*CV*", *and* "χ^2+*CV*+*AT*" represent the approach based on the chi-square, context-vector, anchor-text, chi-square plus context-vector, and chi-square plus context-vector and anchor-text methods, respectively.

The experimental results show that the anchor-text-based and search-result-based approaches are quite complementary. Although either the anchor-text-based approach or the proposed search-result approach (χ^2+*CV*) alone was effective, the approach based on both of them (χ^2+*CV*+*AT*) achieved the best performance in maximizing the inclusion rates in every case. The anchor-text-based approach can achieve higher precision (higher top-1 inclusion rates) for the test queries, and the proposed search-result-based approach can have high coverage of various translation pairs (higher inclusion rates in the top 5 lists). Besides, the approach based on the HK parallel corpus was not suitable due to the limitations of the size and domain of the corpus, and the approach based on either χ^2 or *CV* was not reliable enough. However, the anchor-text-based approach requires powerful spiders and high network bandwidth costs to collect a sufficient corpus. The proposed search-result approach (χ^2+CV) is thus very promising, because it requires little cost to collect an effective and dynamic corpus, and the translation process of it can be finished online and in acceptable access/computing time.

Table 1: Some examples of the title queries in NTCIR-2.

	English Title Queries	Chinese Title Queries
Q06	Kosovar refugees	科索沃難民潮
Q12	Michael Jordan's retirement	麥可喬登退休
Q23	Disneyland	迪士尼樂園
Q28	Cutting down the timber of Chinese cypress in Chilan	棲蘭檜木砍伐
Q43	CIH computer virus	ＣＩＨ電腦病毒
Q46	Ma Yo-yo cello recital	馬友友演奏會

The above Web-based approaches were found to be effective in finding translations of proper names, e.g., some local place names "*Chilan*" (棲蘭), "*Meinung*" (美濃) and foreign names "*Jordan*" (喬登, 喬丹), "*Kosovar*" (科索沃), "*Carter*" (卡特), etc. The accuracy for the unknown query terms might be even higher than that for the others. However, the results also revealed that the proposed approach might not be reliable enough when used to extract translations of some common terms, e.g.,"*victim*" (受難者) and "*abolishment*" (廢止). One of the possible reasons is that the usage of common terms is diverse on the Web and the retrieved search results are not highly relevant. Fortunately, many of these common terms can be found in general translation dictionaries.

4.1.2 CLIR Performance

Another important merit of the proposed approach is its effectiveness in extracting semantically-close translations. As the examples shown in Table 3, an extracted top translation had a higher chance of being a semantically-close translation, even though it was not a translation equivalent of the test query term. We investigated whether these automatically extracted translations could benefit CLIR. The probabilistic retrieval model [21] was adopted in the experiment and defined as:

$$P(Q \mid D) = \prod_{e \in Q} P(e \mid D) = \prod_{e \in Q} [\lambda P(e) + (1-\lambda) \sum_{c} P(e \mid c) P(c \mid D)],$$

where Q is a query, D is a document, e is an English query term in Q, c is a target translation of e in traditional Chinese and λ represents a smoothing parameter. In addition, $P(e)$ is the priori probability of e, which can be estimated based on e's page frequency on the Web. $P(c|D)$ is the probability of c appearing in document D. $P(e|c)$ is the translation probability of e given c, which could be estimated by different translation approaches. Four approaches have been tested, which are the dictionary-based approach (using the LDC English-Chinese lexicon) in which $P(e|c) \approx 1/n_e$, where n_e is the number of possible translations of c and $P(e|c)=0$ if n_e is zero; the proposed search-result-based approach in which $P(e|c) \approx S_{\{\chi 2,CV\}}(e,c)$; the approach combining with search-result corpus and anchor-text corpus in which $P(e|c) \approx S_{\{\chi 2,CV,AT\}}(e,c)$; and the hybrid approach combining all resources (dictionary + anchor-text corpus + search-result corpus) in which $P(e|c) \approx [S_{\{\chi 2,CV,AT\}}(e,c)+1/n_e]/2$.

We used mean average precision (*MAP*) values to evaluate the retrieval performance. The number of translations used for CLIR

[6] http://www.openfind.com.tw

Table 2: Inclusion rates for the NTCIR-2 title query terms using the different approaches.

Query Types	Approaches		Dic			OOV			All		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
NTCIR-2 Title Queries	HK Parallel Corpus		27.4%	35.4%	37.7%	4.54%	4.54%	4.54%	24.8%	31.9%	34.0%
	Web Corpora	χ^2	32.7%	61.4%	71.2%	22.7%	59.1%	72.7%	31.4%	59.4%	71.4%
		CV	47.7%	62.1%	70.6%	40.9%	59.1%	68.2%	46.9%	64.6%	70.3%
		χ^2 + CV	51.0%	68.0%	77.1%	45.4%	68.2%	72.7%	50.3%	68.0%	76.6%
		AT	56.4%	66.7%	68.6%	63.6%	72.7%	77.3%	57.3%	67.4%	69.7%
		χ^2 + CV+ AT	60.3%	77.6%	83.3%	68.1%	81.8%	86.3%	61.2%	78.1%	83.7%

Table 3: Some examples of the test English NTCIR-2 title query terms and their extracted Chinese translations using the proposed search-result-based (χ^2 + CV) approach.

English Queries	Extracted Chinese Translations
Ma Yo-yo	馬友友 (transliteration in Chinese), 大提琴家 (violoncello player), 具聲望華裔 (a famous non-Chinese citizen of Chinese origin), 巴西情迷 (one of Ma's concerts)
Disneyland	迪士尼樂園 (disneyland), 東京迪士尼 (disney Tokyo), 迪士尼, 迪斯耐, 迪斯奈 (disney; transliterations with different Chinese characters), 加州迪士尼 (disney California)
CIH	病毒 (virus), 陳盈豪 (designer's name), 電腦病毒 (computer virus), 防毒 (anti-virus), 防毒軟體 (anti-virus software), 掃毒 (virus scanning), 巨集病毒 (macro virus)
NBA	職籃 (professional basketball game), 籃球 (basketball), 美國職籃 (professional basketball game in U.S.), 美國職業籃球聯盟 (professional basketball league in U.S.)

Table 4: The MAP values obtained by applying different translation approaches to the NTCIR-2 English-Chinese retrieval task with respect to different top-ranked *K* translations.

Approaches of Query Translation		Mean Average Precision (*MAP*)					
		K = 1	*K* = 3	*K* = 5	*K* = 7	*K* = 9	*K* = 11
Monolingual		0.387	none	none	none	none	none
Crosslingual	Dictionary-Based	0.198	0.203	0.207	0.214	0.215	0.215
	Search-Result-Based	0.176	0.218	0.237	0.245	0.259	0.262
	Search Result + Anchor-Text	0.184	0.222	0.241	0.248	0.264	0.267
	Search Result + Anchor-Text + Dictionary	0.241	0.264	0.271	0.273	0.278	0.279

varied from 1 to 11. Table 4 shows the results. The approaches using Web corpora performed better than the dictionary-based approach, and the approach combining with all resources achieved the best performance at 0.279 when *K*=11. Comparing the dictionary-based approach with the hybrid approach, we found that the problem of missing translations from the dictionary was obvious for the NTCIR-2 title queries. The improvements varied from 0.043 (*K*=1) to 0.064 (*K*=11). The only one case that the dictionary-based approach performed better than the approaches using search-result corpus was when *K*=1 because only a small number of the title query terms (12.36%, 22/178) had no translations. In the other cases, the overall retrieval performance was visibly improved due to the fact that the approaches using search-result corpus were effective in finding the correct and relevant translations of the unknown short queries. For example, the query *"Disneyland"* (Q23) increased the MAP value from 0 to 0.721 due to the translations, e.g.,*"迪士尼樂園," "東京迪士尼," "迪士尼," "迪斯耐,"* and *"迪斯奈"* (as shown in Table 3), which were not included in the LDC lexicon. Similarly, the query "*Ma Yo-yo cello recital*" (Q46) was correctly translated into *"馬友友演奏會"* and increased the MAP value from 0.205 to 0.446. Note the achieved MAP values can be further improved. Since the main purpose of these experiments was to examine if the proposed approach can help conventional approaches for CLIR, we simply used some basic techniques of query expansion and phrase translation in our experiments. Many advanced techniques as reported in [7] have not been adopted yet.

4.2 Translation of Web Query Terms

We collected Web queries from two real-world Chinese search engine logs in Taiwan, i.e. *Dreamer* and *GAIS*[7]. The Dreamer log contained 228,566 unique query terms from a period of over 3 months in 1998, while the GAIS log contained 114,182 unique query terms from a period of two weeks in 1999. Two different test query sets were prepared based on the two logs. The first set, called

[7] http://gais.cs.ccu.edu.tw

Table 5: Inclusion rates for different test Web query sets using the different approaches.

Query Types	Approaches		Dic			OOV			All		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Popular Web Queries	HK Parallel Corpus		11.5%	15.7%	16.9%	1.18%	1.78%	2.36%	7.20%	9.77%	10.4%
	Web Corpora	χ^2	42.1%	57.9%	62.1%	40.2%	53.8%	56.2%	41.4%	56.3%	59.8%
		CV	51.7%	59.8%	62.5%	45.0%	55.6%	57.4%	49.1%	58.1%	60.5%
		χ^2 + CV	52.5%	60.4%	63.1%	46.1%	56.2%	58.0%	50.7%	58.8%	61.4%
		AT	47.1%	66.7%	70.1%	60.4%	69.8%	71.6%	52.3%	67.9%	70.7%
		χ^2 + CV +AT	56.7%	71.6%	73.6%	50.3%	75.7%	77.5%	54.2%	73.3%	75.1%
Random Web Queries	HK Parallel Corpus		12.5%	17.5%	27.5%	2.50%	2.50%	2.50%	6.00%	8.00%	12.0%
	Web Corpora	χ^2	35.0%	50.0%	55.0%	35.0%	50.0%	51.7%	35.0%	50.0%	53.0%
		CV	40.0%	52.5%	55.0%	36.7%	55.0%	58.3%	38.0%	54.0%	55.0%
		χ^2 + CV	45.0%	50.0%	57.5%	40.0%	58.3%	60.0%	42.0%	55.0%	59.0%
		AT	42.5%	62.5%	67.5%	26.7%	33.3%	35.0%	33.0%	45.0%	48.0%
		χ^2 + CV +AT	50.0%	67.5%	70.0%	45.0%	60.0%	63.3%	47.0%	64.0%	70.0%

the *popular-query set*, contained a set of 430 frequent English query terms. These query terms were obtained from the 1,230 English terms out of the most popular 9,709 query terms (with frequencies above 10 in both logs). The popular-query set was further divided into two types: type *Dic* (the terms existing 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 100 English query terms, which were randomly selected from the top 19,124 queries in the Dreamer log. About 60% of the randomly-selected English query terms were not included in the LDC English-Chinese lexicon.

Table 5 (with the same format as Table 2) shows the results in terms of the top 1-5 inclusion rates for translation of the Web query terms. The domain-specific corpus (HK parallel corpus) was not suitable for translating unknown queries, especially for diverse Web queries. It only reached 1%-2.5% inclusion rates for OOV. However, using the Web as the corpus the proposed search-result-based approach (*χ2+CV*) achieved 46% top-1 inclusion rate for popular Web queries (58% in the top-5) and 40% top-1 inclusion rate for random Web queries (60% in the top-5). The performance of the translation of popular Web queries was better than that of random Web queries because random Web queries were too diverse. Note that the proposed search-result-based approach produced better translations than the anchor-text-based approach for the random Web queries. This might be resulted from that the collected anchor-text corpus was not sufficient and the translation coverage of anchor texts was limited for some particular domains.

5. DISCUSSION

Flexibility for query specification: In many CLIR applications, it is difficult to specify 'correct' queries in source languages for searching relevant documents in target languages - especially for particular domains such as disease names. For example, suppose we want to locate English documents pertaining to *"severe acute respiratory symptom"* by Chinese. Instead of forming the complex Chinese query *"嚴重急性呼吸道症候群"* we can specify a free text *"發燒 38 度"* (*have a temperature*, $38^{O}C$) and *"肺炎"* (*pneumonia*) as a combined description of the query, the proposed approach might correctly translate it to *"SARS," "severe acute respiratory symptom"* and *"severe acute breathe symptom."* Compared with the conventional query translation approaches, the proposed search-result-based approach provides more flexibility and convenience for query specification. Most search engines return results with confidence or relevancy rankings with respect to a given query. Not only the query but also its relevant terms may frequently co-occur with its correct translations in the search-result pages. Hence, in the proposed approach the query is not restricted to its original expression in the source language.

Moreover, search-result pages are dynamic and allow new words to be effectively translated such as "*SARS*" mentioned above, which only started to appear recently. This is of great practical importance because practical CLIR tasks, e.g., cross-language Web search, still suffer from a major drawback of lacking up-to-date translation dictionaries.

Translation effectiveness: Based on the observations on the experiments, it was found that the search-result-based approach is feasible for translating unknown query terms such as proper names, technical terms, and Web query terms. The query terms may be in the form of word, phrase, abbreviation, or free text.

The proposed approach is applicable to some other language pairs and is not limited to the Chinese and English. To examine its effectiveness in the English-to-Japanese and English-to-Korean translation, we did a preliminary evaluation. 50 scientists' names and 50 disease names in English were randomly selected from 256 scientists (Science/People) and 664 diseases (Health/Diseases and Conditions) in the Yahoo! Directory[8], respectively. The obtained results show that for the English-to-Japanese translation, the top-1, top-3, top-5 inclusion rates were 35%, 52%, and 63%, respectively; for the English-to-Korean translation, the top-1, top-3, top-5 inclusion rates were 32%, 54%, and 63%, respectively.

The proposed approach is also capable of translating a query term with multiple meanings if the occurrence frequency of each of its

[8] http://www.yahoo.com

translations is high enough on the Web. For example, the query term *"Juguar"* can be translated into *"積架"* (*Juguar car*) and *"美洲虎"* (*Juguar*, a kind of large animal of the cat family). This, however, depends on the number of snippets retrieved by the approach. When fewer snippets are considered having possible translations, some translations may be missing.

The proposed search-result approach, unavoidably, has some drawbacks. It might not perform good at the translation of terms that do not frequently co-occur with their translations in the search-result pages such as some common terms, and is dependent on the performance of the employed search engines. The translation extraction process of it might not be effective for language pairs that do not exhibit the mixed language characteristic on the Web.

Application: The search-result-based approach does not require crawling the Web and downloading Web documents. Taking the advantage of search engines' quick responses, we have developed an experimental system with the proposed approach, called *LiveTrans* (http://livetrans.iis.sinica.edu.tw/lt.html), to provide online English translation service of query terms for several Asian languages. The LiveTrans system is an experimental meta-search engine that provides cross-language search for retrieval of both Web pages and images. The system suggests a list of target translations to help users finding the translations of unknown query terms. The response time of query translation is in general within several seconds.

6. CONCLUSION

Finding translations in general dictionaries for CLIR encounters the problems of the translation of unknown queries - especially for short queries and the availability of up-to-date lexical resources. This paper has proposed an approach to automatically translate unknown queries for CLIR using the dynamic Web as the corpus. The obtained experimental results have shown its effectiveness in efficiently generating translation equivalents of various unknown query terms and improving retrieval performance for conventional CLIR approaches.

ACKNOWLEDGES

This work was partially supported by the National Science Council, Taiwan, under contact No. NSC93-2422-H-001-0003. We thank Sukil Kim M.D. and Shih-Jui Lin for their support of this work in examining Japanese and Korean translations.

REFERENCES

[1] Chien, L.-F, Huang, T.-I., and Chien, M.-C. PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. In *Proc. of ACM-SIGIR*, pp. 50-58, 1997.

[2] Fung, P. and Yee, L.Y. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proc. of COLING-ACL*, pp. 414-420, 1998.

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

[4] Hull, D.A. and Grefenstette, G. Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In *Proc. of ACM-SIGIR*, pp. 49-57, 1996.

[5] Kilgarriff, A. and Grefenstette, G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3), pp. 333-348, 2003.

[6] Kupiec, J. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proc. of ACL*, pp. 17-22, 1993.

[7] Kwok, K.L. NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS. In *Proc. of NTCIR Workshop Meeting*, 2001.

[8] Lavrenko, V., Choquette, M., and Croft, W.B. Cross-Lingual Relevance Models. In *Proc. of ACM-SIGIR*, pp. 175-182, 2002.

[9] 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(2), pp. 242-269, 2004.

[10] Melamed, I.D. Models of Translational Equivalence among Words. *Computational Linguistics* 26(2), pp. 221-249, 2000.

[11] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proc. of ACM-SIGIR*, pp. 74-81, 1999.

[12] Oard, D.W. and Diekema, A.R. Cross-Language Information Retrieval. In *Annual Review of Information Science and Technology.* American Society for Information Science, 1998.

[13] Pu, H.-T., Chuang, S.-L., and Yang, C. Subject Categorization of Query Terms for Exploring Web Users' Search Interests. *Journal of the American Society for Information Science and Technology* 53(8), pp. 617-630, 2002.

[14] Rapp, R. Automatic Identification of Word Translations from Unrelated English and German Corpora, In *Proc. of ACL*, pp. 519-526, 1999.

[15] Resnik, P. Mining the Web for Bilingual Text. In *Proc .of ACL*, pp. 527-534, 1999.

[16] 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, pp. 113-132, 1999.

[17] Silverstein, C., Henzinger, M., Marais, H., and Morics, M. Analysis of a Very Large AltaVista Query Log. *Technical Report* 1998-014, Digital Systems Research Center, 1998.

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

[19] Voorhees, E.M. and Harman, D.K. Overview of the sixth Text Retrieval Conference TREC-6. In *Proc. of the 6th Text Retrieval Conference*, 1998.

[20] Wang, J.-H., Teng, J.-W., Cheng, P.-J., Lu, W.-H, and Chien, L.-F. Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries*, 2004

[21] Xu, J., Weischedel, R., and Nguyen, C. Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval. In *Proc. of ACM-SIGIR*, pp. 105-110, 2001.

[22] Yang, C.C. and Li, K.W. Automatic Construction of English/Chinese Parallel Corpora. *Journal of the American Society for Information Science and Technology* 54(8), pp. 730-742, 2003.