LiveTrans: Translation Suggestion for Cross-Language Web Search from Web Anchor Texts and Search Results

Wen-Hsiang Lu[1,2], Lee-Feng Chien[1] and Hsi-Jian Lee[2]

1. Institute of Information Science, Academia Sinica, Taiwan, ROC

2. Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, ROC

{whlu, lfchien}@iis.sinica.edu.tw, {whlu, hjlee}@csie.nctu.edu.tw

Abstract

In this paper we will present a system, called LiveTrans, which can generate translation suggestions for given user queries and provide an English-Chinese cross-language search service for the retrieval of both Web pages and images. The system effectively utilizes two kinds of Web resources: anchor texts and search results. The developed anchor-text-based and search-result-based methods are complementary in the precision and coverage rates and promising in extracting translations of unknown query terms that were not included in general-purpose translation dictionaries. Experimental results demonstrate the feasibility of the system.

1. Introduction

To deal with automatic construction of translation lexicons, conventional research on machine translation (MT) [3] and cross-language information retrieval (CLIR) [1, 5, 7, 10, 13, 18] has generally used statistical techniques to automatically extract word translations from domain-specific parallel/comparable bilingual texts, such as bilingual newspapers [4, 11, 12, 20, 21]. However, only a certain set of their translations can be extracted through corpora with limited domains. In our research, we are interested in extracting translations of technical terms and proper names in diverse subjects, which are especially needed in performing CLIR services for Web users, e.g., “Hussein” (海珊/哈珊/侯賽因), “SARS” (嚴重急性呼吸道症候群). Existing CLIR systems usually rely on bilingual dictionaries for query translation [1, 13, 15]. Unfortunately, our analysis of Dreamer query log collected in Taiwan (see Section 3.1) showed that 74% of the 20,000 high frequent Web queries can not be found in general-purpose English-Chinese dictionaries (they are called *unknown terms* in this paper). How to automatically find translations for unknown terms, therefore, has become a major challenge for cross-language Web search.

Different from previous works, we focus on investigating new approaches to mining multilingual Web resources [19]. We have proposed a novel approach to extracting translations of Web queries through the mining of Web anchor texts and link structures [16, 17]. An anchor text is the descriptive part of an out-link of a Web page used to provide a brief description of the linked page. A variety of anchor texts in multiple languages might link to the same pages from all over the world. For example, Figure 1 shows a typical example, in which there are a variety of anchor texts in multiple languages linking to the Yahoo! from all over the world. Such a bundle of anchor texts pointing together to the same page is called an *anchor-text set*. Web anchor-text sets may contain similar description texts in multiple languages. Thus, for an unknown term appearing in some anchor-text sets, it is likely that its corresponding target translations appear together in the same anchor-text sets.

However, discovering translation knowledge from the Web has not been fully explored. In this paper, we intend to investigate another kind of Web resource, *search results*, and try to combine them with the anchor texts to benefit term translation. Chinese pages on the Web consist of rich texts in a mixture of Chinese (main language) and English (auxiliary language), and many of them contain translations of proper nouns. According to our observations, many search result pages in Chinese Web usually contain snippets of summaries in a mixture of Chinese and English. For example, Figure 2 illustrates the search-result page of the English query “National Palace Museum,” which was submitted to Google for searching Chinese pages, could obtain many relevant results containing both the query itself and its Chinese aliases. To explore search results on extraction of term translation, we have employed two methods: the chi-square test and context-vector analysis.

Based on a novel integration of the developed anchor-text- and search-result-based methods, we implemented an experimental system, called LiveTrans, to provide English-Chinese translation suggestion and cross-lingual retrieval of both Web pages and images. The purpose of this paper is to introduce our experiences in developing the methods and implementing the system.

2. Related Work

Term translation extraction is an important research problem in the context of MT. A number of related researches [12, 21] have used sentence-aligned parallel corpora to extract translations since the advent of statistical translation model [3]. Although high accuracy can be easily achieved by these techniques, sufficiently large parallel corpora for various subject domains and language-pairs are still hard to be available. On the other hand, some work has been done on term translation extraction from comparable or even unrelated texts [11, 20]. However, using non-parallel corpora is more difficult to effectively extract translations than parallel corpora due to the lack of parallel correlation aligned between documents or sentence pairs.

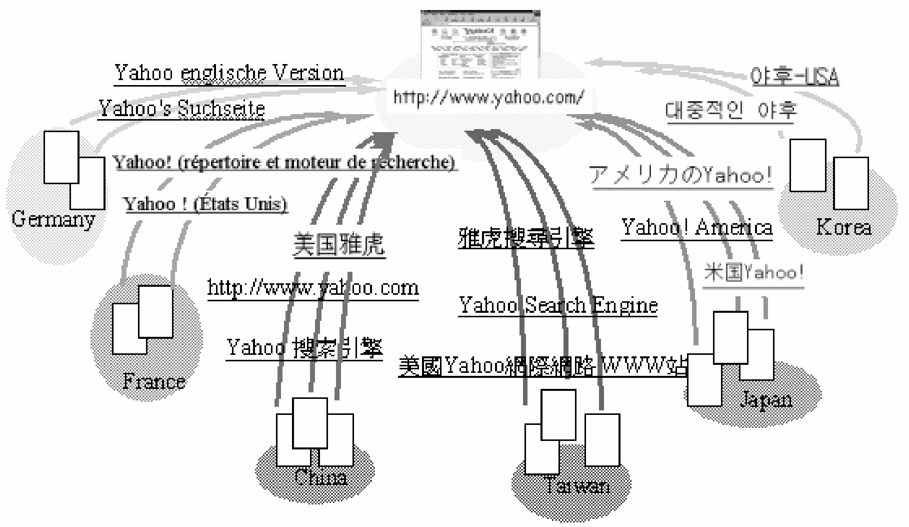
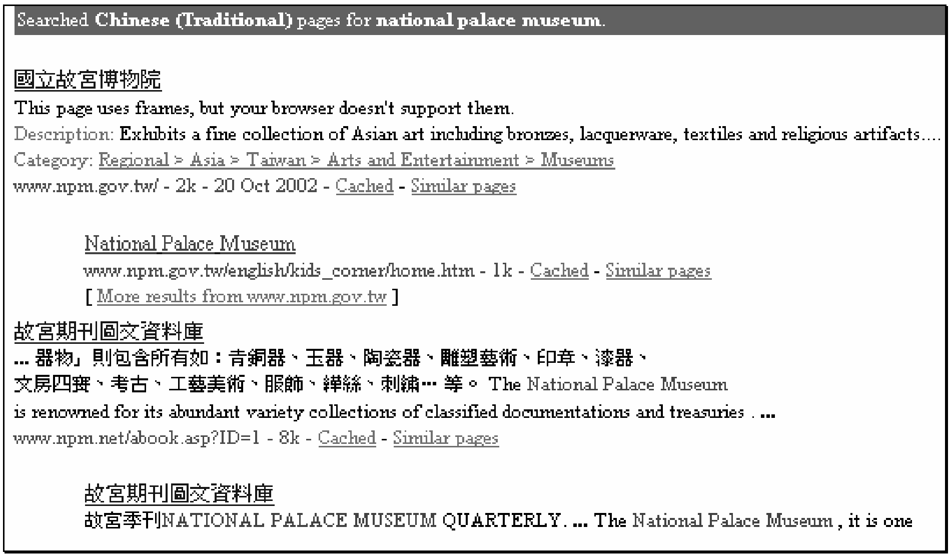


Figure 1. An illustration showing various anchor texts in multiple languages linking to Yahoo! from all over the world [17].



Searched **Chinese (Traditional)** pages for **national palace museum**.

國立故宮博物院
This page uses frames, but your browser doesn't support them.
Description: Exhibits a fine collection of Asian art including bronzes, lacquerware, textiles and religious artifacts....
Category: Regional > Asia > Taiwan > Arts and Entertainment > Museums
www.npm.gov.tw/ - 2k - 20 Oct 2002 - Cached - Similar pages

National Palace Museum
www.npm.gov.tw/english/kids_corner/home.htm - 1k - Cached - Similar pages
[More results from www.npm.gov.tw]

故宮期刊圖文資料庫
... 器物」則包含所有如：青銅器、玉器、陶瓷器、雕塑藝術、印章、漆器、
文房四寶、考古、工藝美術、服飾、緙絲、刺繡… 等。 The National Palace Museum
is renowned for its abundant variety collections of classified documentations and treasuries . ...
www.npm.net/abook.asp?ID=1 - 8k - Cached - Similar pages

故宮期刊圖文資料庫
故宮季刊NATIONAL PALACE MUSEUM QUARTERLY. ... The National Palace Museum , it is one

Figure 2. An illustration showing translation equivalents, such as "National Palace Museum"/國立故宮博物院 (故宮), which are included in a search result page returned from Google.

On the other hand, CLIR has become an important topic in recent research on information retrieval, however, practical cross-language Web search services have not lived up to expectations. This task must face a number of challenges, especially the problem of query translation. To deal with such problem, existing CLIR systems mostly rely on bilingual dictionaries. These dictionary-based techniques are limited in real-world applications since queries often contain unknown query terms, such as personnel names and technical terms [15]. Although some methods integrating dictionary-based techniques with parallel-corpus disambiguation, technology have been proposed and achieved performance improvements [1, 13]. Nonetheless, the unavailability of translations of unknown Web queries in diverse subjects is still a thorny problem.

A page[1] modified by Oard lists some CLIR retrieval systems, which can be either used on the Internet or obtained from commercial sources. For example, the Multilingual Summarization and Translation (MuST[2]) system is a Web-accessible CLIR system that uses English queries to search Indonesian, Spanish, Arabic and Japanese. MTIR is a demonstration search system that accepts queries in Chinese, finds documents in English, and then translates the selected documents into Chinese [1]. These systems generally rely on built-in bilingual dictionaries for query translation. To our knowledge, the proposed LiveTrans system is one of the few CLIR systems which allow the translations of unknown queries to be extracted through the mining of Web resources.

3. LiveTrans System

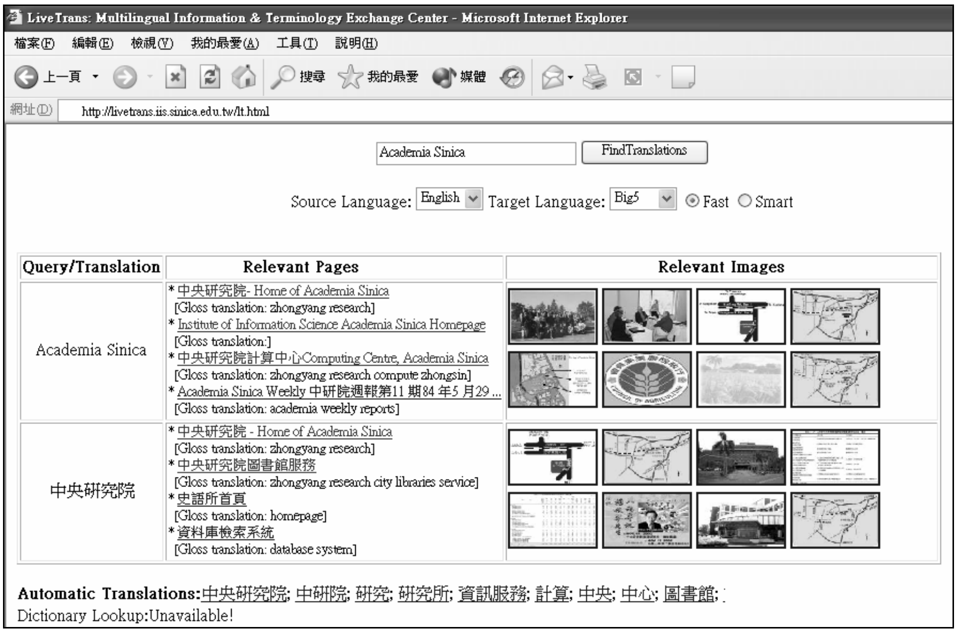
The LiveTrans[3] system is an experimental meta-search engine that provides English-Chinese translation suggestion and cross-language search for retrieval of both Web pages and images. It was implemented based on a novel combination of the developed Web mining methods. To use the system, users may select either English, traditional Chinese or simplified Chinese as the source/target language. For each input source query, the system will suggest a list of target translations. Since real queries are often short, there is a lack of context information needed to perform query translation. The system combines the term translation extraction methods and bilingual lexicons to make suggestions. The users can select the preferred translation and the system will return the retrieved Web pages and images, and sort them in their order of decreasing relevance to the corresponding translated queries. The titles of the retrieved pages are also translated word by word to the source languages for reference (i.e. gloss translation). Like most of the meta-search engines, backend engines can be chosen and the retrieved results can be merged using a data fusion technique. The system has been used to collect translation equivalents of a certain portion of users' queries. Many of the obtained translations are really not easy for human indexers to compile. For example, in the case shown in Figure 3, the user selected English as the source language and Chinese as the target language. In this example, the given query was "Academia Sinica" and its translations were extracted, i.e., 中央研究院 and 中研院.

We sometimes refer to the Web as a globally interconnected information infrastructure. At present, however, for someone who reads only English, it is presently the English-Wide-Web, and a reader of only Chinese sees only the Chinese-Wide-Web. With the LiveTrans system, it is easy to see that there are a number

[1] http://raveb.umd.edu/ddlrg/clir/systems.html

[2] http://www.isi.edu/natural-language/projects/C-ST-RD.html

[3] http://livetrans.iis.sinica.edu.tw/lt.html


LiveTrans: Multilingual Information & Terminology Exchange Center - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 媒體
網址(D)

http://livetrans.iis.sinica.edu.tw/lt.html

Academia Sinica [FindTranslations]

Source Language: English Target Language: Big5 ● Fast ○ Smart

Query/Translation	Relevant Pages	Relevant Images
Academia Sinica	* 中央研究院- Home of Academia Sinica [Gloss translation: zhongyang research] * Institute of Information Science Academia Sinica Homepage [Gloss translation:] * 中央研究院計算中心Computing Centre, Academia Sinica [Gloss translation: zhongyang research compute zhongsin] * Academia Sinica Weekly 中研院週報第11 期84 年5 月29 ... [Gloss translation: academia weekly reports]	
中央研究院	* 中央研究院 - Home of Academia Sinica [Gloss translation: zhongyang research] * 中央研究院圖書館服務 [Gloss translation: zhongyang research city libraries service] * 史語所首頁 [Gloss translation: homepage] * 資料庫檢索系統 [Gloss translation: database system]	

Automatic Translations:中央研究院; 中研院; 研究; 研究所; 資訊服務; 計算; 中央; 中心; 圖書館;

Dictionary Lookup:Unavailable!

Figure 3. An example showing the search results retrieved by the LiveTrans system, where the given query was “Academia Sinica” and its translations extracted were 中央研究院, 中研院.

of cases where Chinese users need English-Chinese cross-language translation. In fact, the LiveTrans system was found to be effective in increasing the recall rate of Web search, especially for the retrieval of Web images. Requests for images often are not limited to the local environment. For example, for the original query 羅浮宮 (Louvre) in Chinese, it could retrieve only hundreds of Web images, but it could retrieve hundreds of thousands images through its English translation.

With the novel combination of the developed Web mining methods (see Section 4), the LiveTrans system could provide effective translation suggestions for users selecting the ‘Smart’ mode; however, it cannot perform efficiently in real time due to its computation complexity. To obtain query translation instantly, the user is recommended selecting the ‘Fast’ mode with a little loss of accuracy. To remain the accuracy, the system can constantly update translations for new queries in the query log in a batch. Therefore, the system can effectively provide translation suggestions and cross-lingual search services.

4. Query Translation from Anchor Texts and Search Results

To implement a query translation process via mining the Web resources: anchor texts and search results, three major processing steps are required:

(1) Corpus collection: Collect bilingual Web data as a comparable corpus.

(2) Translation candidate extraction: Extract translation candidates from the collected corpus.

(3) Translation selection: Estimate the similarity for each candidate and determine the most possible translations.

To effectively handle this process, we have developed two kinds of methods: the anchor-text-based method and the search-result-based method. The details regarding the two methods will be presented in the following.

4.1 The Anchor-Text-Based Method

Query translation from anchor texts contains three major computational modules: anchor-text extraction, translation candidate extraction, and translation selection. The anchor-text extraction module was constructed to collect pages from the Web and build up a corpus of anchor-text sets. For each given query term, the translation candidate extraction module extracts key terms in the target language as the translation candidates from the anchor-text sets containing the query term. The effectiveness of the adopted term extraction methods greatly affects the performance in extracting correct translations. Three different methods have been tested in our previous work [17]: the PAT-tree-based (a statistics-based n-gram model [9]), query-set-based and tagger-based methods. Among them, the query-set-based method has been adopted in this paper because it could extract longer terms (i.e. multi-words) and have less problems of Chinese term segmentation than the other methods. This method uses query logs in the target language as the translation vocabulary set to segment anchor texts and extract key terms. The pre-condition for using this method is that the coverage of the query set should be high. Finally, the translation selection module selects the possible translation that maximizes the estimation based on the probabilistic inference model described below.

4.1.1 The Probabilistic Inference Model

To find the most probable translation *t* for a query term *s*, we have proposed probabilistic inference model to utilize Web anchor texts and hyperlink structures. This model is used to estimate the probability value between a query term and each translation candidate that co-occurs with the query term in the same anchor-text sets. The estimation assumes that anchor texts linking to the same pages may contain similar terms with analogous concepts. Therefore, a candidate has a higher chance of being an correct translation if it is written in the target language and frequently co-occurs with the query term in the same anchor-text sets. In addition, in the field of Web research, it has been proven that link structures can be used effectively to estimate the authority of Web pages [2, 14]. Our model further assumes that the translation candidates in the anchor-text sets of pages with higher authority may be more reliable. For a Web page (or URL) u_i, its anchor-text set $AT(u_i)$ is defined as consisting of all of the anchor texts of the links pointing to u_i, i.e., u_i 's in-links.

The similarity estimation function based on the probabilistic inference model is called model S_{AT} for the sake of usage consistency in the consequent sections and is defined below:

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

The above measure is adopted to estimate the degree of similarity between source term *s* and target translation *t*. The measure is estimated based on their co-occurrence in the anchor text sets 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, $P(u_i)$ is estimated along with the probability of u_i being linked, and its estimation is defined as follows: $P(u_i)= L(u_i)/\Sigma_{j=1,n} L(u_j)$, where $L(u_j)$ indicates the number of in-links of page u_j.

In addition, we assume that *s* and *t* are independent given u_i; then, the joint probability $P(s \cap t|u_i)$ is equal to the product of $P(s|u_i)$ and $P(t|u_i)$, and the similarity measure becomes

$$S_{AT}(s,t) \approx \frac{\sum_{i=1}^{n} P(s \mid u_i) P(t \mid u_i) P(u_i)}{\sum_{i=1}^{n} [P(s \mid u_i) + P(t \mid u_i) - P(s \mid u_i) P(t \mid u_i)] P(u_i)}. \quad (2)$$

The values of $P(s|u_i)$ and $P(t|u_i)$ are estimated by calculating the fractions of the numbers of u_i's in-links containing *s* and *t* over $L(u_i)$, respectively. Therefore, a candidate translation has a higher confidence value for being an effective translation if it frequently co-occurs with the source term in the anchor-text sets of those pages having higher authority. For details about the probabilistic inference model, readers may refer to our previous work [17].

4.2 The Search-Result-Based Method

Query translation from search results also contains three major computational modules: search-result collection, translation candidate extraction, and translation selection. In the search-result collection module, a given source query is submitted to a real-world search engine to collect the top search result pages. In the translation candidate extraction module, we use the same term extraction method adopted in the anchor-text-based method. In the translation selection module, our idea is to utilize co-occurrence and context information between source queries and target translation candidates to estimate their semantic similarity and to determine the most possible translations. We have investigated several different methods of estimation and found that the chi-square test and context vector analysis achieve better performance.

4.2.1 The Chi-Square Test

A number of statistical measures have been proposed for estimating the association between words/phrases based on co-occurrence analysis, including mutual information, the DICE coefficient, and statistical tests, such as the chi-square test and the log-likelihood ratio test [12, 20, 21]. Although the log-likelihood ratio test is suitable for dealing with the data sparseness problem, in our preliminary experiments on 430 popular Web queries (see Section 5.1), we found that the chi-square test performs better than the log-likelihood ratio test. One of the possible reasons is that the required parameters for the chi-square test can be effectively obtained from real-world search engines, and is enough to avoid the data sparseness problem. The chi-square test was, therefore, adopted as the major method for co-occurrence analysis in our work. Its similarity measure is defined as

$$S_{\chi 2}(s,t) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}, \quad (3)$$

where a, b, c and d are the numbers in the four cells of the contingency table (see Table 1) for the source term *s* and target term *t* and are defined as follows:

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*.

Table 1. A contingency table.

	t	*~t*
s	*a*	*b*
~s	*c*	*d*

The required parameters for the chi-square test can be computed using the search results returned from real-world search engines. Most search engines accept Boolean queries and can report the number of pages matched.

4.2.2 The Context-Vector Analysis

Co-occurrence analysis is applicable to frequent query terms because these terms are more likely to appear with their translation candidates. On the other hand, infrequent query terms have little chance of appearing with translation candidates in the same pages. The context-vector-based method has been used to extract translations from comparable corpora [11, 20], and is thus adopted to deal with this problem. Different from previous works using a translation lexicon to bridge the features with the same meaning in different languages, we use only popular query terms as the feature set, because of the advantage of updating the feature set with queries in diverse subjects continuously supplied by Web users. This is a suitable way to provide effective feature sets to represent context vectors of diverse unknown query terms and their translation candidates. For each query or candidate term, we take the co-occurring feature terms as its context vector since translation equivalents may share the same occurring feature terms. The similarity between a query term and each translation candidate can be computed based on their context vectors. Thus, infrequent query terms still have a chance of extracting translations.

Like Fung et al.'s vector space model, we also use the TF-IDF weighting scheme to estimate the significance of each feature in the context vector and use the cosine measure to calculate the translation similarity of each query term and its translation candidates. The weighting scheme is defined as follows:

$$w_{t_i} = \frac{f(t_i,d)}{\max_j f(t_j,d)} \times \log(\frac{N}{\text{n}}), (4)$$

where $f(t_i,d)$ is the frequency of 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 including t_i.

Given the context vectors of a query term and each translation candidate, their similarity measure is estimated as follows:

$$S_{CV}(s,t) = \frac{\sum_{i=1}^{m} ws_i \times wt_i}{\sqrt{\sum_{i=1}^{m}(ws_i)^2 \times \sum_{i=1}^{m}(wt_i)^2}}. (5)$$

It is not difficult to construct context vectors for query terms and their translation candidates. For a query term, we can obtain search results by submitting it as a query to real-world search engines. Basically, we can use a fixed number of the top retrieved results (snippets) to extract translation candidates. The co-occurring feature terms of each query can also be extracted, and their weights calculated based on the retrieved snippets. The context vector of the query is, thus, constructed. The same procedure is used to construct a context vector for each translation candidate.

4.3 The Combined Method

Our previous experiments show that the anchor-text-based method can achieve a good precision rate for popular Web queries in other language pairs besides Chinese and English [17], but it has a major drawback; that is, the cost is relatively high to collect sufficient pages to extract anchor texts. Benefiting from real-world search engines, the search-result-based method can achieve a good coverage rate for diverse query terms. However, method using the chi-square test has difficulty in dealing with infrequent query terms, and the method using context-vector analysis needs to carefully handle the issue of feature selection. Intuitively, a more complete solution is to integrate the three different methods. Under consideration of the large difference of ranges of similarity values among the three methods, we use a linear combination weighting scheme to compute the similarity measure as follows:

$$S_{COMBINED}(s,t) = \sum_{m} \frac{\alpha_m}{R_m(s,t)}, \quad (6)$$

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 translation candidate t with respect to the source term s, is assigned to be from 1 to k (candidate number) in decreasing order by similarity measure $S_m(s,t)$.

5. Experimental Results

5.1 The Test Bed

To determine the effectiveness of the developed methods to Web query translation, we conducted several experiments on extracting English translations for Chinese queries. We collected real query terms along with 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 from a period of over 3 months in 1998, and the GAIS log contained 114,182 unique query terms from a period of two weeks in 1999. A query set, called the *popular-query set*, was prepared to test the translation effectiveness for unknown Web queries. There were 9,709 most popular query terms whose frequencies were above 10 in the two logs, and 1,230 of them were English terms. After checking the logs, we obtained 430 terms whose Chinese translations appeared together in the logs and took their Chinese translations as the popular-query set. Table 2 lists some examples of the test query terms, which were divided into two types, where type Dic (the terms existing in the dictionary) made up about 36% (156/430) of the test queries, and type OOV (out of vocabulary; the terms not in the dictionary) made up about 64% (274/430).

In addition, to further investigate the translation effectiveness for proper names and technical terms, we also prepared two different query sets containing 50 scientist names and 50 disease names in English, which were randomly selected from the 256 scientists (Science/People) and 664 diseases (Health/Diseases and Conditions) in the Yahoo! Directory, respectively. It should be noted that 76% (38/50) scientist names and 72% (36/50) disease names are not included in the general-purpose translation dictionary which contains 202,974 entries collected from the Internet.

Table 2. Some sample test queries.

Type	Number	Sample test queries
Dic	156	銀行 (bank) 亞洲 (Asia) 愛滋病 (AIDS) 白宮 (White House) 世界貿易組織 (WTO)
OOV	274	電子商務 (E-commerce) 個人數位助理(PDA) 雅虎 (Yahoo) 太空總署 (NASA) 星際大戰 (Star War)

5.2 Web Data Collection

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, and which were taken as the anchor-text-set corpus for testing the anchor-text-based method. In addition, for testing the search-result-based method, we obtained search results of queries by submitting them to real-world Chinese search engines, such as Google Chinese[4] and Openfind[5]. Basically, we used only the first 100 retrieved results (snippets) to extract translation candidates. The context vector of each query was also extracted from the snippets. Also, the required parameters for the chi-square test were computed using the search results returned from the utilized search engines.

5.3 Performance of the Proposed Methods for Popular Query Terms

We carried out experiments to determine the performance of the proposed methods in extracting translations for the bilingual query set. To evaluate the performance of translation extraction, we used the *average top-n inclusion rate* as a metric. For a set of test queries, its top-n inclusion rate was defined as the percentage of queries whose effective translations could be found in the first *n* extracted translations. Also, we wished to know if the coverage of effective translations was high enough in the top search result pages for the real queries. The coverage rate was the percentage of queries whose effective translations could be found in the extracted translation candidate set.

Table 3 shows the obtained results in terms of top 1-5 inclusion rates and coverage rate. In this table, CV, χ^2, AT and Combined represent the context-vector analysis, chi-square test, anchor-text-based, and combined methods, respectively. In addition, Dic, OOV and All represent the terms existing in a dictionary, the terms not

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

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

in a dictionary, and the total query set, respectively. It is clear that the AT method and the combined method performed better than the χ^2 and CV methods in almost every case. The weights of the combined method were assigned according to the top-1 inclusion rates achieved by the three other methods, i.e., α_{cv} = 56.3%/(56.3%+49.5%+66.5%) ≈ 0.33. In fact, the obtained coverage rates were very high. This shows that the Chinese Web is rich in texts with a mixture of Chinese and English.

Table 3. Coverage and inclusion rates for popular Chinese queries using the 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%
AT	Dic	67.3%	78.2%	80.8%	89.1%
	OOV	66.1%	74.5%	76.6%	83.9%
	All	66.5%	75.8%	78.1%	85.8%
Combined	Dic	68.6%	82.1%	84.6%	92.3%
	OOV	66.8%	85.8%	88.0%	94.2%
	All	67.4%	84.4%	86.7%	93.5%

Table 4. Coverage and inclusion rates for popular English queries using the 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%
AT	57.1%	70.0%	71.9%	85.4%
Combined	59.4%	74.3%	76.2%	89.9%

The above popular-query set contained only Chinese queries. To determine the performance of the proposed methods in translating English queries into Chinese, we carried out another experiment which used the English translations of the same popular-query set as the test set. The results are shown in Table 4. The achieved performance was a little worse than achieved using the Chinese query set. The reason for this result was that the English queries had to deal with more ambiguous Chinese translation candidates since the search result pages returned from Chinese search engines normally contain mostly Chinese texts.

5.4 Performance of the Combined Method for Proper Names and Technical Terms

To further deal with the translation of proper names and technical terms, we conducted an experiment on the test sets of scientist names and medical terms mentioned in Section 5.1. According to our analysis of the test terms, many of scientist and disease names were not included in our collected query-log set, and some disease names were multi-words, e.g., “Hypoplastic Left Heart Syndrome” (左心發育不全症候群), “Lactose Intolerance” (乳糖不耐症), “Nosocomial Infections” (院內感染). Thus, we slightly modified the method of query-set-based translation candidate extraction by augmenting a simplified technique of unknown term and multi-word identification [6, 8]. As a result, the top-1 inclusion rate was obtained at 40% and 44% for the scientist and disease names, respectively (see Table 5). Some examples of the correct translations extracted using the combined method are shown in Table 6.

Table 5. 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%

5.5 Discussion

The translation accuracy achieved using the combined method is very promising, especially for popular queries. According to our analysis, this good performance was primarily due to the fact that the Chinese Web has a mixed language characteristic: many pages mainly consist of texts in Chinese (main language) with parts of texts in English (auxiliary language). The Chinese Web is considerably rich in texts containing English-Chinese translations of proper nouns, such as personal names and technical terms. As a result, this characteristic makes it possible to automatically extract English-Chinese translations of a large number of unknown query terms.

In fact, the translation process based on the search-result-based method might not be very effective for language pairs that do not exhibit the language-mixed characteristic on the Web. For this reason, the anchor-text-based method is still attractive while it achieves good precision rates for popular queries in other language pairs besides Chinese and English, even though not every particular pair of languages has sufficient texts on the Web.

The performance achieved using the combined method looks very promising, but it still has limitations. For example, it is less reliable in extracting translations of multi-word terms. To enhance the accuracy in translating multi-word or unknown terms, it should be worthy to employ more effective techniques, such as word segmentation and language model, to filter out noise terms and extract complete translation candidates. Currently, the LiveTrans system cannot perform efficiently in real time due to its computation complexity. This is a real challenge to improve the response time of query translation in our future work. However, the system can constantly update translations for new queries in the query log in a batch. Therefore, the system still can provide translation suggestions and cross-lingual search services.

Table 6. Some examples of the test proper names and technical terms, and their extracted translations.

Query Type	English Query	Extracted Chinese Translations
Scientist Name	Aldrin, Buzz (Astronaut)	艾德林
	Hadfield, Chris (Astronaut)	哈德菲爾德
	Galilei, Galileo (Astronomer)	伽利略/伽里略/加利略
	Ptolemy, Claudius (Astronomer)	托勒密
	Earhart, Amelia (Aviators)	鄂哈特
	Tibbets, Paul (Aviators)	第貝茲/迪貝茨
	Crick, Francis (Biologists)	克立克/克里克
	Drake, Edwin Laurentine (Earth Scientist)	德拉克
	Aryabhata (Mathematician)	阿耶波多/阿利耶波多
	Kepler, Johannes (Mathematician)	克卜勒/開普勒/刻卜勒
	Dalton, John (Physicist)	道爾頓/道耳吞/道耳頓
	Feynman, Richard (Physicist)	費曼
Disease Name	Ganglion Cyst	腱鞘囊腫
	Gestational Diabetes	妊娠糖尿病
	Hypoplastic Left Heart Syndrome	左心發育不全症候群
	Lactose Intolerance	乳糖不耐症
	Legionnaires' Disease	退伍軍人症
	Muscular Dystrophy	肌肉萎縮症
	Nosocomial Infections	院內感染
	Shingles	帶狀皰疹/帶狀庖疹
	Stockholm Syndrome	斯德哥爾摩症候群
	Sudden Infant Death Syndrome (SIDS)	嬰兒猝死症

6. Conclusion

Practical cross-language Web search services have not lived up to expectations since they suffer from a major problem where up-to-date multilingual lexicons containing the translations of popular Web queries, such as proper names and technical terms, are lacking. In this paper we present a promising system, called LiveTrans, which can generate translation suggestions for given user queries and provide an English-Chinese cross-language search service for the retrieval of both Web pages and images. The system effectively utilizes two kinds of live Web resources: anchor texts and search results, which are contributed continuously by a huge number of volunteers (page authors) around the world. The developed anchor-text-based and search-result-based methods are complementary in the precision and coverage rates and promising in extracting translations of query terms that were not included in general-purpose translation dictionaries.

References

[1] Bian, G. W. and Chen, H. H. (2000) Cross-Language Information Access to Multilingual Collections on the Internet, Journal of the American Society for Information Science, 51(3), 281-296.

[2] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of the 7th International World Wide Web Conference, 107-117.

[3] Brown, P., Pietra, S. A. D., Pietra, V. D. J., Mercer, R. L. (1993) The Mathematics of Machine Translation, Computational Linguistics, 19(2), 263-312.

[4] Chang, J. S., Yu, D., Lee, C. J. (2001) Statistical Translation Model for Phrases, Computational Linguistics and Chinese Language Processing, 6(2), 43-64.

[5] Chang, J. S., Ker, S. J. and Chen, M. H. (1998) Cross Language Information Retrieval and Data Mining, Proceedings of the Conference on Information Science and Technology-1998: Perspectives in the 21st Century, 153-166.

[6] Chang, J. S. and Su, K. Y. (1997) A Multivariate Gaussian Mixture Model for Automatic Compound Word Extraction, Proceeding of ROCLING X, 123-142.

[7] Chen, K. H. and Chen, H. H. (2001) The Chinese Text Retrieval Tasks of NTCIR Workshop 2, Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization.

[8] Chen, K. J. and Bai, M. H. (1998) Unknown Word Detection for Chinese by a Corpus-Based Learning Method, International Journal of Computational Linguistics and Chinese Language Processing, 3(1), 27-44.

[9] Chien, L. F. (1997) PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval, Proceedings of ACM-SIGIR ’97, 50-59.

[10] Dumais, S. T., Landauer, T. K., Littman, M. L. (1996) Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing, Proceedings of ACM-SIGIR’96 Workshop on Cross-Linguistic Information Retrieval, 16-24.

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

[12] Gale, W. A. and Church, K. W. (1991) Identifying Word Correspondances in Parallel Texts, Proceedings of DARPA Speech and Natural Language Workshop.

[13] Hiemstra, D. and de Jong, F. (1999) Disambiguation Strategies for Cross-language Information Retrieval, Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, pp. 274-293.

[14] Kleinberg, J. (1998) Authoritative Sources in a Hyperlinked Environment, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 46(5), 604-632.

[15] Kwok, K. L. (2001) NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS, Proceedings of NTCIR workshop meeting.

[16] Lu, W. H., Chien, L. F., Lee, H. J. (2001) Anchor Text Mining for Translation of Web Queries, Proceedings of the 2001 IEEE International Conference on Data Mining, 401-408.

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

[18] Lvarenko, V., Choquette, M., Croft, W. B. (2002) Cross-lingual Relevance Model, Proceedings of ACM-SIGIR 2002 Conference, 175-182.

[19] Nie, J. Y., Isabelle, P., Simard, M., and Durand, R. (1999) Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, Proceedings of ACM-SIGIR’99 Conference, 74-81.

[20] Rapp, R. (1999) Automatic Identification of Word Translations from Unrelated English and German Corpora, Proceedings of the 37th Annual Conference of the Association for Computational Linguistics.

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