LiveTrans-Cross-Language Web Search through Live Mining of Query Translations

Lee-Feng Chien
Institute of Information Science, Academia Sinica, Taiwan, ROC
E-mail：lfchien@iis.sinica.edu.tw

【Abstract】

Enabling users to find effective translations automatically for query terms not included in dictionary is one of the major goals of a practical cross-language Web search service. This paper presents a cross-language Web search system called LiveTrans, which is an experimental meta-search engine that provides English-Chinese cross-lingual retrieval of both Web pages and images. The system has been implemented based on a novel integration of Web mining approaches, including anchor-text-based and search-result-based approaches, to extract bilingual translations of real English and Chinese queries through the mining of Web resources. Experimental results demonstrate the feasibility of the system for providing a practical cross-language search service.

INTRODUCTION

Although cross-language information retrieval (CLIR) [6], which enables users to query in one language and retrieve relevant documents written or indexed in another language, has become an important topic in recent research on information retrieval, practical cross-language Web search services have not lived up to expectations. This paper presents a cross-language Web search system called LiveTrans (http://livetrans.iis.sinica.edu.tw/), which is an experimental meta-search engine that provides English-Chinese cross-lingual retrieval of both Web pages and images. The system has been implemented based on a novel integration of Web mining approaches, including anchor-text-based and search-result-based approaches, to extract bilingual translations of real English and Chinese queries through the mining of Web resources.

One of the major bottlenecks stems from the fact that up-to-date bilingual lexicons containing the translations of popular query terms, such as proper names and new terminology, are lacking in these services [2]. Unfortunately, our analysis of a query log showed that 74% of the top 20,000 popular Web queries in Taiwan, which formed 81% of the search requests in the log, could not be obtained from common English-Chinese translation dictionaries. How to find effective translations automatically for query terms not included in a dictionary, therefore, has become a major challenge for practical Web search services. To deal with the translation of unknown words, conventional research on machine translation has generally used statistical techniques to automatically extract word translations from domain-specific parallel bilingual texts, such as bilingual newspapers. Web queries are often diverse and dynamic. Only a certain set of their translations can be extracted through corpora with limited domains. Different from the previous works, we propose an alternative approach to performing Web query translation directly through mining of the Web's multilingual and wide-range resources. The Web is becoming the largest data repository in the world. Chinese pages on the Web (the Chinese Web) consist of rich texts in a mixture of Chinese

and English, and many of them contain bilingual translations of proper nouns, such as company names and personal names. This nice characteristic makes it possible for the English-Chinese bilingual translations of a large number of query terms to be automatically extracted.

To utilize such live sources for query translation, which are being added to by a huge amount of volunteers daily, we have developed several Web mining approaches to effectively exploiting two kinds of Web resources: anchor texts and search results. In these approaches, we employ several different term similarity estimation methods, such as the probabilistic inference model, context vector analysis and the chi-square test, to extract translation equivalents for unknown query terms. The purpose of this paper is to introduce our experiences obtained in developing the approaches and implementing the LiveTrans cross-language Web search system.

LiveTrans SYSTEM

The LiveTrans system is an experimental meta-search engine that provides English-Chinese cross-language search for retrieval of both Web pages and images. As shown in Figure 2, it was implemented based on a novel combination of the developed Anchor text mining and search result mining approaches. 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 approaches 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. Like most of the meta-search engines, backend engines can be chosen and the retrieved results can be merged using a data fusion technique.

To operate a practical system, the response time of query translation must be close to real time. However, neither the anchor-text-based approach nor the search-result-based approach can perform in real time. The system can generate translation equivalents for many queries in the query log using a batch mode and constantly update the effective translations for each new query term. It takes time to extract translations for a new query term with the combined approach. If it is necessary to extract query translations in an online mode, the system normally generates translations using the context vector approach. That is because the required feature terms can be fixed with a predefined query log and their feature vectors can be constructed in advance. In this way, it is possible to execute a query translation process within only a few seconds on a PC server. The system has been used to collect effective translations of a certain portion of users' queries. Most of the obtained translations are really not easy for human indexers to compile. For example, in the case shown in Fig. 1, the user selected English as the source language and Chinese as the target language. In this example, the given query was "national palace museum" and the extracted translations were "**國立故宮博物院**", "**故宮**" and "**故宮博物院**."

THE WEB MINING APPROACHES

To deal with the problem of term translation, we have developed two kinds of approaches: the anchor-text-based approach and the search-result-based approach.

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 a query 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

Figure 1 An example showing the search results retrieved by the LiveTrans system, where the given query was "national palace museum" and its translations extracted were "國立故宮博物院", "故宮" and "故宮博物院."

page. Such a bundle of anchor texts pointing together to the same page is called as an anchor-text set. In fact, Web anchor-text sets may contain similar description texts (or concepts) in multiple languages. It is likely that a number of word (or phrase) translations and synonyms can be extracted from them. Such anchor text sets that contain a given query term can be considered as composing a comparable corpus of translated texts for that query term. We have proposed an anchor-text-based approach to extracting translations of Web queries through the mining of Web anchor texts and link structures.

Although the anchor-text-based approach has been proven effective in extracting the translations of proper nouns in multiple languages, it nevertheless has a drawback that the translation process is not applicable for some query terms if the size of the collection of anchor texts is not large enough. For this reason, this paper presents search-result-based approaches to fully exploiting Web resources. These approaches take the search result pages of queries submitted to real search engines as the corpus for extracting query translations. Web search engines normally return search result pages with a long ordered list of relevant documents and snippets of summaries to help users locate interesting documents. According to our observations, a number of search result pages usually contain snippets of summaries in a mixture of Chinese and English, in which many translation equivalents of query terms are included. In our research, we seek to find out if the number of effective translations is high enough in the top search result pages for real queries. If it is, the search-result-based approaches can alleviate the difficulty encountered by the anchor-text-based approach. To determine the usefulness of search-result-based approaches, we have also investigated several different similarity estimation methods. The details of the two types of approaches will be presented in the following.

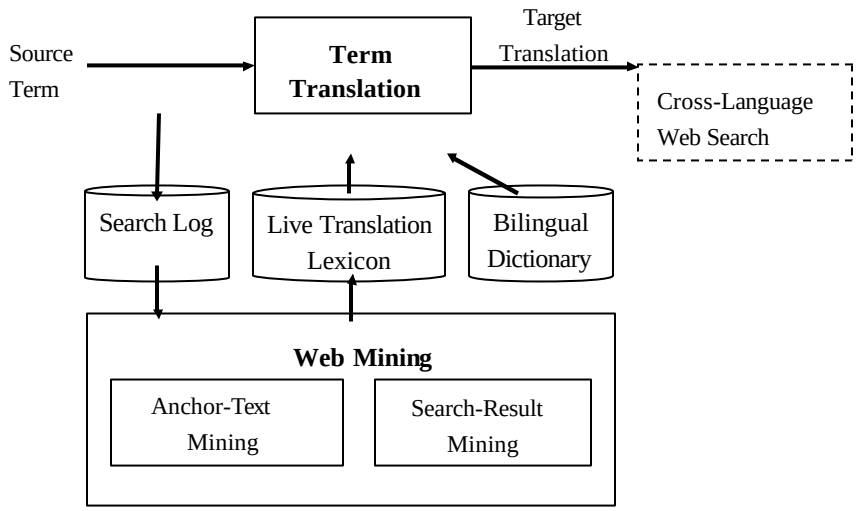

Figure 2 An abstract diagram showing the system architecture of the LiveTrans system

The Anchor-Text-Based Approach

To determine the most probable target translation *t* for source query term *s*, we have developed an anchor-text-based approach based on integrating hyperlink structure into probabilistic inference model. This model is used to estimate the probability value between a source query and all the translation candidates that co-occur 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 translation has a higher chance of being an effective translation if it is written in the target language and frequently co-occurs with the source 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. 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, ... 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: $P(u_i) = L(u_i)/S_{j=1,n} L(u_j)$, 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:

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

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.

The estimation process based on the probabilistic inference model contains three major computational modules: anchor-text extraction, translation candidate extraction, and translation selection. The anchor-text extraction module was constructed in order to collect pages from the Web and build up a corpus of anchor-text sets. For each given source term, the translation candidate extraction module extracts key terms in the target language as the translation candidate set from the anchor-text sets of those pages containing the source term. The effectiveness of the adopted term extraction methods greatly affects the performance in extracting correct translations. Three different methods have been tested: the PAT-tree-based, query-set-based and tagger-based methods. Among them, the query-set-based method has been strongly recommended because it has no problem with term segmentation. This method uses a query log in the target language as the translation vocabulary set to segment key terms in the anchor-text sets. The pre-condition for using this method is that the coverage of the query set should be high. Finally, the translation selection module extracts the translation that maximizes the similarity estimation. For details about the anchor-text-based approach, readers may refer to our previous work [3, 4].

The Search-Result-Based Approach

The estimation process based on the search-result-based approach 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 approach. In the translation selection module, our idea is to utilize co-occurrence relation and context information between source queries and target translations to estimate their similarity in semantics and to determine the most promising translations. We have investigated several different methods of estimation and found that the chi-square test and context vector analysis achieve the best performance.

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. Although the log-likelihood ratio test is theoretically more suitable for dealing with the data sparseness problem than the other measures, in our experiment, 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, which alleviates the data sparseness problem. The chi-square test was, therefore, adopted as the major method of co-occurrence analysis in our study. 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 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.

Context Vector Analysis

Co-occurrence analysis is applicable to higher frequency query terms because these terms are more likely to appear with their translation candidates. On the other hand, lower frequency query terms have little chance of appearing with other candidates in the same pages. The context vector method is thus adopted to deal with this problem. As translation equivalents may share similar occurring terms, for each query term, we take the co-occurring feature terms as its 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 of extracting correct translations. The context-vector-based method has also been used to extract translations from comparable corpora. Different from previous works, such as the use of Fung et al.'s seed word [1], we use real users' popular query terms as the feature set. This can help avoid the need to use 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 and use the cosine measure to calculate the translation probabilities of each source query and their target candidates. The weighting scheme is defined as follows:

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

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 source query term and each target translation candidate, their similarity measure is estimated as follows:

$$S_{cv}(s,t) = \frac{\sum_{i=1}^{m} f(w_{s_i}, w_{t_i})}{\sqrt{\sum_{i=1}^{m} (w_{s_i})^2 \times \sum_{i=1}^{m} (w_{t_i})^2}}. \quad (5)$$

It is not difficult to construct context vectors for source query terms and their translation candidates. For a source 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 using the snippets. The context vector of the query is, thus, constructed. The same procedure is used to construct a context vector for each of the extracted translation candidates.

The Combined Approach

Our previous experiments show that the anchor-text-based approach can achieve a good precision rate for popular queries and extract longer translations in other languages besides Chinese and English [23, 24], but it has a major drawback; that is, the cost of collecting a sufficient number of anchor texts, including the required software (e.g. spider) is very high to collect sufficient pages to extract anchor texts. Benefiting from real-world search engines, the search-result-based approach using the chi-square test can reduce the work of corpus collection but has difficulty in dealing with

low-frequency query terms. Although the search-result-based approach using context vector analysis can deal with the difficulties encountered by the above two approaches, it is not difficult to see it needs to carefully handle the feature selection issue. Intuitively, a more complete solution is to integrate the above three different approaches. Under consideration of the various ranges of similarity values among the above approaches, we use a linear combination weighting scheme to compute the similarity measure as follows:

$$S_{av}(s,t) = \sum_m \frac{\alpha_m}{R_m(s,t)}, \quad (6)$$

where a_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 (candidate number) in decreasing order by similarity measure $S_m(s,t)$.

EXPERIMENTS

The Test Bed

To determine the effectiveness of the proposed approaches to Web query translation, we conducted several experiments on extracting English target translations for source Chinese queries. We collected real query terms with a log from a real-world Chinese search engine in Taiwan, i.e., Dreamer. The Dreamer log contained 228,566 unique query terms from a period of over 3 months in 1998. We prepared the two different test query sets based on the two logs. A query set, called the *random-query set*, was prepared to test the translation effectiveness for common queries. The query set contained 50 query terms in Chinese, which were randomly selected from the top 20,000 queries in the Dreamer log. Forty of them were found to not be included in common translation dictionaries. It should be noted that the topics of the query terms could be very local, and that not all of them had target translations.

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 approach.

In addition, for testing the search-result-based approaches, we obtained search results of queries by submitting them to real world Chinese search engines, such as Google Chinese (http://www.google.com) and Openfind (http://www.openfind.com). 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. In addition, the required parameters for the chi-square test were computed using the search results returned from the utilized search engines.

Performance of the Proposed Approaches

We carried out experiments to determine the performance of the proposed approaches in extracting translations for the random-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 2 shows the obtained results for random queries in terms of top 1-5 inclusion rates and coverage rate. In this table, CV, X2, ATS and Combined represent the context vector analysis, chi-square test, anchor-text-based, and combined approaches, respectively.

Table 2 Coverage and top 1~5 inclusion rates obtained with the four different approaches for the random-query set

Approach	Top-1	Top-3	Top-5	Coverage
CV	40.0%	54.0%	54.0%	68%
X2	36.0%	50.0%	52.0%	68%
ATS	20.0%	32.0%	32.0%	32%
Combined	44.0%	64.0%	66.0%	72%

As shown in Table 2, except for the ATS approach, the approaches were reliable. The top-1 inclusion rate obtained with the ATS approach was only 20%. The main reason, we found, was that many test query terms did not appear in the test anchor-text-set corpus, not to mention their translations. The merit of the search-result-based approaches is, thus, obvious. Their performance degradation was small. The anchor-text-based and search-result-based approaches are quite complementary. This can be seen from the performance obtained using the combined approach, where the achieved top-1 inclusion rate reached 44% and the coverage rate reached 72%.

Discovering useful knowledge in Web data for CLIR has not been fully explored in our study. In fact, the translation process based on the search-result approaches might not be very effective for language pairs that do not exhibit the mixed language characteristic on the Web. The anchor-text approach is, therefore, also attractive for this reason. In our experience, the anchor-text-based approach achieves good precision rates for popular queries and may extract longer translations in other languages besides Chinese and English. However, although Web anchor texts undoubtedly are live multilingual resources, not every particular pair of languages has sufficient texts on the Web. To deal with the above problems, we have extended the Web mining approaches by adding a phase consisting of indirect translation via an intermediate language. For a query term which can not be translated, the extended approach will translate it into a set of translation candidates in an intermediate language and then to seek the most likely translation from among the candidates that are translated from the intermediate language into the target language. We, therefore, have proposed a *transitive translation model* to further exploit anchor text mining for translating Web queries [5].

CONCLUDING REMARKS

Practical cross-language Web search services have not lived up to expectations since they suffer from a major problem in that up-to-date bilingual lexicons containing the translations of popular query terms, such as proper nouns, are lacking. The LiveTrans system effectively utilizes live Web sources for query translation, that are contributed by a huge number of volunteers on a daily basis. It can generate translation equivalents for many queries in a query log in a batch mode and constantly updates the effective translations for each new query term. By combining Web mining approaches, the system can generate effective translation equivalents for many Web queries and provide a practical English-Chinese cross-language search service for the retrieval of both Web pages and images.

REFERENCES

[1] 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.

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

[3] 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.

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

[5] Lu, W. H., Chien, L. F., Lee, H. J. (2002b) A Transitive Model for Extracting Translation Equivalents of Web Queries through Anchor Text Mining, To appear in proceedings of the 19th International Conference on Computational Linguistics (COLING2002), 584-590.

[6] Oard, D. W. (1997) Cross-language Text Retrieval Research in the USA. Proceedings of the 3rd ERCIM DELOS Workshop., Zurich, Switzerland.