Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach

WEN-HSIANG LU
Academia Sinica and National Chiao Tung University
LEE-FENG CHIEN
Academia Sinica
and
HSI-JIAN LEE
National Chiao Tung University

To discover translation knowledge in diverse data resources on the Web, this article proposes an effective approach to finding translation equivalents of query terms and constructing multilingual lexicons through the mining of Web anchor texts and link structures. Although Web anchor texts are wide-scoped hypertext resources, not every particular pair of languages contains sufficient anchor texts for effective extraction of translations for Web queries. For more generalized applications, the approach is designed based on a transitive translation model. The translation equivalents of a query term can be extracted via its translation in an intermediate language. To reduce interference from translation errors, the approach further integrates a competitive linking algorithm into the process of determining the most probable translation. A series of experiments has been conducted, including performance tests on term translation extraction, cross-language information retrieval, and translation suggestions for practical Web search services, respectively. The obtained experimental results have shown that the proposed approach is effective in extracting translations of unknown queries, is easy to combine with the probabilistic retrieval model to improve the cross-language retrieval performance, and is very useful when the considered language pairs lack a sufficient number of anchor texts. Based on the approach, an experimental system called LiveTrans has been developed for English–Chinese cross-language Web search.

Categories and Subject Descriptors: H.2.4 [**Database Management**]: Systems—*textual databases*; H.2.8 [**Database Management**]: Database Applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.7.2 [**Document and Text Processing**]: Document Preparation—*hypertext / hypermedia*

This research was partially supported by a research grant under contract NSC91-2219-E-001-010 from the National Science Council, Republic of China.
Authors' present addresses: W.-H. Lu, National Cheng Kung University, No. 1, TaHsueh Road, Tainan 701, Taiwan, ROC; email: whlu@mail.ncku.edu.tw; L.-F. Chien, Institute of Information Science, Academia Sinica, Nangang 115, Taiwan; email: lfchien@iis.sinica.edu.tw; H.-J. Lee, Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 300, Taiwan; email: hjlee@csie.nctu.edu.tw.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 1515 Broadway, New York, NY 10036 USA, fax: +1 (212) 869-0481, or permissions@acm.org.
© 2004 ACM 1046-8188/04/0400-0242 $5.00

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Multilingual translation, anchor text mining, cross-language information retrieval, cross-language Web search, competitive linking algorithm

1. INTRODUCTION

The Web is becoming the largest data repository in the world. How to discover knowledge in diverse data resources on the Web and benefit Web information systems is being studied in almost every intelligent information processing area. Multilingual terminological resources, such as multilingual lexicons or thesauri, are valuable for conducting academic researches or developing applications, such as machine translation [Brown et al. 1993; Knight 1997], cross-language information retrieval (CLIR) [Dumais et al. 1996; Oard and Diekema 1998], or even information exchange in electronic commerce [Silverman et al. 2001]. However, manual lexicography is time-consuming and not cost-effective. It is worthwhile to automatically construct multilingual lexicons or even thesauri via mining the Web content, which consists of huge amounts of multilingual and wide-scoped hypertext resources.

To deal with the above problem, we have proposed a preliminary approach to extracting translations of Web queries through the mining of Web anchor texts and link structures [Lu et al. 2001, 2002a]. This novel approach exploits Web anchor texts as bilingual corpora to alleviate the existing difficulties of cross-language Web search, and has been proven particularly effective for extracting multilingual translation equivalents of query terms[1] containing proper names or new terminology. For example, Figure 1 shows a typical example, in which there are a variety of anchor texts in multiple languages linking to the URL at http://www.yahoo.com/ from all over the world. Such a bundle of anchor texts pointing together to the same page is called as an *anchor-text set* (refer to Subsection 3.1). It is not difficult to find some of Yahoo's regional aliases in this 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.

Discovering useful knowledge in Web anchor texts has not been fully explored. In accordance with our previous experiments, the extracted translation equivalents might not be reliable when a query term's corresponding translations either appear infrequently in the same anchor text sets or even do not appear together. In addition, the translation process will be inapplicable if there is a lack of sufficient anchor texts for a particular language pair. Although Web anchor texts undoubtedly are multilingual resources, not every particular pair of languages contains sufficient anchor texts. To deal with the above problems, this paper extends the previous anchor-text-based approach by adding a phase consisting of indirect translation via an intermediate language. For a query term which can not be translated, our idea is to translate it into a set of translation candidates in an intermediate language, and then to seek the most likely

[1] In our collected query logs, most of the user queries contain only one or two words, so we use query term, query or term interchangeably in this article.

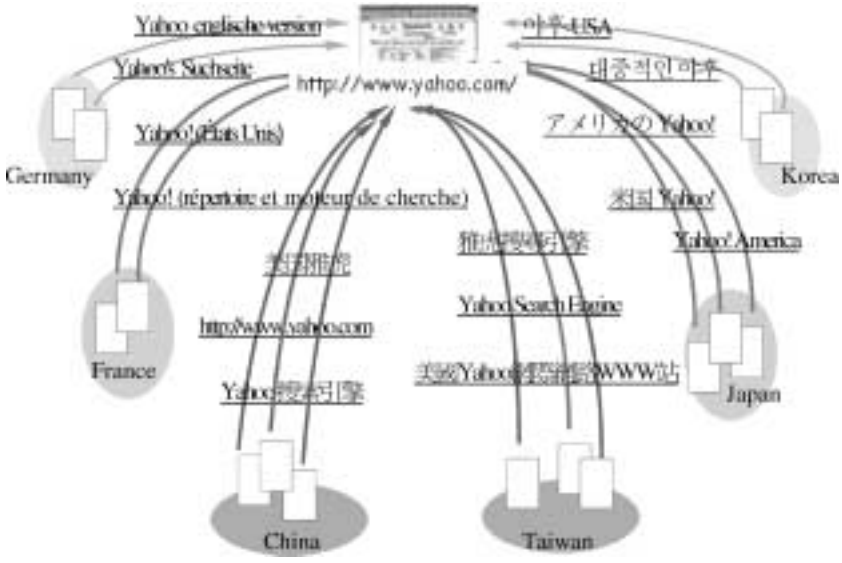

Fig. 1. An illustration showing various anchor texts in multiple languages linking to site Yahoo! from all over the world [Lu et al. 2002a].

translation from the candidates that are translated from the intermediate language into the target language. We therefore propose a *transitive translation approach* to further exploit anchor text mining for translating Web queries [Lu et al. 2002b].

To reduce interference from translation errors, the proposed transitive translation model further integrates a *competitive linking algorithm* into the process of determining the most likely translation [Melamed 2000]. The algorithm is designed based on an approximation of the maximum a posteriori matching for all considered translation pairs. The overall approach is composed of several techniques, including anchor text mining, the probabilistic inference model, the transitive translation model, and the competitive linking algorithm. A series of experiments has been conducted, including performance tests on term translation extraction, cross-language information retrieval, and translation suggestions for practical Web search services, respectively. The obtained experimental results on term translation extraction show that the transitive translation model is very useful when the considered language pairs lack a sufficient number of anchor texts. In addition, the competitive linking algorithm is effective in reducing the number of *indirect association errors*, that is, erroneous translations caused by highly relevant source terms. To determine the performance of the proposed approach when applied to CLIR, we conducted some experiments on the NCTIR-2 English–Chinese IR task using both CL-LSI and the probabilistic retrieval model. The proposed approach was found to be effective in extracting translations of unknown queries, and easy to combine with the probabilistic retrieval model to improve the retrieval performance. On the other hand, although the proposed approach cannot be integrated into the CL-LSI model in a straightforward manner, we found that it could augment

the translation lexicon and aid the retrieval process. Based on the proposed approach, an experimental system, called *LiveTrans*, has been developed for English–Chinese cross-language Web search.

2. RELATED WORK

2.1 Translation Extraction and CLIR

Cross-language information retrieval (CLIR), 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 [Ballesteros and Croft 1998]. However, this presents a number of challenges, especially the problem of query translation. To make query translation possible, existing IR systems mostly rely on bilingual dictionaries for cross-lingual retrieval. In these systems, queries submitted in a source language normally have to be translated into a target language by means of simple dictionary lookup. These techniques are limited in real-world applications since queries given by users often contain proper nouns that do not exist in general-purpose bilingual dictionaries. Another popular approach to dealing with query translation is based on corpus-based techniques. A number of related researches on natural language processing have used a parallel corpus, containing aligned sentences whose translation pairs correspond to each other, to extract translation equivalents [Brown et al. 1993; Dagan et al. 1993; Smadja et al. 1996]. In addition, an alternative approach using comparable or unrelated text corpora was presented by Rapp [1999] and Fung et al. [1998].

On the other hand, many research works have been focused on the development of cross-language retrieval models to improve retrieval performance. Some of them were based on probabilistic retrieval models [Xu et al. 2001; Hiemstra and de Jong 1999; Lavrenko et al. 2002]. For example, Lavrenko et al. [2002] developed a formal probabilistic model for integrating query expansion into the relevance estimation of documents. In addition, some other works were based on cross-language latent semantic indexing (CL-LSI), which uses the Singular Value Decomposition (SVD) to discover important associative relationships among source query terms and target documents without the need for query translation [Dumais et al. 1996, 1997; Furnas et al. 1988].

Practical cross-language Web search services have not lived up to expectations, since they suffer from a major bottleneck due to the fact that up-to-date bilingual lexicons containing the translations of popular query terms such as proper nouns are lacking [Kwok 2001; Hull and Grefenstette 1996]. Web query terms are often diverse and dynamic. Only a certain set of the translations of a query term can be extracted using corpora with limited domains. Different from previous works, Nie et al. [1999] developed an efficient method for extracting parallel text corpora from bilingual Web sites instantly, based on use of the Web's multilingual nature and a wide range of hypertext resources to collect bilingual corpora. Similar to utilizing parallel texts extracted from the Web, Resnik et al. [2001] proposed a backoff translation technique that combines

evidence from dictionary-based and corpus-based approaches, and improves the coverage and accuracy of translation lexicons.

2.2 Transitive Translation via a Third Language

For CLIR, transitive translation via an extra language might be necessary and beneficial for query translation if the direct query translation is unreliable or even unavailable due to the lack of sufficient parallel (comparable) corpora or a bilingual dictionary. Borin [2000] tried to use various sources of information to improve the alignment of word translation and proposed the pivot alignment, which combines direct translation and indirect translation via a third language. The method increases the recall rate without lowering precision.

In general, using an intermediate language for transitive translation doubles the number of translations and thus increases the likelihood of translation errors due to incorrect identification of ambiguous words. Gollins and Sanderson [2001] proposed a feasible approach of translating in parallel across multiple intermediate languages and fusing the results. Such a technique eliminates errors and raises the effectiveness of retrieval. In addition, Simard [2000] developed an approach to exploiting the transitive properties of translations to improve quality of multilingual text alignment.

2.3 Text Mining

Text mining, also known as text data mining [Hearst 1999], concerns the discovery of knowledge in huge amounts of unstructured textual data. It is a relatively new research area in data mining in which most researchers have focused on knowledge discovery in structured databases [Fayyad et al. 1996; Deogun et al. 1997]. Feldman and Dagan [1995] created the Knoweledge Discovery in Texts (KDT) system, the precursor to the text mining, to find patterns between concept distributions in textual data. A variety of related studies have focused on different subjects, such as automatic extraction of terms or phrases [Feldman et al. 1997; Ahonen et al. 1999], the discovery of rules for the extraction of specific information patterns [Soderland 1997], and ontology construction [Fensel 2001].

On the other hand, some recent works have proposed practical algorithms, such as HITS [Kleinberg 1998] and PageRank [Brin and Page 1998], for modeling Web topology or exploring authoritative or popular pages in Web search based on link structures and their related content information [Chakrabarti et al. 1998]. Amitay and Paris [2001] also applied both hyperlink structures and the contexts of their associated anchor texts to automatically summarize Web pages and sites. Unlike the previous researches, we exploit Web anchor texts and link structures as multilingual corpora to incrementally discover knowledge for extracting multilingual translations of query terms.

3. TRANSLATION EXTRACTION

The proposed transitive translation approach is designed mainly to find exact translation equivalents of query terms via anchor text mining. The overall approach incorporates several techniques, including anchor text mining, the

probabilistic inference model, the transitive translation model, and the competitive linking algorithm. In this section, the basic concepts behind the probabilistic inference model and the transitive translation model will be introduced.

3.1 Anchor-Text Set

An anchor text is the descriptive part of an out-link of a Web page to represent a brief description of the linked Web page. The anchor-text set of a Web page (or URL) u_i is defined as all of the anchor text of the links, that is, u_i's in-links, pointing to u_i. In general, the anchor-text set records u_i's alternative concepts and textual expressions, such as titles and headings, which are cited by other Web pages. With different preferences, conventions and levels of language competence, the anchor-text set can be composed of multilingual phrases, short texts, acronyms, or even u_i's URL. For a query term appearing in an anchor-text set,it is likely that its corresponding translations will also appear together. A collection of anchor-text sets can be considered as composing a comparable corpus of translated texts, from the viewpoint of translation extraction.

Conventional statistical translation models perform estimation mostly based on a parallel corpus. These models have a fundamental problem in that they are not able to efficiently deal with the translation of semantically close or opposite terms in CLIR applications. That is, because semantically close or opposite terms have a lower chance of occurring together with the source term in the same sentences; as a result, their translations cannot be extracted through co-occurrence with the source term in a parallel corpus. A collection of anchor-text sets can be considered as being composed of a comparable corpus of translated texts. In accordance with our observations, there is a higher chance of finding semantically close translations for a source word from anchor-text sets containing it. The above drawback can be to some degree improved by using anchor-text sets as the corpus if a proper translation model can be employed.

3.2 Probabilistic Inference

Different from conventional statistical translation models or rule induction used in data mining, our proposed probabilistic model uses the symmetric similarity estimation function $P(s \leftrightarrow t)$ to replace the conditional estimation function $P(s \rightarrow t)$ and takes into account the significance of the authority of Web pages linked by other related pages.

To determine the most probable target translation t for source query term s, we have developed a probabilistic inference model [Wong et al. 1995]. 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 Webpages [Kleinberg 1998; Chakrabarti et al. 1998]. Our model further assumes that

the translation candidates in the anchor-text sets of the pages with higher authority may be more reliable. The similarity estimation function based on the probabilistic inference model is defined below:

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

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. This estimation is simplified from the HITS algorithm [Kleinberg 1998].

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

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

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 the pages having higher authority.

In our previous work, we showed that the link structure and page authority information are useful for improving the translation accuracy [Lu et al. 2002a]. We found that text expressions associated with the links of pages with higher authority are usually more reliable and professional. This phenomenon is especially obvious in the case of the Chinese Web, where bilingual pages are often provided by organizations with higher authority. In fact, using the probabilistic inference model, we found it easier to consider the characteristics of anchor texts and to integrate the model with other techniques to improve the translation accuracy. Also, as will be shown in Section 6.2.1, the model could be integrated with the probabilistic retrieval model to improve the performance of cross-lingual retrieval, especially for queries with unknown query terms. For further information about the probabilistic inference model, readers may refer to our previous works [Lu et al. 2001, 2002a].

3.3 Direct Translation

For each source term, the probabilistic inference model extracts the most probable translation that maximizes the estimation. The estimation process based on the model was developed to extract term translations through the mining

of real-world anchor-text sets. The process contains three major computational modules: anchor-text extraction, term extraction and term translation extraction. The anchor-text extraction module was constructed to collect pages from the Web and build up a corpus of anchor-text sets. The coverage and corresponding domains of the collected Web pages U determine the reliability of the estimation process. In fact, the proposed approach can be extended to the extraction of domain-specific term translations if the module focuses on collecting domain-specific documents on the Web.

On the other hand, for each given source term s, the term extraction module extracts key terms as the translation candidate set T from the anchor-text sets of the pages containing s, that is, AT_s. The effectiveness of the adopted term extraction methods greatly affects performance in extracting correct translations. Three different methods have been tested: the PAT-tree-based, query-set-based and tagger-based methods [Lu et al. 2002a]. Among them, the query-set-based method has been strongly recommended because it can avoid some problems with term segmentation. The method uses a query log in the target language as the translation vocabulary set V_t to segment key terms in AT_s. The precondition for using this method is that the coverage of the query set should be high. Last, the term translation extraction module extracts the translation that maximizes the similarity estimation. To reduce the computational cost, the similarity estimation between source term s and each translation candidate t is estimated based only on the page set $U_{\{s\}\cup\{t\}}$ that contains either s or t in AT_s or AT_t. To distinguish it from the translation process that uses an intermediate language, the above process is called direct translation, and the adopted model will be called the direct translation model hereafter. Meanwhile, we use function P_{direct} in Eq. (3) for the estimation of the direct translation:

$$P_{direct}(s, t) = P(s \leftrightarrow t) \tag{3}$$

The direct translation process is designed to extract the top k most probable translations as the output. Its major operations are summarized in Algorithm 1 (see Figure 2).

3.4 Transitive Translation

As mentioned above, for those query terms whose corresponding translations either appear infrequently in the same anchor text sets or do not appear together, the estimation obtained using Eq. (2) is basically unreliable. To increase the possibility of translation extraction, especially for source terms whose corresponding translations do not co-occur, we add an indirect translation phase using an intermediate language. For example, as shown in Figure 3, our idea is to obtain the corresponding target translation "索尼" in simplified Chinese by translating the source term "新力" in traditional Chinese into an intermediate term "Sony" in English, and then trying to translate "Sony" into a target term "索尼" in simplified Chinese. For both the source query and the target translation, we assume that their translations in the intermediate language are the same and can be found.

The above assumption is actually not unrealistic. For example, it is possible to find the Chinese translation of the name of a Japanese movie star

Algorithm 1: ***Direct_Translation*** (s, U, V_t)

Input: source term s, Web pages of concern U, translation vocabulary set V_t.

Output: target translation set R.

1. Extract translation candidate set T.
   - 1.1. Let AT_s be the anchor-text sets of the pages in U containing s. Use V_t as the translation vocabulary set to segment terms appearing in AT_s and take them as translation candidates.
2. For each translation candidate t in T, perform the following sub-steps:
   - 2.1. Extract the page set $U_{\{s\}\cup\{t\}}$ that contains either s or t in its anchor text sets (pages containing neither s nor t in their anchor-text sets have no effect for the estimation in Step 2.2).
   - 2.2. Calculate $P_{direct}(s, t)$ based on Equation (2) and page set $U_{\{s\}\cup\{t\}}$.
3. Select up to k most probable translations from T as the target translation set R and return R.

Fig. 2. Algorithm for the estimation of direct translation1

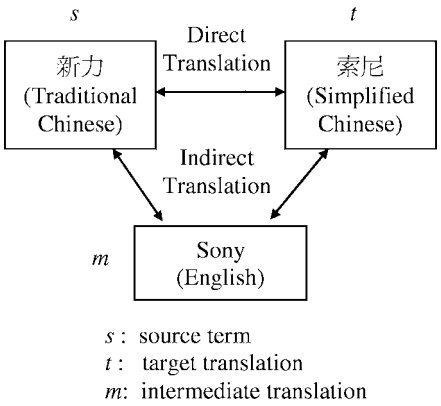

Fig. 3. An abstract diagram showing the concepts of direct translation and indirect translation.

by submitting his/her English name to a search engine and browsing the retrieved Chinese pages containing the English name. Based on this assumption, we extend the probabilistic inference model and propose an indirect translation model as shown in the following formula:

$$P_{indirect}(s, t) = \sum_m P(s \leftrightarrow m, m \leftrightarrow t)P(m)$$
$$\approx \sum_m P(s \leftrightarrow m) \times P(m \leftrightarrow t) \times P(m), \quad (4)$$

where m is one of the top k most probable intermediate translations of s in the intermediate language, and $P(m)$ is the confidence value of m's accuracy, which can be estimated based on m's probability of occurring in the corpus.

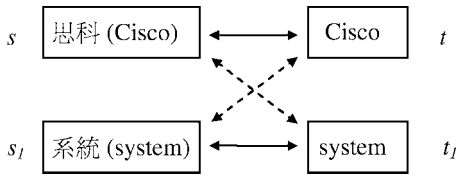

Fig. 4. An illustration showing the indirect association problem, in which the dashed arrows indicate possible indirect association errors.

In addition, $P(s \leftrightarrow m)$ and $P(m \leftrightarrow t)$ are the probability values, which can be obtained by using the direct translation model and calculated using Eq. (2).

The transitive translation model combines both the direct and indirect translation models. By combining Eqs. (3) and (4), the transitive translation model can be defined as follows:

$$P_{trans}(s, t) = \begin{cases} P_{direct}(s, t), & \text{if} \quad P_{direct}(s, t) > \theta \\ P_{indirect}(s, t), & \text{otherwise,} \end{cases} \tag{5}$$

where θ is a predefined threshold value. The transitive translation model can be used to enhance the ability of the proposed approach to deal with the translation of a language pair when an adequate corpus is lacking. Under this circumstance, the direct translation process is normally not reliable enough. Via translation into an intermediate language, the possibility of translation extraction can be increased. However, the indirect translation process might unavoidably result in more errors in some cases. For a language pair with an adequate corpus, the importance of indirect translation in the estimation process can be reduced.

4. ROBUST ESTIMATION USING THE COMPETITIVE LINKING ALGORITHM

4.1 Indirect Association Problem

As mentioned above, the estimation obtained using the proposed models is not reliable if the translation of the source term appears infrequently in the same anchor text sets. A so-called *indirect association problem* might arise [Melamed 2000]. An illustration in Figure 4 shows this problem. Assume that t is s's corresponding translation but appears infrequently with s. An indirect association error might arise when t_1, the translation of s's highly relevant term s_1, co-occurs often with s. To reduce such translation errors and enhance the reliability of the estimation, a *competitive linking* (CL) algorithm that is extended from Melamed's work is developed and integrated into the process of determining the most probable translation [Melamed 2000].

4.2 Concept of CL Algorithm

Let A be a one-to-one mapping between a set of source terms S and a set of possible target translations T, and let the mapping for a source term to be associated with a translation be called an assignment. The competitive linking algorithm determines the most probable translation pairs for the two term sets. The algorithm can be viewed as a heuristic search for the most likely

assignment in the space of all possible assignments. For the convenience of discussion, we use a bipartite graph to show the concept. The bipartite graph can be defined as $G = (S \cup T, E)$, in which all the considered source terms and target translations are vertices (the vertices on one side correspond to unique source terms, and the vertices on the other side to unique target terms), and the edges are weighted with the similarity values of the corresponding source terms and target translations. Let $e_{ij} = \langle s_i, t_j, w_{ij} \rangle$ indicate the edge from the vertex of source term s_i to the vertex of target term t_j and its weight $w_{ij} \cdot w_{ij}$ can be $P_{direct}(s_i, t_j)$ or $P_{indirect}(s_i, t_j)$, depending on the model employed.

Let **A** be the set of all possible assignments between S and T. The CL algorithm is a heuristic search that proceeds using an iterative process of assignment elimination. In the first search iteration, all the assignments that do not contain the most likely edges (links) are discarded. In the second iteration, all the assignments that do not contain the second most likely link are discarded, and so on until only one assignment remains. The algorithm greedily selects the most likely edges first, and then selects less likely edges only if they do not conflict with previous selections. The algorithm is based on the one-to-one assumption: each source term is translated into at most one term.

For the example shown in Figure 4, s, s_1 can be taken as source vertices, and t, t_1 as target vertices. Based on the one-to-one assumption, the example conceptually constitutes two possible assignments (graphs), that is, $G_1 = (\{s, s_1\} \cup \{t, t_1\}, \{\langle s, t \rangle, \langle s_1, t_1 \rangle\})$, and $G_2 = (\{s, s_1\} \cup \{t, t_1\}, \{\langle s, t_1 \rangle, \langle s_1, t \rangle\})$. Assume that $\langle s_1, t_1 \rangle$ is the most likely edge, that is, $w_{s_1 t_1}$ is the highest weight. In the first search iteration, assignment G_2 is discarded immediately. Therefore, assignment G_1 is the remaining one, and s's translation is, thus, determined as t. For the sake of efficiency in implementation, all of the possible graphs can be merged into one, that is, $G_3 = (\{s, s_1\} \cup \{t, t_1\}, \{\langle s, t \rangle, \langle s_1, t_1 \rangle, \langle s, t_1 \rangle, \langle s_1, t \rangle\})$. The above search process is then modified and re-interpreted as follows. In the first search iteration, the edge $\langle s_1, t_1 \rangle$ whose weight is the highest will be selected, and all other edges linking to either s_1 or t_1 (i.e., $\langle s, t_1 \rangle$ and $\langle s_1, t \rangle$) will be discarded. Therefore, the weights of the remaining edges will be re-estimated and the second iteration performed. However, in our case, edge $\langle s, t \rangle$ is the only remaining edge after the first iteration, and no further iterations are necessary.

4.3 Problems of Using CL Algorithm

Before the above algorithm can be used in real applications, there are some problems that need to be further investigated. The CL algorithm is designed based on an approximation of the maximum a posteriori matching for the term pairs in S and T. According to Melamed [2000], the employed one–one mapping assumption is not as restrictive as it may appear. As the experimental results presented in Section 5.4 will demonstrate, the CL algorithm is effective in reducing the number of indirect association errors and useful in dealing with the translation of language pairs when an adequate corpus is not available.

On the other hand, the computing time is another possible drawback. Let a bipartite graph $G = (S \cup T, E)$ with $v = |S \cup T|$ and $e = |E|$; the lowest currently known upper bound on the computational complexity of this problem

Algorithm 2: ***Bipartite_Graph_Construction*** (s, U, V_t)

Input: source query term s, Web pages of concern U, translation vocabulary set V_t.

Output: bipartite graph G for s.

1. Based on translation vocabulary set V_t, extract a set of s's possible translations $T = \{t_1, t_2, \ldots, t_n\}$ from AT_s, the anchor-text sets of the pages in U containing s.
2. For each t_i in T, find a set of possible translations St_i which appear in both ATt_i and AT_s. Let $ST = St_1 \cup St_2 \cup \ldots \cup St_n$, and $S = \{s\} \cup ST$.
3. Construct a bipartite graph $G = (V_1 \cup V_2, E)$, where $V_1 = S$ and $V_2 = T$, and E contains the edges from every vertex in V_1 to vertex in V_2.

Fig. 5. Algorithm for bipartite graph construction.

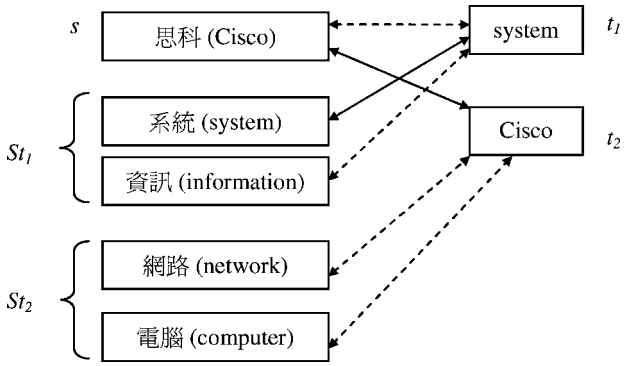

Fig. 6. An illustration showing a bipartite graph generated by using Algorithm 2.

is $O(ve + v^2 \log v)$. The Web is live, and users' query terms are dynamic and huge in number. Although this upper bound is polynomial, it is still too high for typical Web query sets. To reduce the computation cost, the number of vertices (the number of considered source terms and translation candidates) should be limited.

To extract the most likely translation of each given source term, it is not realistic to consider matching with all of the source terms and their possible translations in the corpus. We, therefore, develop a method to determine a proper set of candidate terms for the estimation. Algorithm 2 shown in Figure 5 is a method designed for both direct and indirect translations, in which s's translation set $T = \{t_1, \ldots, t_n\}$ contains only the candidates extracted from AT_s, that is, the anchor-text sets of the pages in U containing s. In addition to source term s, the source term set contains additional source terms $ST = St_1 \cup \cdots \cup St_n$, which is the union of all translation sets of s's translation candidates that appear in T_s. These source terms are assumed to be likely to cause indirect associations and should be included when constructing a bipartite graph for further estimation. Following the previous example, Figure 6 shows an example graph that might be generated using Algorithm 2.

Algorithm 3: ***Direct_Translation_with_CL*** (s, U, V_t)

Input: source term s, Web pages of concern U, translation vocabulary set V_t.

Output: target translation set R.

1. Call *Bipartite_Graph_Construction* (s, U, V_t) to construct bipartite graph $G = (S \cup T, E)$ for source term s and its possible translations in V_t.
2. For each edge e_{ij} in E, compute the translation probability $P_{direct}(s_i, t_j)$ as its initial weight w_{ij}. The probability estimation is based on the anchor-text sets of the pages in U containing at least one term in $S \cup T$, rather than all of the collected pages.
3. Sort all w_{ij} from highest to lowest.
4. Let $e_i{}^*{}_j{}^*$ be the edge with the highest weight. If $s_i{}^* = s$, let $R = R \cup \{t_j{}^*\}$. Return R if $|R| = k$.
5. If $s_i{}^* \neq s$, remove all edges linking to either vertex $s_i{}^*$ or vertex $t_j{}^*$. Re-estimate w_{ij} for each remaining edge e_{ij}. The probability estimation is based on the original anchor-text sets but excluding sets that contain either $s_i{}^*$ or $t_j{}^*$.
6. If $s_i{}^* = s$, remove all edges linking to vertex $t_j{}^*$. Re-estimate w_{ij} for each remaining edge e_{ij}. The probability estimation is based on the original anchor-text sets but excluding sets that contain $t_j{}^*$.
7. Repeat beginning with Step 3 until $|E| = 0$.
8. Return R as the output.

Fig. 7. Algorithm for the estimation of direct translation using the CL algorithm.

4.4 Integration with the CL Algorithm

The CL algorithm can be implemented based on the bipartite graph constructed by Algorithm 2. Algorithm 3 shown in Figure 7, which is an extension of Algorithm 1, describes the operations for estimating direct translation using the CL algorithm. Note that the algorithm extracts up to the k most probable translations as the output.

In addition, Algorithm 4 shown in Figure 8 describes the operations for estimating indirect translation using the CL algorithm, which is an approximation considering the high computational cost that the CL algorithm might cause. The algorithm divides the estimation process into two stages. In the first stage, it extracts up to the k most probable intermediate translations. It constructs a bipartite graph containing the source term, the possible intermediate translations and their source translations. With these intermediate translations, it extracts the most probable target translations in the second stage. For each intermediate translation, an additional bipartite graph needs to be constructed. The graph consists of the intermediate translation, its possible target translations and their translations in the intermediate language. It calls Algorithm 3 to extract

Algorithm 4: ***Indirect_Translation_with_CL*** (s, U, V_t, V_m)

Input: source term s, Web pages of concern U, target translation vocabulary set V_t, and transitive translation vocabulary set V_m.

Output: target translation set R.

1. Call *Direct_Translation_with_CL* (s, U, V_m) to extract the most probable intermediate translation set $M = \{m^*\}$ for s.
2. For each m^*, Call *Direct_Translation_with_CL* (m^*, U, V_t) to extract the most probable target translation set $TR = \{t^{**}\}$.
3. For each t^{**}, calculate $P_{indirect}$ (s, t^{**}) based on Equation (4) and page set $U_{\{s\}\cup\{t^{**}\}\cup\{m^*\}}$.
4. Select up to the k most probable target translations from all t^{**} as the output set R and return R.

Fig. 8. Algorithm for the estimation of indirect translation using the CL algorithm.

Algorithm 5: ***Transitive_Translation_with_CL*** (s, U, V_t, V_m)

Input: source term s, Web pages of concern U, target translation vocabulary set V_t, and intermediate translation vocabulary set V_m.

Output: target translation set R.

1. Call *Direct_Translation_with_CL* (s, U, V_t) to obtain the most probable target translation set R_1.
2. Select up to k most probable translations $\{t_1^*\}$ from R_1, where P_{direct} (s, t^*) is larger than the threshold. Let R = $\{t_1^*\}$, and return R if $|R| = k$.
3. Call *Indirect_Translation_with_CL* (s, U, V_t, V_m) to obtain up to the k - $|R|$ most probable target translations $\{t_2^*\}$. Let $R = R \cup \{t_2^*\}$ and return R.

Fig. 9. Algorithm for the estimation of transitive translation using the CL algorithm.

a set of the most probable target translation candidates. For each target translation candidate generated by the intermediate translations, it uses the function $P_{indirect}$ in Eq. (4) to estimate the similarity value between the source term and the target translation candidate. The top k most probable target translations will be selected as the output. A drawback of the above estimation process is that the most probable target translations are determined without considering all the possible intermediate translations of the source term. This is, however, a tradeoff.

With the above three algorithms, estimation with the transitive translation model using the CL algorithm is straightforward. Algorithm 5 (see Figure 9) shows the process.

Table I. Some Examples of Test Queries

Type	Number	Examples of test queries
Dic	96	銀行 (bank) 亞洲 (Asia) 愛滋病 (AIDS) 書店 (bookstore) 照相機 (camera)
OOV	162	電子商務 (e-commerce) 地理資訊系統 (GIS) 雅虎 (Yahoo) 美國職籃 (NBA) 喬丹 (Jordan) 威而剛 (viagra)

5. EXPERIMENTAL RESULTS

5.1 Analysis of Anchor-Text Sets and Query Logs

In the initial experiments, we took traditional Chinese and simplified Chinese as the source and target language, respectively, and used English as the intermediate language. We had collected 1,980,816 traditional Chinese Web pages in Taiwan. Among these pages, 109,416 pages whose anchor-text sets contained both traditional Chinese and English terms were taken as the anchor-text set corpus. We had also collected 2,179,171 simplified Chinese Web pages in China and extracted 157,786 pages whose anchor-text sets contained both simplified Chinese and English terms. In addition, after merging the two Web page collections into a larger one, we extracted only 4,516 pages whose anchor-text sets contained both traditional and simplified Chinese terms. The three comparable corpora provided a potential resource of translation pairs for some Web queries. In order to realize the feasibility of translating query terms via transitive translation, we aimed to find out the corresponding simplified Chinese translations of traditional Chinese query terms using English as the intermediate language.

We also collected popular query terms with the logs from two real-world Chinese search engines in Taiwan, that is, Dreamer and GAIS.[2] 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. There were 9,709 most popular query terms whose frequencies were above 10 in both logs, and 1,230 of them were English terms. After filtering out the terms used locally in Taiwan, we obtained 513 terms, and 430 of which had Chinese translations that appeared together in the logs. For retrieving the desired pages, these terms are supposed to be useful in cross-language search. We randomly selected 258 terms from them and took their corresponding Chinese translations from the logs, which were determined manually, as *the source query set* in the following experiments. Table I lists some

[2]Both search engines are second-tier portals in Taiwan, whose logs contain queries that are popular in Chinese communities, and whose URL's are as follows: http://www.dreamer.com.tw/ and http://gais.cs.ccu.edu.tw/.

Table II. Top-*n* Inclusion Rates Obtained with the Direct Translation Model and the Three Specific Language Pairs Corpora

Type	Top1	Top2	Top3	Top4	Top5
TC => SC	35.7%	43.0%	46.9%	49.6%	51.2%
TC => ENG	68.6%	82.2%	85.7%	88.0%	88.8%
ENG => SC	45.3%	55.8%	59.3%	61.6%	64.0%

examples of the test query terms, which were divided into two types, where type Dic (the terms existing in dictionary) made up about 37% (96/258) of the test queries, and type OOV (out of vocabulary; the terms not in the dictionary) made up about 63% (162/258). The dictionary used for the experiments was complied with the CEDICT[3] Chinese–English electronic dictionary, which contains 23,948 word/phrase entries with the translations in traditional Chinese, simplified Chinese and English.

To evaluate the performance of translation extraction, we used the *average top-n inclusion rate* as a metric. For a set of test query terms, its top-*n* inclusion rate was defined as the percentage of query terms whose effective translation(s) could be found in the top-*n* extracted translations. Normally, the determination of whether or not extracted translations are effective is based on whether the translations can be used to perform similar search requests in the target language, which needs to be judged manually. There is a major difference between query-term translation in Web search and common-term translation in machine translation. Query terms usually indicate certain information requests. Their translated terms should be effective for the same requests. In the experiments, three volunteers with library information service backgrounds helped prepare effective translations.

In our previous work [Lu et. al. 2002a], we performed a number of experiments to determine the performance of the approach based on the direct translation model. We showed that the link structure and page authority information can be used to improve translation accuracy. Also, three different term extraction methods were compared, and it was found that the query-set-based method could achieve better performance than either the PAT-tree or POS-tagger-based methods. Nevertheless, since we had not yet collected any query logs in simplified Chinese, in the following experiments, we adopted the PAT-tree-based keyword extraction method, which is an efficient statistics-based approach that can automatically extract repeated text patterns from documents indexed with a PAT-tree data structure [Chien 1997].

5.2 Performance of the Direct Translation Model

In order to determine the feasibility of the transitive translation model, we carried out experiments based on the direct translation models and three different anchor-text set corpora in the first step. Table II shows the results of the obtained top-5 inclusion rates, where the terms “TC,” “SC” and “ENG” represent traditional Chinese, simplified Chinese and English terms, respectively. The

[3]http://www.mandarintools.com/cedict.html.

Table III. Top-*n* Inclusion Rates Obtained with Different Models for Translating Traditional Chinese into Simplified Chinese

Model	Top1	Top2	Top3	Top4	Top5
Direct	35.7%	43.0%	46.9%	49.6%	51.2%
Indirect ($k = 1$)	44.2%	55.1%	58.0%	59.7%	60.5%
Indirect ($k = 3$)	46.5%	57.0%	60.4%	62.0%	62.8%
Transitive ($k = 1$)	49.2%	58.1%	60.9%	61.6%	62.0%
Transitive ($k = 3$)	50.0%	60.1%	62.8%	63.9%	64.3%

performance of translating TC into SC was worse than that of using the other two corpora since the size of the anchor-text set corpus containing both TC and SC was relatively small in comparison with the others. This is why we combined direct translation with indirect translation via a third language. However, the performance of direct translation from TC to SC was used as a reference for comparison with our proposed models in the following experiments.

5.3 Performance of the Indirect and Transitive Translation Models

To determine the improvement obtained using the transitive translation model, we conducted further experiments. As shown in Table III, the indirect and transitive translation models outperformed the direct translation model. As mentioned above, the size of the anchor-text corpus that contained both TC and SC was small. The indirect translation model was, therefore, helpful for finding corresponding translations for some terms with low-frequency values in the corpora. For example, the traditional Chinese term "西門子," we found its corresponding translation equivalent "西门子" in simplified Chinese via the intermediate translation "Siemens," which could not be found using only direct translation.

The use of multiple intermediate translations was found to improve the translation accuracy of both the indirect and transitive translation models. The top-1 inclusion rates ranged from 44.2% when the indirect translation model was used with only one intermediate translation ($k = 1$) to 50.0% when the transitive translation model was used with at most three intermediate translation candidates ($k = 3$). Table V shows some examples of the translations extracted using the transitive translation model.

5.4 Performance of the Models Using the CL Algorithm

The competitive linking algorithm was integrated into the transitive translation model to deal with the indirect association problem to increase the accuracy rate. An additional experiment was also conducted to compare the models without use of the CL algorithm.

Table IV shows that the top-1 inclusion rate achieved using the CL algorithm was slightly better at 52.7%. Detailed analysis of the procedure for processing test queries shows that the CL algorithm is actually helpful in removing some indirect association errors based on correct translation pairs with reliably high co-occurrence frequency. For example, as shown in Table V, the simplified Chinese "蓝鸟" could reach the first rank from the third rank of the translation

Table IV. Top-*n* Inclusion Rates Obtained with the Transitive Translation Model and the Approach Integrating the CL Algorithm

Model	Top1	Top2	Top3	Top4	Top5
Direct + CL	38.0%	43.8%	47.3%	49.6%	51.2%
Indirect + CL ($k = 1$)	48.0%	57.0%	59.4%	60.1%	60.9%
Indirect + CL ($k = 3$)	48.7%	58.1%	60.8%	62.0%	63.1%
Transitive + CL ($k = 1$)	52.7%	60.1%	62.5%	63.1%	63.9%
Transitive + CL ($k = 3$)	52.7%	61.6%	63.9%	64.3%	65.1%

Table V. Two Examples of Extracted Target Translations by the Three Different Models. (The asterisk indicates correct translations)

Source terms (Traditional Chinese)	Top-5 extracted target translations (Simplified Chinese)		
	Direct	Transitive	Transitive with CL
藍鳥 (Bluebird)	Not available	视点 (focus) 电影 (movie) 蓝鸟 (Bluebird)* 试点 (test point) 快车 (express)	蓝鸟 (Bluebird)* 视点 (focus) 电影 (movie) 试点 (test point) 快车 (express)
迪士尼 (Disney)	乐园 (amusement park) 迪士尼 (Disney)* 狮子王 (Lion King) 狄斯尼 (Disney)* 世界 (world)	乐园 (amusement park) 迪士尼 (Disney)* 狮子王 (Lion King) 狄斯尼 (Disney)* 世界 (world)	迪士尼 (Disney)* 乐园 (amusement park) 狄斯尼 (Disney)* 世界 (world) 动画 (anime)

candidates after the CL algorithm was used since the correct high-frequency pairs (focus, 视点) and (movie, 电影) could be chosen first, then the error pairs of indirect association (Bluebird, 视点) and (Bluebird, 电影) could be removed, and finally the correct translation pair (Bluebird, 蓝鸟) could be found as the best pair. As a result, for the traditional Chinese term "藍鳥," its corresponding translation "蓝鸟" in simplified Chinese could be obtained using the intermediate translation "Bluebird."

5.5 Discussion

To realize the abilities of transitive translation model, we further took traditional Chinese and Japanese as the source and target language, respectively, and used English as the intermediate language. We had also collected 801,091 Japanese Web pages and extracted 183,582 pages whose anchor-text sets contained both Japanese and English terms were taken as the Japanese–English anchor-text set corpus. In addition, after merging both collections of 801,091 Japanese pages and 1,980,816 traditional Chinese pages (see Subsection 6.1), we extracted only 1,371 pages whose anchor-text sets contained both traditional Chinese and Japanese terms as the Chinese–Japanese anchor-text set corpus. We continued to use the same 258 traditional Chinese query terms as the test set and the 109,416 Chinese–English anchor-text set corpus to make the following experiments as shown in Table VI. Considering the shortage of a sufficiently large Chinese–Japanese anchor-text corpus on the Web, the performance using

Table VI. Top-*n* Inclusion Rates Obtained with Different Models for Translating Traditional Chinese Queries into Japanese

Model	Top1	Top2	Top3	Top4	Top5
Direct	10.5%	12.8%	14.3%	15.1%	15.1%
Indirect	40.2%	49.4%	56.6%	58.6%	59.6%
Transitive	42.9%	51.4%	58.6%	61.3%	61.9%

Table VII. Selected Source Query Terms in Traditional Chinese and Their Translations in English, Simplified Chinese and Japanese Respectively, Which were Extracted by Our Transitive Translation Model

Source terms (Traditional Chinese)	Extracted target translations		
	English	Simplified Chinese	Japanese
新力	Sony	索尼	ソニー
耐吉	Nike	耐克	ナイキ
史丹佛	Stanford	斯坦福	スタンフォード
雪梨	Sydney	悉尼	シドニー
網際網路	internet	互联网	インターネット
網路	network	网络	ネットワーク
首頁	homepage	主页	ホームページ
電腦	computer	计算机	コンピューター
資料庫	database	数据库	データベース
資訊	information	信息	インフォメーション

both the indirect and transitive models actually surpassed our expectations. Some of multilingual translations of the test queries, including Japanese, extracted by using our transitive translation model are shown in Table VII.

It is important to understand what kinds of queries are suitable when the proposed approach is used to find translations. In fact, the correct translations of a source term are hard to be found when their frequencies are low in the anchor-text sets. This implies that the correct translations of popular terms might more easily appear in the anchor-text sets and be extracted. However, many proper names, such as company names and personal names, have bilingual translations on the Web. If there are no other noisy translations with higher frequencies in the anchor-text sets containing the terms, then their correct translations can be also extracted even though these are not popular terms employed by typical users. In our additional test of the top 19,124 terms, which occupied 80% of the query requests in the Dreamer log, it was found about one-third of the test queries could obtain effective target translations in the top-10 candidates.

On the other hand, it will also be important to determine the performance of the proposed approach when the corpus size changes. A large anchor-text corpus can improve the coverage rates of translation candidates, but it might result in more noises as well. This problem will be further investigated in our future research.

Table VIII. Examples of Title Queries in NTCIR-2

	English title query	Chinese title query
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
Q29	Black-faced Spoonbill in Taiwan	台灣黑面琵鷺
Q30	El Nino and infectious diseases	聖嬰現象與傳染病
Q34	Side effects of Viagra	威而鋼之副作用
Q38	Chunghwa NO. 1 Satellite (ROCSAT-1)	中華衛星一號
Q43	CIH computer virus	CIH 電腦病毒
Q45	Cloud Gate Dance Theatre of Taiwan	雲門舞集
Q46	Ma Yo-yo cello recital	馬友友演奏會
Q47	Jin Yong kung-fu novels	金庸武俠小說

6. CLIR APPLICATION

In this section, we will investigate the performance achieved when the proposed transitive translation approach is applied to CLIR. We conducted experiments with the NTCIR-2 English–Chinese task of retrieving Chinese documents using English queries. The test collection contained 132,173 traditional Chinese news documents (200MB) and 50 English query topics [Chen and Chen 2001]. There were four types of tests designed based on different combinations of descriptive sections of query topics, including "Long Query" (all the sections in a topic were used), "Short Query" (titles, queries and concepts), "Very Short Query" (titles and concepts) and "Title Query" (title sections only). Our experiments focused on testing the title queries because the average length of the title queries was 3.8 English words (after removing stop words), which is close to the length of real Web queries. Previous analysis revealed that the average length of Web queries was about 2.3 words [Silverstein 1998]. Table VIII lists several examples of test title queries.

In previous research [Kwok 2001], the obtained retrieval results for "title" only queries were not good enough, reaching only about 55% of the monolingual MAP (mean average precision) value. According to Kwok, the results demonstrated that users will be disappointed if they use short queries for CLIR. We were interested in finding out whether the proposed approach could help with the retrieval of such short queries, including the extraction of translations of unknown queries, and improve retrieval performance.

6.1 Translation Extraction

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[4] English-to-Chinese lexicon with about 120K entries. The 178 English query terms were tested by using the proposed approach to extract their traditional Chinese translations. The approach was based on Algorithm 5 ($k = 1$). We used simplified Chinese as the intermediate language, and the three anchor text corpora mentioned in Section 5.1 as

[4]The LDC English-to-Chinese lexicon (http://www.ldc.upenn.edu/Projects/Chinese) was adopted since it was also employed in related works [Kwok 2001].

Table IX. Top-*n* Inclusion Rates Obtained with the Proposed Approach for Extracting the Traditional Chinese Translations of the 178 Unique English Title Query Terms

Type	Number	Top1	Top2	Top3	Top4	Top5
Terms existing in LDC	156	56.4%	62.8%	66.7%	67.9%	68.6%
Terms not included in LDC	22	63.6%	68.1%	72.7%	77.3%	77.3%
Total	178	57.3%	63.5%	67.4%	69.1%	69.7%

the test corpora. Table IX shows the achieved translation accuracy. The proposed approach was found to be effective in extracting translations of the test queries, especially the unknown query terms. The obtained inclusion rates for the unknown query terms were even higher than those for the existing query terms. However, the results also revealed that the proposed approach might not be reliable enough when used to extract translations of common terms.

6.2 Retrieval Performance

Both the probabilistic retrieval model [Lavrenko et al. 2002; Xu et al. 2001] and CL-LSI model [Dumais et al. 1996, 1997; Furnas et al. 1988] are major approaches to CLIR. We tried to investigate whether the proposed transitive translation approach could be effectively integrated with these models and applied to CLIR.

6.2.1 *The Probabilistic Retrieval Model.* We first combined the proposed approach with the probabilistic retrieval model. Referring to Xu et al.'s work [Xu et al. 2001; Hiemstra and de Jong 1999], the following model was adopted in our experiment:

$$P(Q|D) = \prod_{e \in Q} P(e|D) = \prod_{e \in Q} \Big[\lambda P(e) + (1-\lambda) \sum_{c} P(e|c) p(c|D)\Big], \qquad (6)$$

where Q is a query, D is a document, e is a composed English query term of Q, c is a target translation of e in traditional Chinese and λ represents a smoothing parameter. In addition, $P(e)$ is the background probability of e, which can be estimated based on e's occurrence in the anchor-text set corpus. $P(c|D)$ is the probability of c appearing in document D. $P(e|c)$ is the translation probability of e given c, which was estimated using three different estimation methods listed below:

—The transitive translation model: we used Equation (5) to estimate the translation probability, that is, $P(e|c) = P_{trans}(e, c)$.

—The dictionary-based method: we used the LDC lexicon to look up the translations of e. For each translation candidate c, its translation probability $P(e|c)$ is defined as $P(e|c) = P_{dic}(e|c) = 1/n_e$, where n_e is the number of possible translations of c. $P(e|c)$ will become zero if n_e is 0.

—The combined method: this approach was designed based on a linear combination of the above methods, that is, $P(e|c) = [P_{trans}(e, c) + P_{dic}(e|c)]/2$.

We used mean average precision (MAP) values to evaluate the retrieval performance. Table X shows the obtained results. The transitive translation approach performed slightly better than the dictionary-based approach, but the

Table X. The MAP Values Obtained by Applying Three Different Translation Probability Estimation Methods to the NTCIR-2 English–Chinese Retrieval Task

Method	Mean average precision
Transitive translation model	0.155
Dictionary-lookup method	0.143
Combined method	0.235

combined approach achieved the best performance. However, the achieved MAP values still can be improved. In Kwok's work, the best MAP value obtained was 0.316 [Kwok 2001]. In fact, our experiment was conducted to check the possibility of combining the probabilistic retrieval model with the proposed approach for CLIR applications. To realize the performance of the proposed approach clearly, we did not employ advanced techniques as reported in Kwok's work to deal with the difficult retrieval task. Although the overall retrieval performance was not obviously improved, the proposed approach was found to benefit some queries through the correct translations of unknown query terms; for example, the MAP value of the query "Ma Yo-yo cello recital" (Q46) increased from 0.205 to 0.446 due to the correct translation "馬友友演奏會."

6.2.2 *The CL-LSI-based Model.* As described in Section 2.1, the CL-LSI model can automatically construct a multilingual semantic space using latent semantic indexing (LSI) from a collection of multilingual documents. The following formula was adopted to rank documents related to queries [Furnas et al. 1988; Yang et al. 1997]:

$$M = \begin{bmatrix} A \\ B \end{bmatrix} = U \Sigma V^t \approx U_k \Sigma_k V_k^t \tag{7}$$

$$Sim(Q, D) = \cos(U_k^t Q, U_k^t D) \tag{8}$$

M is an input bilingual term-document matrix constituted by A and B, where A is a term-document matrix for the training documents in the source language, and B is a term-document matrix for the training documents in the target language. In addition, Σ_K, U_K and V_K are the matrices computed using singular value decomposition of M.

We used the program las2 in SVDPACKD [Berry et al. 1993] to perform singular value decomposition [Mori et al. 2001] and obtained a 400 dimension CL-LSI space. Actually, it was not straightforward how the CL-LSI model could be integrated into the proposed transitive translation approach. There were two ways to proceed: (1) using only anchor-text sets as the training corpus to construct the CL-LSI model, and (2) augmenting the translation lexicon using the proposed transitive translation approach. To evaluate them, three CL-LSI-based methods were, therefore, developed:

—The first method used the anchor-text sets as the corpus: Conventional CL-LSI models are normally trained via parallel texts, but the anchor-text set corpus was a comparable corpus. It was necessary to determine whether it was appropriate for constructing CL-LSI models.

Table XI. The MAP Values Obtained using Three Different CL-LSI-Based Methods for the NTCIR-2 English–Chinese Retrieval Task

Method	Mean average precision
CL-LSI using the anchor-text-set corpus	0.090
CL-LSI using the pseudo-parallel corpus	0.192
CL-LSI using the pseudo-parallel corpus plus the augmented lexicon	0.219

—The second method used pseudo-parallel texts as the corpus: To compare this method with the previous method, we randomly selected 33,043 (about 25%) documents from the NTCIR-2 Chinese collection and created a pseudo-parallel corpus by translating these documents into English.

—The third method used the pseudo-parallel corpus plus an augmented lexicon: In addition to using the pseudo-parallel corpus, this method also employed an augmented lexicon, which consisted of the top-5 translations of unknown query terms obtained using the proposed transitive translation approach.

The first two methods employed the LDC lexicon as the word list to segment the Chinese documents into words and generate the term-document matrix in the target language. The third method employed both the LDC lexicon and the augmented lexicon as the word list. Table XI shows the obtained retrieval performance. The obtained MAP values were not good enough. It is obvious that the anchor-text set corpus was not suitable for training CL-LSI models due to its nonparallel text property. But the proposed transitive translation approach might be useful for performing text segmentation and augmenting the translation lexicon, which is useful for Oriental languages. However, the above experiments revealed that the proposed approach is more effective in combination with the probabilistic retrieval model to improve the retrieval performance.

7. AN APPLICATION: TRANSLATION SUGGESTION FOR CROSS-LANGUAGE WEB SEARCH

As mentioned above, the need among real users for cross-language Web search is increasing. We have developed a cross-language Web search (CLWS) system called LiveTrans[5] based on the proposed approach. The LiveTrans system is a prototype of a meta-search engine that provides English–Chinese query translations for the retrieval of both Web pages and images in the greater China area. For example, as shown in Figure 10, for the user query “national palace museum,” three effective Chinese translations which were not included in general-purpose bilingual lexicons could be automatically extracted, that is, “國立故宮博物院” (National Palace Museum) “故宮” (an abbreviation of National Palace Museum) and “故宮博物院” (Palace Museum).

Since real Web queries are often short, it was not easy to adopt conventional cross-language retrieval models for use in the system. The system, therefore, combines the proposed approach with dictionary lookup and produces translation suggestions in response to input queries. For each translation t of source query s, the system assigns a weight, $W(t)$, and only a certain number of the

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

Fig. 10. 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" (National Palace Museum), "故宮" (an abbreviation of National Palace Museum), "故宮博物院" (Palace Museum), etc.

translations with higher weights are sent to the backend search engines for search. The weighting value is obtained by using the weighted sum of the inverse values of the ranks obtained using anchor text mining and dictionary lookup, respectively. The function $W(t)$ is defined below:

$$W(t) = (1-\alpha)(AR_t)^{-1} + \alpha(DR_t)^{-1}, \tag{9}$$

where $0 \leq \alpha \leq 1$.

AR_t is basically the rank of t in the set of all possible translations of s that are extracted through the mining of the anchor-text sets containing s. The rank is obtained by sorting the similarity values of the translations with s. The estimation of the values is obtained using Algorithm 5 ($k = 1$) as described in Section 4.4. AR_t is the default value if t cannot be found using the proposed approach.

On the other hand, DR_t is the rank of t in the set of all possible translations of s that are extracted through dictionary lookup. The rank is obtained by sorting the weights of these translations with s. DR_t will be the default value if t cannot be found using dictionary lookup. Since the used translation dictionary has no predefined rank for each of the translations, the method used to determine the ranks of the translations is based on their approximate frequencies on the Web.

A preliminary experiment, as shown in Table XII, was also conducted to compare the performance of anchor-text mining, dictionary lookup and the

Table XII. Top-*n* Inclusion Rates Obtained with Different Approaches

Method	Top1	Top2	Top3	Top4	Top5
Dictionary lookup	12.4%	20.2%	23.3%	28.7%	30.2%
Anchor-text mining	49.2%	58.1%	60.9%	61.6%	62.0%
Anchor-text mining plus CL	52.7%	60.1%	62.5%	63.1%	63.9%
Combined approach	55.8%	60.8%	64.0%	65.9%	67.8%
Combined plus CL approach	59.7%	63.2%	66.0%	68.2%	70.1%

combined approach in translating traditional Chinese into simplified Chinese, respectively. The dictionary mentioned in Section 5.1 was used. By examining the top-1 translations for the 258 test queries introduced in Section 5.1, it was found that the inclusion rates ranged from 12.4% for dictionary lookup to 59.7% for the combined approach plus the use of the CL algorithm. The achieved performance shows that anchor-text mining approach is really more suitable for practical Web search than dictionary lookup. Of course, combining these techniques could be the best choice.

8. CONCLUSION

In this article, we have proposed a transitive translation approach to extracting multilingual translations of Web queries through the mining of Web anchor texts and link structures. The translation equivalents of a query term can be extracted via its translations in an intermediate language. To reduce interference due to translation errors, the approach further integrates a competitive linking algorithm into the process of determining the most probable translation. This approach is particularly useful when the considered language pair lacks a sufficient number of anchor texts.

Preliminary experimental results show the feasibility of the proposed approach for term translation extraction. To determine the performance of the proposed approach when applied to CLIR, some experiments were conducted on the NCTIR-2 English–Chinese IR task using both CL-LSI and the probabilistic retrieval model. The proposed translation extraction approach was found to be effective in extracting translations of unknown queries, and it was easy to combine with the probabilistic retrieval model to improve the retrieval performance. In addition, we found that it could augment the translation lexicon and aid construction of a CL-LSI model for Oriental languages. Furthermore, with our novel combination of the anchor-text mining technique and bilingual dictionary lookup, the prototype CLWS system has been shown to be able to generate effective translation suggestions for popular queries. However, there are still some problems that need to be further investigated; for example, the performance of the proposed approach when the corpus size changes needs to be studied. In addition, there are large amounts of multilingual information embedded in other Web resources, such as bilingual and multilingual Web pages, search results from search engines, etc. Finding ways to exploit these resources is one of our future research directions.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Mark Sanderson and the anonymous reviewers for their valuable comments and suggestions. Many thanks are given to Mr. Jin-Shea Kuo, Shui-Lung Chuang, Shih-Jui Lin, Jei-Wen Teng and Chien-Chung Huang for their very helpful discussions and supports in preparing the experiments.

REFERENCES

AHONEN, H., HEINONEN, O., KLEMETTINEN, M., AND VERKAMO, A. 1999. Finding co-occurring text phrases by combining sequence and frequent set discovery. In *Proceedings of IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications*, 1–9.

AMITAY, E. AND PARIS, C. 2001. Automatically summarizing web sites—Is there a way around it? In *Proceedings of ACM 9th International Conference on Information and Knowledge Management*. ACM, New York, 173–179.

BALLESTEROS, L. AND CROFT, W. B. 1998. Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of ACM-SIGIR '98*. ACM, New York, 64–71.

BERRY, M., DO, T., O'BRIEN, G., KRISHNA, V., AND VARADHAN, S. 1993. *SVDPACKC* (*Version 1.0*)*User's Guide, Computer Science Department, University Tennessee.*

BORIN, L. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th COLING*, 97–103.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, 107–117.

BROWN, P., PIETRA, S. A. D., PIETRA, V. D. J., AND MERCER, R. L. 1993. The mathematics of machine translation. *Comput. Ling. 19*, 2, 263–312.

CHAKRABARTI, S., DOM, B., GIBSON, D., KLEINBERG, J., RAGHAVAN, P., AND RAJAGOPALAN, S. 1998. Automatic resource list compilation by analysing hyperlink structure and associated text. In *Proceedings of the 7th World Wide Web Conference/Computer Networks and ISDN Systems*, *30*, 65–74.

CHEN, K. H. AND CHEN, H. H. 2001. The Chinese text retrieval tasks of NTCIR workshop 2, In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

CHIEN, L. F. 1997. PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of ACM-SIGIR '97*. ACM, New York, 50–59.

DAGAN, I., CHURCH, K. W., AND GALE, W. A. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, 1–8.

DEOGUN, J. S., RAGHAVAN V. V., AND SERVER, H. 1997. Data mining: Research trends, challenges, and applications. In *Rough Sets and Data Mining: Analysis of Imprecise Data*. Kluwer Academic Publishers, 9–45.

DUMAIS, S. T., LANDAUER, T. K., AND LITTMAN, M. L. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proceedings of ACM-SIGIR'96 Workshop on Cross-Linguistic Information Retrieval*. ACM, New York, 16–24.

DUMAIS, S. T., LETSCHE, A., LITTMAN, M. L., AND LANDAUER, T. K. 1997. Automatic Cross-Linguistic Retrieval Using Latent Semantic Indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 15–21.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., AND UTHURUSAMY, R. 1996. *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

FELDMAN, R. AND DAGAN, I. 1995. KDT—Knowledge discovery in texts. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, 112–117.

FELDMAN, R., AUMANN, Y., AMIR, A., KLOESGEN, W., AND ZILBERSTIEN, A. 1997. Maximal association rules: A new tool for mining for keyword co-occurrences in document collections. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 167–170.

FENSEL, D. 2001. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, New York.

FUNG, P. AND YEE, L. Y. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Conference of the Association for Computational Linguistics*, 414–420.

FURNAS, G. W., DEERWESTER, S., DUMAIS, S. T., LANDAUER, T. K., HARSHMAN, R. A., AND LOCHBAUM, K. E. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of ACM-SIGIR'88*. ACM, New York, 465–480.

GOLLINS, T. AND SANDERSON, M. 2001. Improving cross language information with triangulated translation. In *Proceedings of ACM-SIGIR00*. ACM, New York, 90–95.

HEARST, M. 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 3–10.

HIEMSTRA, D. AND DE JONG, F. 1999. Disambiguation strategies for cross-language information retrieval. In *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, 274–293.

HULL, D. A. AND GREFENSTETTE, G. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the ACM-SIGIR'96*. ACM, New York, 49–57.

KLEINBERG, J. 1998. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, 668–677.

KNIGHT, K. 1997. Automating knowledge acquisition for machine translation. *AI Mag. 18*, 4.

KWOK, K. L. 2001. NTCIR-2 Chinese, cross language retrieval experiments using PIRCS. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

LU, W. H., CHIEN, L. F., AND LEE, H. J. 2001. Anchor text mining for translation of web queries. In *Proceedings of the 2001 IEEE International Conference on Data Mining*. IEEE Computer Society Press, Los Alamitos, Calif., 401–408.

LU, W. H., CHIEN, L. F., AND LEE, H. J. 2002a. Translation of web queries using anchor text mining. *ACM Trans. Asian Lang. Inf. Proc. 1*, 159–172.

LU, W. H., CHIEN, L. F., AND LEE, H. J. 2002b. A transitive model for extracting translation equivalents of web queries through anchor text mining. In *Proceedings of the 19th International Conference on Computational Linguistics* (*COLING2002*). 584–590.

LAVRENKO, V., CHOQUETTE, M., AND CROFT, W. B. 2002. Cross-lingual relevance model. In *Proceedings of ACM-SIGIR 2002*. ACM, New York, 175–182.

MELAMED, I. D. 2000. Models of translational equivalence among words. *Comput. Ling. 26*, 2, 221–249.

MORI, T., KOKUBU, T., AND TANAKA, T. 2001. Cross-lingual information retrieval based on LSI with multiple word spaces. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

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. In *Proceedings of ACM-SIGIR'99*, 74–81.

OARD, D. AND DIEKEMA, A. 1998. Cross-language information retrieval. *Ann. Rev. Inf. Sci. Tech. 33*, 223–256.

RAPP, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics*, 519–526.

RESNIK, P., OARD, D., AND LEVOW, G. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the 1st International Conference on Human Language Technology Research*.

SILVERMAN, B. G., BACHANN, M., AND AKHARAS, K. 2001. Do what I mean: Online shopping with a natural language search agent. *IEEE Intel. Syst.* Jul/Aug, 48–53.

SILVERSTEIN, C., HENZINGER, M., MARAIS, J., AND MORICZ, M. 1998. Analysis of a very large altavista query log. Tech. Rep. 1998–014. Digital Systems Research Center.

SIMARD, M. 2000. Multilingual text alignment. In *Parallel Text Processing*, J. Veronis, Eds. Kluwer Academic Publishers, The Netherlands, 49–67.

SMADJA, F., MCKEOWN, K., AND HATZIVASSILOGLOU, V. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Comput. Ling. 22*, 1, 1–38.

SODERLAND, S. 1997. Learning to extract text-based information from the world wide web. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 251–254.

WONG, S. K. M., AND YAO Y. Y. 1995. On Modeling Information Retrieval with Probabilistic Inference. *ACM Trans. Inf. Syst. 13*, 38–68.

XU, J., WEISCHEDEL, R., AND NGUYEN, C. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of ACM-SIGIR 2001*. ACM, New York, 105–110.

YANG, Y., CARBONELL, J. G., BROWN, R. D., AND FREDERKING, R. E. 1997. Translingual information retrieval: Learning from bilingual corpora. *Artif. Intel. J. 103*, 323–345.

Received August 2002; revised April 2003; accepted August 2003