Chinese OOV Translation and Post-translation Query Expansion in Chinese–English Cross-lingual Information Retrieval

YING ZHANG, PHIL VINES, and JUSTIN ZOBEL
School of Computer Science and Information Technology, RMIT University

Cross-lingual information retrieval allows users to query mixed-language collections or to probe for documents written in an unfamiliar language. A major difficulty for cross-lingual information retrieval is the detection and translation of out-of-vocabulary (OOV) terms; for OOV terms in Chinese, another difficulty is segmentation. At NTCIR-4, we explored methods for translation and disambiguation for OOV terms when using a Chinese query on an English collection. We have developed a new segmentation-free technique for automatic translation of Chinese OOV terms using the web. We have also investigated the effects of distance factor and window size when using a hidden Markov model to provide disambiguation. Our experiments show these methods significantly improve effectiveness; in conjunction with our post-translation query expansion technique, effectiveness approaches that of monolingual retrieval.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: General

General Terms: Algorithms, Languages

Additional Key Words and Phrases: CLIR, query translations, OOV terms, translation disambiguation, web mining, HMM, post-translation query expansion, mutual information

1. INTRODUCTION

The web contains documents written in many languages and as many languages are used in queries. Although English is the most widely used language on the web, the use of Chinese continues to grow; it is now the third or fourth most commonly used query language. In addition, many Chinese speakers have some knowledge of English, making it attractive to develop effective methods for Chinese–English cross-lingual information retrieval (CLIR).

Earlier versions of portions of this work have previously appeared in Zhang and Vines [2004b] and Zhang and Vines [2004a].

Authors' address: Ying Zhang, Phil Vines, and Justin Zobel, School of Computer Science and Information Technology, RMIT, GPO Box 2476V, Melbourne 3001, Australia.

There are several approaches to implementation of CLIR. Perhaps the most popular is to use a dictionary to translate the queries into the target language and then use monolingual retrieval. As with other CLIR language pairs, in Chinese–English CLIR the accuracy of dictionary-based query translation is limited by two factors: the presence of out-of-vocabulary (OOV) words and translation ambiguity. The OOV problem arises from the fact that some Chinese query terms are not found in translation resources, such as bilingual dictionaries and parallel corpora. For example, the query may concern current affairs and thus contain new words or translated words that are outside the scope of the translation dictionary; or the query may contain proper nouns—such as brand names, place names or personal names—that are not included in the translation dictionary. Although only some queries contain OOV terms, incorrect translation of such terms almost inevitably leads to disastrous results. In some literature, especially in relation to English-to-Chinese translation [Mend et al. 2004], a further problem, the need for phrase detection, is also discussed. Groups of two or more words may have a special meaning when translated as a unit that differs from that obtained by translating them individually. However, when translating from Chinese, the need to correctly handle such phrases can be treated as part of the OOV problem, as there are no explicit delimiters between words and the difference between a word and phrase is not well defined. The major distinction is between individual characters and "words," which are strings of characters.

Existing systems tackle the OOV problem in several ways. We have observed that some systems sometimes do not translate an OOV term at all (such as BabelFish), while others translate character by character using Pinyin, with disastrous results. Some research prototypes appear to require manual intervention [Chen et al. 2000] in order to correctly segment OOV terms. In our approach, we exploit juxtaposition of English text and Chinese text on the web to identify OOV terms and thus determine the appropriate segmentation. The technique is segmentation-free as we do not segment a query into words when searching for OOV terms. This has proved to be a "catch-22" problem in the past. If the word is unknown, it can not be segmented correctly. If the word cannot be segmented correctly, it was assumed to be impossible to search for a translation.

In this paper we examine the problem of OOV translation in the context of Chinese–English CLIR. We circumvent segmentation difficulties in OOV translation by using the entire Chinese query to search on the web. By mining the web to collect a sufficient number of Chinese–English co-occurrences and applying statistical techniques, we are then able to infer the OOV term and its appropriate English translation with reasonable confidence. The idea of using the web to search for translations is not new [Chen et al. 2000; Mend et al. 2004]; however, our technique can extract translations that were previously undetected, or only detected after manual intervention to provide correct segmentation. For example: suppose the query is "$x_1x_2x_3x_4x_5x_6x_7$" and the correct segmentation is "$x_1x_2x_3|x_4|x_5x_6x_7$." However "$x_5x_6x_7$" is not in any available dictionary and thus it is an OOV term. If we first applied a segmenter, we might obtain "$x_1x_2x_3|x_4|x_5|x_6x_7$." We could be tempted to conclude that two or more characters “x_4x_5” are an OOV term, but that would be wrong. It is unwise to assume anything about the segmentation of a query when trying to identify OOV terms. A potential drawback to our technique is that it might obtain candidate translations for all query terms. However, our technique works because unusual Chinese terms borrowed from English tend to be accompanied by the original English terms on the web, but this is not the case for common terms.

After using the entire Chinese query to fetch web documents written in Chinese, we collect the English text that is preceded by any substring of the original Chinese query. By applying simple statistical techniques, we are then able to detect Chinese OOV terms and appropriate English translations with reasonable confidence. This OOV translation technique leads to substantial improvement in effectiveness. We also tested several query expansion techniques, including a novel use of mutual information to select additional query terms. This technique provided a further improvement in retrieval effectiveness.

Extending our original NTCIR-4 paper [Zhang and Vines 2004a], we have used a series of additional experiments to identify the value of the individual components of our process and to explore the sensitivity to parameter settings.

Together, these methods show that CLIR using the web and public dictionaries can approach the effectiveness of monolingual information retrieval.

2. BACKGROUND

2.1 Chinese OOV Term Translation

NTCIR is a forum for evaluating information retrieval techniques for Asian languages (see `www.research.nii.ac.jp/ntcir/`). The NTCIR-4 task we participated in involves using Chinese queries to retrieve English documents [Kishida et al. 2004]. The English document collection from the NTCIR-4 Workshop CLIR task contains 347,376 news articles from 1998 to 1999. There are 58 Chinese topics, each containing four parts: *title*, *description*, *narrative*, and *key words*. The principal translation problem we investigated was Chinese query terms that are missing from translation dictionaries or out-of-vocabulary (OOV) terms.

The OOV problem can be divided into two classes, each requiring different approaches. By understanding the different kinds of OOV terms, we can classify the previous approaches and see which types of terms they are likely to provide solutions for. In particular, it is useful to distinguish between OOV terms that have been rendered into the target language by means of transliteration and those that have been rendered by some form of semantic translation. In addition, there are terms that have been translated by a combination of these approaches, such as “起亞汽車” (Kia Motors).

2.1.1 *Transliteration.* In Chinese, each character represents a syllable. Chinese has relatively few distinct phonemes and many sounds in English are not present in Chinese (and vice versa), so Chinese terms transliterated from English often do not closely resemble the original English pronunciation. Since, many English phonemes often map to one Chinese phoneme, backward transliteration—recovery of the English term from the Chinese—is difficult. An additional problem with this process is that a given English term may have more than one Chinese transliterated equivalent. For example, “Michael Jordan” is transliterated into “麥可喬丹” in Taiwan, but into “米高佐敦” and “邁爾克.喬丹” in Hong Kong. This is because different Chinese communities transliterate English terms in different ways.

Mend et al. [2004] developed a system that incorporates transliteration from English to Chinese to deal with English OOV terms. (The results of such a system can be added to Chinese-to-English translation dictionaries, as a further enhancement to the methods discussed below.) Their system appears successful where transliteration rules have been strictly applied and little variation exists, but is not always able to pick the best transliteration when several are in use.

Lin and Chen [2002] developed a backward OOV transliteration process that attempts to recover the original English term from the transliterated Chinese term. This approach requires a candidate set of English terms. Chinese terms and the candidate English terms were converted into a common form using the International Phonetic Alphabet and a similarity function was then applied to select the closest match. Although there are many sounds in English that have no equivalent Chinese sound and vice versa, they exploited the degree of consistency used in the transliteration process. Their system is trained on sample data to learn phonetic similarities, then produces a rank list of possible English name equivalents. Lin and Chen report that the average rank of correct English term was 2.04. In their experiments they usually had the correct transliteration available to them. In previous work we found that by mining the web to collect OOV terms and then using the web to search for translations, we were able to translate 61% of terms correctly and 31% of terms approximately [Zhang and Vines 2004b]. These results are not directly comparable, but show that it is rarely necessary to have the correct translation available.

2.1.2 *Semantic Equivalence.* When OOV term translation is based on meaning rather than sound, transliteration techniques fail. For example, “胚胎乾細胞” (embryonic stem cell) and “國際太空站” (international space station) are semantic translations and cannot be connected using transliteration. Also, backward transliteration of Japanese and Korean names will generally fail, such as “黑澤明” (Akira Kurosawa), as they have been transliterated via different rules.

In this case, schemes such as web mining [Lu et al. 2002] and approaches based on parallel corpora [McEwan et al. 2002; Yang and Li 2002; Chen and Nie 2000] can be applied. Lu et al. [2002] exploited the existence of web pages written in different languages that had anchor text pointing to the same page. By applying statistical techniques, the top-ranked translation proved to be correct in 53% of cases. While this technique is useful, the key drawback is that it requires a web page relating to the Chinese OOV term and sufficient interest to cause linking from a foreign language site. Their technique found several company names, but did not appear to find names of individuals, place names, and other such terms that are rarely the subject of a web page.

McEwan et al. [2002] attempted to locate parallel documents on the web and used these to build bilingual dictionaries. However, such approaches suffer from lack of sufficient high-quality parallel texts. Yang and Li [2002] successfully mined parallel Chinese–English documents from the web, but, as is common with parallel mining, considered only a small domain—press releases from the Hong Kong Government. Chen and Nie [2000] also obtained good results in alignment of English–Chinese documents, but only 427 documents from the Hong Kong government were used in their experiments.

2.2 Resolution of Translation Ambiguity

Dictionary-based query translation is prone to errors, because of the possibility of selecting the wrong translation of a query term from among the translations provided by the translation dictionary. This is the *translation ambiguity* problem. It is particularly severe when users enter short queries (often two or three words), a situation in which it may not be possible for even a human to determine the intended meaning from the available context.

There have been several approaches to the ambiguity problem. These have included using co-occurrence statistics in the target document collection [Ballesteros and Croft 1998; Gao et al. 2002], using mutual information [Mirna 2000; Maeda et al. 2000], and probabilistic methods based on a language model [Federico and Bertoldi 2002; Sun et al. 2003].

2.2.1 *Co-occurrence Statistics.* Ballesteros and Croft [1998] describe a technique that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary-based translation. Their hypothesis is that the correct translations of query terms should co-occur in target language documents and incorrect translation should tend not to co-occur. They measured the importance of co-occurrence of the elements in a set by the *em* metric, which is a variation of EMIM [van Rijsbergen 1977]. Gao et al. [2002] used a technique based on mutual information and showed that closer words tend to have stronger relationships and improved the basic co-occurrence approach by adding a distance factor. In this paper, we explore the effectiveness of using this approach in combination with a hidden Markov model.

2.2.2 *Mutual Information.* Mirna [2000] proposed a term-sense disambiguation technique for selecting the best translation sense of a term from all possible senses given by a bilingual dictionary. Given a set of original query terms, they select the best translation for each of the terms such that resulting set of selected translations contains translations that are mutually related or statistically similar with one another. The degree of similarity or association relation between terms was calculated with a term association measure, the Dice similarity coefficient, which is also used in document or term clustering. Maeda et al. [2000], working on Japanese–English CLIR, have used a search engine to collect the cooccurrence information between terms in web documents, and applied a modified Dice coefficient to calculate the mutual information between terms. They used one document as the window of co-occurrence.

2.2.3 *Language Modeling.* Federico and Bertoldi [2002] have used N-best translations provided by the translation dictionary, together with a term weighting adjustment scheme, with good results. Where the motivation for doing this is language mismatch—that is, several words in the target language having a similar meaning to the original word in the source language—then the effect is similar to query expansion and can improve recall effectiveness. However, when trying to discover the most appropriate translation of an OOV term it is possible that some of the candidate translations are entirely wrong, thus leading to considerable loss of retrieval effectiveness if they are included in the query.

3. CHINESE OOV TERM DETECTION AND TRANSLATION

When looking for English translations of Chinese OOV terms, they need to be appropriately detected in the query. Many existing systems use a segmenter to determine Chinese word boundaries. However, if the Chinese OOV term is currently unknown, there is no information to indicate how it should be segmented. In other work [Chen et al. 2000], this problem appears to have been overcome by manual intervention to provide appropriate segmentation. However, it is clearly desirable that the segmentation be either automatic or, as in the case of the technique we describe, unnecessary.

When a large corpus of Chinese text is available, it is possible to apply statistical techniques to identify named entities that are not present in translation dictionaries. Sun et al. [2003] used a trigram stochastic model to detect named entities. Their technique had a success rate of approximately 80%; however, they did not attempt translation. Such a technique is not practical in our situation, as we do not have a Chinese corpus to work with, and in any case it is unlikely that a corpus would contain many of the OOV terms that occur in news and current affairs.

The basis of our approach is the observation that most translated English terms tend to accompanied by the original English terms on the web, typically immediately after the Chinese text, but general terms do not. For example, the text might contain "世紀之毒戴奧辛 (Dioxin)" where "世紀之毒戴奧辛" is a sequence of Chinese characters and "Dioxin" is the original English term for "戴奧辛". By mining the web to collect a sufficient number of such instances for any given word and applying statistical techniques, we hypothesize that we are then able to infer an appropriate translation with reasonable confidence. In formulating our approach, we also considered English text that was not immediately adjacent to the Chinese query terms. However, we found that such text was only rarely a reliable translation. In some cases, we found only a small number of Chinese–English co-occurrences and our approach proved to be robust in such situations.

OOV translation is only the first step in the retrieval process. We add these terms into both a segmentation and a translation dictionary. We then use the standard dictionary-based query translation steps of segmentation and translation disambiguation. In summary, our procedure consists of three steps: extraction of the web text, collection of co-occurrence statistics, and translation selection.

... 北野武性愛狂想曲【Geetting Any】...
.. 雖然之前還爲石井隆的《Gonin》作過序..
... 因爲我這個只對大衛・林區(David Lynch)等幾位導演或犯罪..
... 歡迎到eWorld 好站...
... ... 導演北野武(Takeshi Kitano) ... ...
... ... 北野武(Takeshi Kitano) ...
... ...導演北野武Takeshi Kitano ... ...
... 他是BeatTakeshi ...
... 菊次郎的夏天Kikujiro ..
... 【盲劍俠】ZATOICHI-劍客側寫...
... 【盲劍俠】ZATOICHI-劍客側寫...
...【那個凶暴的男子】(Violent Cop) ...
... 【3比4X十月】(Boiling Point) ...
... 《３比４Ｘ十月》（BOILING POINT 1990）...
... 《奏鳴曲》（SONATINE 1993）... 《恣在年少》（KIDS RETURN 1996）...
... 北野武(1948 -) ...
... 導演：北野武Takeshi Kitano ... 演員：北野武Takeshi Kitano ...
... 淺野忠信Tadanobu Asano ...
... The Spinning Image 「視覺暫留」作者：Daniel Auty ...
... ... 北野武（Takeshi Kitano ）... ...
... ... 導演(Director) ...
... 電影DVD ... ...
... ... 導演Takeshi Kitano ...
... ... 電影: CHARLIE AND THE CHOCOLATE FACTORY ...
... ... 北野武(Takeshi Kitano) ... ...

Fig. 1. Web text retrieved using "北野武導演的電影" (Director Takeshi Kitano's films).

3.1 Extraction of Web Text

First, we query the web to identify strings that contain the Chinese query terms and some English text.

1. Use a search engine to fetch the top 100 Chinese documents, using the entire Chinese query. (A side effect of using a reliable search engine is that only good-quality web text is returned. This reduces the likelihood of noisy translations being collected.)
2. For each returned document, the title and the query-biased summary are extracted.

For example, consider a query "北野武導演的電影" (Director Takeshi Kitano's films) composed of four Chinese terms: "北野武" (Takeshi Kitano), "導演" (Director), "的 and "電影" (films). Suppose that "北野武" is an OOV term, "的" is a structural particle, and "導演" and "電影" are in-vocabulary terms. We used this query to retrieve a series of titles and query-biased summaries of web text that contain English text, as shown in Figure 1, and as can be seen, "Takeshi Kitano" is the most common English text and "北野武" is the Chinese text most commonly observed in the context.

3.2 Collection of Co-occurrence Statistics

We then collect co-occurrence information from the data we obtained. Although much English text is present in web pages written in Chinese, not all of it is useful in OOV translation. We only consider the English text that occurs immediately after the Chinese query substrings, because, if such an English

Table I. Frequency of Co-occurrence of English Terms and Chinese Query Substrings

e	f_e	c_j	$\|c_j\|$	$f(e, c_j)$
Takeshi Kitano	8	北野武	3	7
		導演	2	1
Director	1	導演	2	1
DVD	1	電影	2	1
CHARLIE AND THE CHOCOLATE FACTORY	1	電影	2	1
1948	1	北野武	3	1
Gonin	1	的	1	1

term occurs at a high frequency, it almost invariably serves as the translation of that Chinese string.

1. Where English text occurs, check the immediately preceding Chinese text to see if it is a substring of the Chinese query.
2. Collect the frequency of co-occurrence of each distinct English string and all Chinese query substrings that appear immediately prior.

For each distinct English string e with the frequency f_e, we obtained a group of associated Chinese query substrings c_j with the length $|c_j|$ and the co-occurrence frequency $f(e, c_j)$. Extending the example from in Figure 1, this information is summarized in Table I.

3.3 Translation Selection

Incorrect translations can, as discussed earlier, greatly degrade effectiveness. For this reason, we select only the best translation for a Chinese OOV term from Table I as follows:

1. Select the English text e with the highest f_e, since the remaining English text that occurs with the highest frequency is more likely to provide the correct translation compared to other text with the lower frequency.
2. For this English text e, select the associated Chinese query substring c_j with the highest $f(e, c_j)$. In the event of a tie, we use $|c_j|$ to discriminate.
3. If the selected Chinese query substring c_j cannot be found in the Chinese segmentation dictionary, we treat it as OOV term and add it into the Chinese segmentation dictionary and (c_j, e) into the translation dictionary.

In this case, "北野武" is identified as a Chinese OOV term and "Takeshi Kitano" is extracted as its English translation.

A given Chinese term may have more than one English transliteration. For example, "Osama" and "Usama" are transliterated into the same Chinese term "奧薩瑪". However, this phenomenon is rare. Our methods tend to choose the most common form.

4. TRANSLATION DISAMBIGUATION

Each set of English translations E is a sequence of words $e_1, e_2, e_3, ...e_n$. We use a probability model $P(E) = P(e_1, e_2, e_3, ...e_n)$ to estimate the maximum likelihood of each sequence of words. We select English translations E with the highest $P(E)$ among all possible translation sets.

Our disambiguation technique is based on hidden Markov models (HMM) [Miller et al. 1999], which have been used widely for probabilistic modeling of sequence data.

$$P(e_1, e_2 \ldots e_n) = P(e_1) \prod_{a=2}^{n} P(e_a | e_{a-1})$$

To compute the probability of a sequence of words, we need to calculate the quantities $P(e)$, the probability of word e, and $P(e|e')$, the probability of e in the context of e':

$$P(e) = \frac{f(e)}{N}, P(e|e') = \frac{P_w(e, e')}{\sum_{e''} P_w(e'', e')}$$

where $f(e)$ is the collection frequency of term e, N is the number of terms in the document collection, and $P_w(e, e')$ is the probability of term e' occurring after term e within a window of size w.

The zero-frequency problem arises in the context of probabilistic language models, when the model encounters an event in a context in which it has not been seen before. Smoothing provides a way to estimate and assign the probability to that unseen event. We use the following absolute discounting and interpolation formula, which applies the smoothing method proposed by Federico and Bertoldi [2002]. In this method,

$$P(e|e') = \max\left\{\frac{f_w(e, e') - \beta}{N}, 0\right\} + \beta P(e) P(e')$$

where $f_w(e, e')$ is the frequency of term e' occurring after term e within a window size w.

Federico and Bertoldi [2002] successfully used this formula to compute the frequency of term e' and e within a text window of fixed size through an order-free bigram language model in their work. However, they did not give detailed information about the size of the text window. The absolute discounting term β is equal to the estimate proposed by Ney et al. [1994]:

$$\beta = \frac{n_1}{n_1 + 2n_2}$$

where n_k representing the number of terms with the collection frequency k.

We have observed that two words being in close proximity generally provides stronger correlation and produces more credible results for disambiguation of translation than does cooccurrence of two words in a large window. Gao et al. [2002] applied a decaying factor to the mutual information calculation; their experiments showed that the decaying factor can be used to discriminate strong and weak term correlation.

$$D(e, e') = \mathbf{e}^{-\alpha(\mathrm{Dist}(e, e') - 1)}$$

Table II. Effect of Window Size and Distance Factor on Translation Disambiguation in CLIR, Mean Average Precision Values for Title Queries

	α					
w	*0.0*	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
2	0.2112	0.2112	0.2112	0.2112	0.2112	0.2112
4	0.2166	0.2166	0.2166	0.2166	0.2166	0.2166
6	0.2152	0.2152	0.2152	0.2152	0.2152	0.2152
8	0.2152	0.2152	0.2152	0.2152	0.2152	0.2152
10	0.2161	0.2161	0.2161	0.2161	0.2161	0.2161
14	0.2160	0.2160	0.2160	0.2160	0.2160	0.2160

where $\text{Dist}(e, e')$ is the average distance between e and e' in the document collection. Therefore, we have added this distance factor $D(e, e')$ into the probability calculation, to give:

$$P(e|e') = \left[\max\left\{\frac{f_w(e,e') - \beta}{N}, 0\right\} + \beta P(e)P(e')\right] \times D(e,e')$$

Gao found a value of $\alpha = 0.8$ gave the best results when combined with their mutual information model. However there was not much difference for values of α between 0.2 and 1.0. The major difference was for $\alpha = 0.0$, that is, no distance factor. We investigated the effect of this parameter on disambiguation performance. (The full details of our experimental setup is described in Section 6.) The results are shown in Table II. It can be seen that our test collection is quite insensitive to this parameter, even for $\alpha = 0.0$. We believe that the reason for this is that the HMM model is a superior technique where sequence data is involved, and is not significantly improved by the addition of a decaying distance factor. Note that variation in window size used to collect word association information has a small effect on the outcome, with $w = 4$ producing the best results.

5. POST-TRANSLATION QUERY EXPANSION

Query expansion has been widely investigated in monolingual retrieval [Robertson and Jones 1976; Xu and Croft 2000; Ruthven 2003]. It has generally provided improvement in retrieval effectiveness, whereas other approaches that use document structure or thesauri expansion have been less successful [Mandala et al. 1999]. We aimed, first, to measure the effect of using mutual information on post-translation query-expansion and, second, to investigate the effect of parameter value variability on query-expansion retrieval effectiveness. We also compare the query-expansion techniques we used at NTCIR-4 with those of other participants. We note that query expansion involves selection of parameter values that are not necessarily consistent from one collection to another [Billerbeck and Zobel 2004].

We applied an automatic feedback query-expansion approach that adds t terms from the top d retrieved documents to the translated query. We tested a two-stage process that first selects a set of candidate terms using standard term weighting metrics and then applies a mutual information procedure to select the final set of terms to be added. The motivation for this approach was

Table III. Post-translation Query Expansion Using tf and tf.idf, Mean Average Precision Values for Title Queries

		d					
t		*5*	*10*	*20*	*30*	*40*	*50*
0	—	0.2166					
5	*tf*	0.2157	0.2218	0.2259	0.2280	0.2280	0.2104
	tf.idf	0.2060	0.2286	0.2211	0.2224	0.2254	0.2239
10	*tf*	0.2158	0.2258	0.2175	0.2102	0.2022	0.2036
	tf.idf	0.2098	0.2202	0.2106	0.2041	0.2016	0.1933

that, not only should terms be "important" in terms of having a high weight, but they should be related to the query. Using the co-occurrence of the candidate term and all query terms in the collection is a plausible way to measure the word association. In the following sections we explain how these parameters were calculated.

5.1 Term Weighting

To provide a baseline for our query expansion experiments using mutual information, we experimented with two approaches to the selection of the set of candidate terms: *tf* and *tf.idf*, without the additional mutual information step. *tf* is the frequency with which the term occurs in the top d retrieved documents. *idf* is calculated as $\log(N/d_f)$, where d_f is the document frequency of the term and N is the total number of documents in the document collection. English stop words were removed from the retrieved documents prior to term selection.

We experimented with adding either 5 or 10 top ranked terms from the top-ranked 5, 10, 20, 30, 40, and 50 documents. The results of these experiments are shown in Table III. It can be seen that there is not a great difference in the results, but in all cases using *tf* to select the top terms was more effective than using *tf.idf*. It can also be seen that using more documents provided a slight improvement up to 30 documents, but no improvement after that.

5.2 Mutual Information

As explained above, our mutual information procedure involved selecting the top t terms from the set of candidate terms that have the highest degree of mutual information with all translated query terms. In order to select the best t terms, we need to calculate the mutual information of a term and a term set. The mutual information of a term x and a set S of terms is the sum of x with every term in the set S.

$$MI(x, S) = \sum_{s \in S} MI(x, s)$$

To measure the mutual information (MI) between a given term x and a term s within a window size of w, we used:

$$MI(x, s) = \log\left(\frac{f_w(x, s)}{f_x f_s} + 1\right)$$

Table IV. Effect of Number of Documents on Post-translation Query Expansion, Mean Average Precision Values for Title Queries

		w					
		4		16		20	
d		$t = 5$	$t = 10$	$t = 5$	$t = 10$	$t = 5$	$t = 10$
5	*tf*	0.2264	0.1965	0.2249	0.1983	0.2246	0.2022
	tf.idf	0.1947	0.1947	0.1815	0.0.1936	0.1848	0.1906
10	*tf*	0.2314	0.2154	0.2272	0.2127	0.2244	0.2210
	tf.idf	0.2185	0.2119	0.2114	0.2101	0.2111	0.2073
20	*tf*	0.2294	0.2051	0.2386	0.2075	0.2341	0.2097
	tf.idf	0.2171	0.2117	0.2049	0.1964	0.2068	0.2016
30	*tf*	0.2254	0.2090	0.2247	0.2084	0.2271	0.2109
	tf.idf	0.2119	0.2054	0.2092	0.1982	0.2105	0.1985
40	*tf*	0.2153	0.2055	0.2168	0.2043	0.2185	0.2049
	tf.idf	0.2163	0.1929	0.2193	0.1941	0.2201	0.1962

where $f_w(x, s)$ is the frequency with which x and s co-occur within a window size of w in the document collection; f_x is the collection frequency of x and f_s is the collection frequency of s. Addition of 1 to the frequency ratio means that a zero co-occurrence frequency corresponds to zero mutual information.

In addition to selecting the number of terms t and documents d that participate in query expansion, using of mutual information also requires selection of a window size w used to collect mutual information statistics. Further, the fact that we are using a two-step process requires that we decide how many terms to collect in the first stage, which we call the candidate set c, to consider in the second stage. We express this as a proportion of the number of terms added in the final stage. For example, if we add t terms to the query, we might collect an initial candidate set of $c = 2t$. Our post-translation query expansion using mutual information thus involves four parameters:

d: the number of the top-ranked documents returned
t: the number of the terms added
w: the window size used in mutual information
c: the size of the candidate set

There are many possible combinations of these parameters. We investigated the interaction of these parameters, as follows.

5.2.1 *Effect of Adding Documents.* As neither *tf* nor *tf.idf* was clearly superior, we experimented with using both of these to select terms for the next phase. From further experiments (not presented here), we determined that choosing $w = 4, 16$, or 20 and $t = 5$ or 10 gave slightly better results. The results are shown in Table IV. From this table we can observe slight improvements up to 20 documents and a decline after that. We also note the using 5 terms is always superior to using 10 terms, something we explore further below. Finally, using *tf* together with mutual information is more effective than using *tf.idf* with mutual information.

5.2.2 *Effect of Adding Terms.* Although our previous experiments (Table IV) suggest that $t = 5$ may provide the best results, we experimented

Table V. Effect of Number of Terms on Post-translation Query Expansion, Mean Average Precision Values for Title Queries

	t			
w	*5*	*10*	*20*	*30*
4	0.2294	0.2051	0.2027	0.1838
8	0.2263	0.2019	0.2057	0.1875
12	0.2279	0.2065	0.2052	0.1862
16	0.2386	0.2075	0.2054	0.1903

Table VI. Effect of Window Size on Post-translation Query Expansion, Mean Average Precision Values for Title Queries

	w						
t	*4*	*8*	*12*	*16*	*18*	*20*	*22*
5	0.2294	0.2263	0.2279	0.2386	0.2293	0.2241	0.2242
10	0.2051	0.2019	0.2065	0.2075	0.1974	0.2097	0.2074

with adding larger numbers of terms to see if this had any effect. As *tf* had proven superior to *tf.idf*, we persevered only with *tf*. The results of these experiments are shown in Table V and confirm that $t = 5$ produces the best results.

5.2.3 *Effect of Window Size.* Table VI shows the effect of window size on query expansion using mutual information. Again we have chosen other parameter values that appear to give optimal results. From the results in Table VI, we can see that there is no consistent trend, although $w = 16$ gives the best results when $d = 20$. Once again $t = 5$ performs best.

5.2.4 *Effect of Candidate Set Size.* A final issue is the number of terms collected in the first phase for consideration in the second phase. In the above experiments, we used a candidate set that was twice the number of terms ultimately required, namely $c = 2t$. We wondered if collecting more terms in the first phase might improve results, so we experimented with using larger candidate sets, namely $t = 3$ and $t = 4$. However, as can be seen in Table VII, this only led to a deterioration in performance.

5.3 Significance of Mutual Information in Query Expansion

After testing a large number of combinations, as outlined above, we selected the best results from query expansion using only *tf* to compare with the best results provided by query expansion using *tf with mutual information*. While this represents tuning, it allowed us to test whether using mutual information in query expansion is likely to provide any benefit.

We used the Wilcoxon-ranked signed to examine the statistical significance of our results. Our baseline title run *T-do* using disambiguation and OOV translation achieved a MAP of 0.2166. As shown in Table VIII, using query expansion based on *tf* with $t = 5$ and $d = 30$, we achieved 0.2280 which represents a 5% improvement; using *tf* and mutual information, with $d = 20$, $t = 5$, $c = 2t$, and $w = 16$, we achieved 0.2386, which represents

Table VII. Effect of Candidate Set Size on Post-translation Query Expansion, Mean Average Precision Values for Title Queries

	d					
	5		10		20	
c	$t = 5$	$t = 10$	$t = 5$	$t = 10$	$t = 5$	$t = 10$
$t \times 2$	0.2249	0.1983	0.2272	0.2127	0.2386	0.2075
$t \times 3$	0.1941	0.1974	0.2188	0.2151	0.2245	0.2012
$t \times 4$	0.1877	0.1957	0.2114	0.2018	0.2052	0.2042

Table VIII. Post-translation Query Expansion Using tf and tf with Mutual Information, Mean Average Precision Values for Title Queries

		d			
	w	5	10	20	30
tf with mutual information	*4*	0.2264	0.2314	0.2294	0.2254
	16	0.2249	0.2272	0.2386	0.2247
	20	0.2246	0.2244	0.2341	0.2271
tf	–	0.2157	0.2218	0.2259	0.2280

a 10% improvement. However, neither of these improvements is statistically significant.

6. QUERY TRANSLATION EXPERIMENTS

We used two dictionaries in our experiments: ce3 from the Linguistic Data Consortium (see `www.ldc.upenn.edu`), and the CEDICT Chinese–English dictionary (see `www.mandarintools.com/cedict.html`) to translate Chinese queries into English.

As described in Section 3, we detect Chinese OOV terms using a segmentation-free process and add them into a Chinese segmentation dictionary for later use. In the dictionary-based query translation phase, we used the updated Chinese segmentation dictionary to segment the queries and replace each query term using a set of English translations through a bilingual translation dictionary lookup.

6.1 Preprocessing

English stop words were removed from the English document collection. We used a stop list that contains 477 entries and the Porter stemmer [Porter 1980] to reduce words to stems. The Chinese queries were processed as follows:

1. In NTCIR-4, each Chinese query was presented as a list of comma-separated Chinese text. Our assumption is that each of these text strings is either a phrase or a word. Sixty-six of 163 Chinese strings cannot be found in the translation dictionaries. We treat all of these as potential Chinese OOV terms, although some of them could be translated word by word.
2. Using these 66 Chinese strings as queries, we applied our translation extraction technique and added extracted translation pairs into the translation dictionary.

Table IX. Run Descriptions

		Translation	*OOV*	*Query*
	RunID	*disambiguation (d)*	*translation (o)*	*expansion (q)*
Cross-lingual runs	*T-BabelFish*	×	×	×
	T-BabelFish+q	×	×	√
	T-d	√	×	×
	T-dq	√	×	√
	T-do	√	√	×
	T-doq	√	√	√
	D-BabelFish	×	×	×
	D-BabelFish+q	×	×	√
	D-d	√	×	×
	D-dq	√	×	√
	D-do	√	√	×
	D-doq	√	√	√
Monolingual runs	*T-mono*	×	×	×
	T-mono+q	×	×	√
	D-mono	×	×	×
	D-mono+q	×	×	√

3. We compiled a segmentation dictionary using the two translation dictionaries and updated it at run time. We used dictionary-based segmentation with greedy-parsing to segment the Chinese queries.
4. The translation dictionary was used to replace each query term by all English translations.
5. Our translation disambiguation technique was used to select the most appropriate translation for each Chinese query.

6.2 Experimental Design

Our retrieval experiments consist of 16 runs. In *T-runs*, we have used the *titles* of the Chinese topics as queries and in *D-runs* the *description* fields are used as queries to retrieve the documents from the English document collection. The relevance judgments provided by NTCIR are at two levels— strictly relevant documents known as *rigid relevance*, and likely relevant documents, known as *relaxed relevance*. In this paper, we used only *rigid relevance* to report our results. Our CLIR experiments used the Zettair search engine developed by the Search Engine Group (see `www.seg.rmit.edu.au`) at RMIT University.

To provide a baseline for our CLIR results, we used BabelFish to "manually" translate each Chinese query. The retrieval results are shown as runs *T-BabelFish* and *D-BabelFish*. Kraaij [2001] showed successful use of the BabelFish translation service based on Systran. We established a monolingual reference (*T-mono* and *D-mono*) by which we can measure our CLIR results. If the Chinese queries were translated perfectly, we would expect to achieve the same retrieval effectiveness as monolingual retrieval. We then tested disambiguation and OOV translation. After each stage, we have also tested query expansion. These experiments allow us to separately gauge the improvement contributed by each of our techniques. A brief description of the runs is shown in Table IX.

Table X. Effect of Disambiguation, OOV Translation, and Post-translation Query Expansion, Separately and in Combination

RunID	*MAP (% Monolingual)*	*Recall*	*P@10*
T-BabelFish	0.1696	2996	0.2983
T-BabelFish+q	0.1906	3586	0.3000
T-d	0.1459	3073	0.2534
T-dq	0.1830	3415	0.3000
T-do	0.2166	3409	0.3448
T-doq	0.2386	3652	0.3759
T-mono	0.2473	3642	0.4017
T-mono+q	0.2490	3788	0.3931
D-BabelFish	0.1226	2495	0.1828
D-BabelFish+q	0.1332	2726	0.2155
D-d	0.1381	3230	0.2569
D-dq	0.1678	3577	0.3000
D-do	0.1932	3573	0.3293
D-doq	0.2147	3794	0.3534
D-mono	0.2186	3676	0.3603
D-mono+q	0.2214	3753	0.3397

7. RESULTS AND DISCUSSION

In the previous sections we discussed the individual techniques developed as part of our dictionary-based query-translation process. In this section, we explore the combinations of these techniques and investigate the improvement contributed by each component. The results of these experiments are shown in Table X.

7.1 BabelFish and Disambiguation

In previous work [Zhang and Vines 2003] using the NTCIR-3 query set, we found that disambiguation alone was always more effective than the BabelFish baseline. This was not the case with the NTCIR-4 query set. Although the MAP for the *description* runs was slightly higher using the disambiguation technique, the MAP for the *title* runs were lower than for use of BabelFish. Examination of the translations gives the explanation. In the NTCIR-3 queries, many of the OOV terms were incorrectly translated syllable by syllable into completely wrong terms. This resulted in a number of incorrect terms being added to a relatively short query. The effect of this was often that many incorrect documents were retrieved. In the NTCIR-4 query set, we observed that in the majority of cases where BabelFish was unable to translate an OOV term, it was simply omitted from the translation. In many cases, there was still enough information in the other query terms to retrieve some relevant documents, especially at high levels of recall.

7.2 Disambiguation Combined with OOV Translation

As shown in Table IX, the OOV translation runs *T-do* and *D-do* combine translation disambiguation and OOV detection techniques. The results showed that our OOV translation technique provided an improvement of 18.4 and 15.4%, respectively, compared to the runs *T-d* and *D-d* that only applied disambiguation. This improvement was statistically significant at the 95% confidence level and emphasizes the importance of a good OOV translation technique. The rigid relevance assessment MAP values for *title* and *description* runs were 0.2166 (*T-do*) and 0.1932 (*D-do*), respectively, representing 87.6 and 88.4% of monolingual retrieval effectiveness.

7.3 Query Expansion

With the addition of query expansion as described in Section 5, results were further improved in all cases. For disambiguation only (no OOV translation) combined with query expansion, our *T-dq* run result (0.1830) was slightly lower than those obtained by applying query expansion to the BabelFish $T\text{-}BabelFish + q$ results (0.1906), while our *D-dq* run results (0.1678) were higher. More importantly, combining components of our technique—disambiguation, OOV translation, and query expansion—produced results that were statistically significantly higher than those obtained by applying query expansion to BabelFish for both *T-doq* (0.2386) and *D-doq* (0.2147) runs. The *title* run (*T-doq*) achieved 95.8% of the monolingual query expansion run ($T\text{-}mono + q$) and *description* run (*D-doq*) achieved 97.0 of the monolingual query expansion run ($D\text{-}mono + q$).

Although query expansion only gave improvements of 0.8 and 1.3% for the monolingual runs, it provided improvements of 10 and 11% for cross-lingual post-translation query expansion runs. This shows that query expansion can usefully improve retrieval effectiveness for imperfectly translated queries, although it was not helpful for monolingual retrieval.

Kwok et al. [2004] tested pre-translation query expansion in NTCIR-4 and found that it degraded the MAP using both rigid and relaxed assessment. By contrast, our post-translation query expansion technique provided an improvement of 10 and 11% of *title* and *description* runs, respectively.

7.4 Translation Quality

Our CLIR results were close to our monolingual benchmark. In comparison to other participants of this task in NTCIR-4, we obtained the second highest results at high levels of precision. The PIRCS retrieval system from City University [Kwok et al. 2004] achieved better overall results. Interestingly, their monolingual benchmark was higher then ours: 0.3175 and 0.3055 for rigid *title* and *description* runs. This suggests the underlying search engine retrieval effectiveness was superior to ours. In CLIR runs they only achieved 75 and 73% of mono-lingual retrieval effectiveness for rigid assessment [Kwok et al. 2004].

This shows that our OOV translation technique has been effective in detecting Chinese OOV terms and extracting English translations and, thus, significantly improves CLIR effectiveness. Table XI shows the translations extracted from the web. Among 66 potential Chinese OOV terms, 38 instances can be translated word by word using the translation dictionary. Of 28 Chinese OOV terms, we were able to successfully translate 20. The remaining eight cases failed for one of two reasons: first, our search technique did not return any English terms associated with some Chinese OOV terms; second, some

Table XI. NTCIR4: Extracted English Translations of Chinese OOV Terms

Query ID	*Chinese query*	*Detected Chinese OOV terms*	*Extracted English translations*	*Given English translations*
001	秋門	秋門	—	Chiutou
002	約翰走路	約翰走路	Johnnie Walker	Johnnie Walker
003	胚胎乾細胞	胚胎乾細胞	Embryonic Stem Cell	Embryonic Stem Cells
004	葛瑞菲絲 喬納 花蝴蝶	葛瑞菲絲	Griffith — —	Griffith Joyner Flojo
005	戴奧辛	戴奧辛	Dioxin	Dioxin
006	麥可喬丹	麥可喬丹	Michael Jordan	Michael Jordan
007	巴拿馬運河 卡杜條約	巴拿馬運河	Panama Canal —	Panama Canal Torrijos-Carter Treaty
008	威而鋼	威而鋼	Viagra	Viagra
009/025	南韓	南韓	South Korea	South Korea
012	黑澤明	黑澤明	Akira Kurosawa	Akira Kurosawa
013	小淵惠三		—	Keizo Obuchi
014	環境荷爾蒙	環境荷爾蒙	environmental hormone	Environmental Hormone
017	後天免疫缺乏症候群	後天免疫缺乏症候群	AIDS	AIDS
021	電子商務交易	電子商務	Electronic Commerce	Electronic Commercial Transaction
022	起亞汽車	起亞汽車	Kia Motors Corp	Kia Motors
030	動物複製技術	複製	clone	Cloning
034	東京都知事		—	Tokyo provincial governor
038	奈米科技	奈米科技	Nanotechnology	Nanotechnology
046	基因治療	基因治療	Genetic Treatment	Genetic Treatment
048	國際太空站	國際太空站	ISS	International Space Station
051	隱形戰鬥機	隱形戰鬥機 戰鬥機	stealth fighter F117	StealthFighter —
052	皇太子妃 雅子		— —	Crown Princess Masako
053	網際網路	網際網路	Internet	Internet
058	非接觸式智慧卡	非接觸式智慧卡	Contactless Smart Cards CSC	Contactless SMART Card

personnel names that relate to events are no longer topical and could not be found on the web, such as “花蝴蝶” (Flojo). As mentioned in Section 2, a system that periodically crawls the web to discover new terms would overcome this problem.

We compared our results to those of the LiveTrans system [Cheng et al. 2004]. The LiveTrans system returns a list of up to 20 alternative translations. In 21 cases, the most appropriate translation were present somewhere in the list. In only five cases was the top-ranked translation the most appropriate. Our system only returns the most appropriate translations, and thus correct in 20 instances. In three cases, our system produced the translations that might be considered more appropriate than LiveTrans. For example, for the Chinese OOV terms “胚胎幹細胞,” “基因治療,” and “非接觸式智慧卡,” our system extracted the English translations “Embryonic Stem Cell,” “Genetic Treatment” and “Contactless Smart Cards CSC”, respectively; whereas LiveTrans extracted “embryonic stem/stem cells,” “gene theraph,” and “contactless/smart card,” in each case.

8. CONCLUSIONS

We have developed a new segmentation-free technique to identify Chinese OOV terms and extract English translations, which has been demonstrated using the NTCIR-4 test collection [Kando 2004]. This technique can improve the retrieval effectiveness by 18.4% and can be used to improve Chinese segmentation accuracy. We also investigated the effects of distance factor and window size when using a hidden Markov model to provide disambiguation. Contrary to what has been noted when using mutual information techniques to provide disambiguation, we found that using a window distance factor has no benefit when combined with a hidden Markov model. We also evaluated a novel use of mutual information to select additional query terms in post-translation query expansion. Although this technique did not work for monolingual retrieval, it provided an improvement of up to 11% in cross-lingual retrieval effectiveness and allowed us to achieve up to 97% of monolingual retrieval effectiveness.

In conclusion, our OOV translation technique leads to a significant improvement in retrieval effectiveness. Post-translation query expansion can be used to improve effectiveness especially for imperfectly translated queries. In addition, the web proved to be a rich resource of potential translations for topic-specific terms.

REFERENCES

BALLESTEROS, L. AND CROFT, W. B. 1998. Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia. ACM Press New York. 64–71.

BILLERBECK, B. AND ZOBEL, J. 2004. Questioning query expansion: An examination of behaviour and parameters. In *Proceedings of the 15th Australasian Database Conference*, K. D. Schewe and H. E. Williams, Eds. Dunedin, New Zealand, pp. 69–76.

CHEN, A., JIANG, H., AND GEY, F. 2000. Combining multiple sources for short query translation in Chinese–English cross-language information retrieval. In *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, Hong Kong, China. ACM Press, New York. 17–23.

CHEN, J. AND NIE, J. Y. 2000. Parallel Web Text Mining for Cross-Language IR. In *Proceedings of RIAO-2000: Content-Based Multimedia Information Access*. CollCge de France, Paris, France. 188–192.

CHENG, P. J., TENG, J. W., CHEN, R. C., WANG, J. H., LU, W.-H., AND CHIEN, L.-F. 2004. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK. ACM Press New York. 146–153.

FEDERICO, M. AND BERTOLDI, N. 2002. Statistical cross-language information retrieval using N-Best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland. ACM Press New York. 167–174.

GAO, J., ZHOU, M., NIE, J., HE, H., AND CHEN, W. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland. ACM Press, New York. 183–190.

KANDO, N. 2004. Overview of the Fourth NTCIR Workshop. In *Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4)*. National Institute of Informatics, Tokyo, Japan. i–viii.

KISHIDA, K., HUA CHEN, K., LEE, S., KURIYAMA, K., KANDO, N., CHEN, H.-H., MYAENG, S. H., AND EGUCHI, K. 2004. Overview of CLIR Task at the Fourth NTCIR Workshop. In *Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4)*. National Institute of Informatics, Tokyo, Japan.

KRAAIJ, W. 2001. TNO at CLEF-2001: Comparing translation resources. In *Proceedings of the CLEF 2001 Workshop*. Springer, Darmstadt, Germany. 79–83.

KWOK, K. L., DINSTL, N., AND CHOI, S. 2004. NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS. In *Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4)*. National Institute of Informatics, Tokyo, Japan. 186–192.

LIN, W.-H. AND CHEN, H.-H. 2002. Backward machine transliteration by learning phonetic similarity. In *Proceedings of CoNLL-2002*, D. Roth and A. van den Bosch, Eds. Taipei, Taiwan. 139–145.

LU, W., TUNG, C., CHIEN, L., AND LEE, H. 2002. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing 2,* 1, 159–172.

MAEDA, A., SADAT, F., YOSHIKAWA, M., AND UEMURA, S. 2000. Query term disambiguation for Web cross-language information retrieval using a search engine. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, China. ACM Press, New York, 25–32.

MANDALA, R., TOKUNAGA, T., AND TANAKA, H. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA. ACM Press, New York. 191–197.

MCEWAN, C. J. A., OUNIS, I., AND RUTHVEN, I. 2002. Building bilingual dictionaries from parallel web documents. In *Proceedings of the 24th European Colloquium on Information Retrieval Research*, Glasgow, Scotland. Springer-Verlag, New York. 303–323.

MEND, H., CHEF, B., KIDNAPER, S., LEVEE, G., LO, W., OARED, D., SCONE, P., TANG, K., WANG, H., AND WANG, J. 2004. Mandarin-English Information (MEI): Investigating translingual speech retrieval. *Computer Speech and Language 18,* 2, 163–179.

MILLER, D. R., LEEK, T., AND SCHWARTZ, R. M. 1999. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA. ACM Press New York. 214–221.

MIRNA, A. 2000. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information Retrieval 2*, 1, 67–68.

NEY, H., ESSEN, U., AND KNESER, R. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language 8*, 3, 1–38.

PORTER, M. F. 1980. An algorithm for su#x stripping. *Automated Library and Information Systems 14*, 3, 130–137.

ROBERTSON, S. AND JONES, K. S. 1976. Relevance weighting of search terms. *The American Society for Information Science 27*, 3, 129–146.

RUTHVEN, I. 2003. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, Toronto, Canada. ACM Press, New York. 213–220.

SUN, J., ZHOU, M., AND GAO, J. F. 2003. A Class-based language model approach to chinese named entity identification. *International Journal of Computational Linguistics and Chinese Language Processing, 8,* 2, 1–28.

VAN RIJSBERGEN, C. J. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation 33*, 106–119.

XU, J. AND CROFT, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS) 18,* 1, 79–112.

YANG, C. C. AND LI, K. W. 2002. Mining English/Chinese Parallel documents from the world wide web. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii. ACM Press, New York. 188–192.

ZHANG, Y. AND VINES, P. 2003. Improved Cross-Language Information Retrieval via Disambiguation and Vocabulary Discovery. In *Proceedings of the 8th Australasian Document Computing Symposium*. CSIRO ICT Centre, Canberra, Australia. 3–7.

ZHANG, Y. AND VINES, P. 2004a. RMIT Chinese–English CLIR at NTCIR-4. In *Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4)*. National Institute of Informatics, Tokyo, Japan. 60–64.

ZHANG, Y. AND VINES, P. 2004b. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK. ACM Press, New York, 162–169.

Received July 2004; revised November 2004; accepted July 2005