Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval

Ying Zhang Phil Vines
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne, Australia, 3001.
{yzhang,phil}@cs.rmit.edu.au

ABSTRACT

There have been significant advances in Cross-Language Information Retrieval (CLIR) in recent years. One of the major remaining reasons that CLIR does not perform as well as monolingual retrieval is the presence of out of vocabulary (OOV) terms. Previous work has either relied on manual intervention or has only been partially successful in solving this problem. We use a method that extends earlier work in this area by augmenting this with statistical analysis, and corpus-based translation disambiguation to dynamically discover translations of OOV terms. The method can be applied to both Chinese-English and English-Chinese CLIR, correctly extracting translations of OOV terms from the Web automatically, and thus is a significant improvement on earlier work.

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

General Terms
Algorithms, Languages

Keywords
Cross-Language IR, OOV problem, query translation

1. INTRODUCTION

Successful translation of OOV terms is one of the challenges of CLIR. Particular difficulties exist in languages where there are no clearly defined boundaries between words as is the case with Chinese text. When translating from Chinese to English, a standard first step is to segment the text into words based on an existing segmentation dictionary. However where an OOV term occurs, it will not be recognized, and segmented into either smaller sequences of characters or individual characters. In this case the constituent components will usually be translated into terms in the target language that have little relationship to the original meaning. We describe a technique to detect and correct this situation via means of a probability value generated by a hidden Markov model (HMM). When translating from English to Chinese, existing systems are able to detect an English OOV term since no segmentation is required for English terms. They are either present in the translation dictionary or not. However when trying to discover appropriate Chinese translations for these terms, segmentation again comes into play, and previous work has suffered from inability to correctly identify new terms automatically. We propose a segmentation free method based on frequency and length analysis and corpus-based disambiguation, and show that it is successful in the vast majority of cases. We have concentrated on short queries as they represent typical web queries and have proved difficult to translate due to lack of context.

The structure of the paper is a as follows. In Section 2, we describe the various components of CLIR systems, existing approaches to the OOV problem, and explain the ideas behind the extensions we have developed. In Section 3, we describe our algorithm for extracting English translations of Chinese OOV terms, while in Section 4, we give our algorithm for extracting Chinese translations of English OOV terms. In Section 5, we detail our experiments and the results we obtained; and Section 6 concludes the paper.

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

2. PREVIOUS WORK

Dictionary-based query translation is a widely used approach in CLIR [3, 6, 7, 10], because of its simplicity and the increasing availability of machine readable dictionaries. However, dictionary-based translation schemes need to address three major issues; phrase identification and translation, translation ambiguity and out of vocabulary (OOV) terms.

Phrase identification refers to the identification of groups of words which have a special meaning when they co-occur that is different from the individual meanings of the words, for example *non proliferation treaty* and *cross straits*. It has been noted that correct phrase translation may improve retrieval effectiveness by up to 25%[2]. Often a given word has multiple translations. *Translation disambiguation* refers to selecting the most appropriate translation in the given context. A number of schemes have been developed to resolve the translation ambiguity problem, using techniques such as *term co-occurrence* [3, 7], *mutual information* [12] or *language modeling* [6]. *OOV* terms are typically new terms from current affairs, such as personnel names, place names and translated words. Several authors [4, 5, 1, 11] have proposed techniques to deal with OOV terms in CLIR, and we summarize these below.

While each of the above phases involve different tech-

niques, they are all inter-related. For example, if the term *cross straits* is missing from the phrase dictionary, it will be translated word by word, and the meaning will be lost. It would be much more appropriate if it could be identified as an OOV term and translated as a phrase. Similarly, OOV term translation extraction sometimes produces more than one candidate translation. However, once added to the dictionary, a sound disambiguation technique will usually be able to select the most appropriate translation. We have developed a disambiguation technique based on language modeling using a HMM [6] together with a decay factor [7], and extended it by adding the concept of window size effect. Although the focus of this paper is on the OOV problem, it must be used in conjunction with a translation disambiguation technique if sensible results are to be achieved. However when measuring the effectiveness of any OOV term translation extraction technique, it is necessary to carefully design experiments so as to isolate the improvement separately contributed by disambiguation and OOV term translation extraction.

2.1 Existing Approaches to the OOV Problem

Depending on the language, it may be possible to deduce appropriate transliterated translations automatically. For example, AbdulJaleel and Larkey describe a transliteration technique [1] that they successfully applied in English-Arabic CLIR. However the issue is more difficult in Chinese as many characters have the same sound, and many English syllables do not have equivalent sounds in Chinese, meaning that selecting the correct characters to represent a transliterated word can be problematic. Meng et. al. [11] describe a similar technique for transliteration of English names to Chinese. However, the technique appears to only be partially successful. In their example, they derive "基里斯特弗" as the translation of "*Christopher*", where the dictionary [9] and the Web gives "克里斯托弗" as the translation. Moreover, when an OOV term translation is based on meaning rather than sound, transliteration techniques will fail. For example, "*NASD*" and "*Mars rover*" cannot be transliterated. Chen et. al. used Yahoo-China search engine to find translations of OOV terms [5]. Our approach extends this idea as we explain below.

2.2 Segmentation Free Translation Extraction

Our approach stems from the observation that when new terms, foreign terms, or proper nouns are used in Chinese web text, they are sometimes accompanied by the English translation in the vicinity of the Chinese text, for example 西雅图水手队...*Seattle Mariners*. By mining the Web to collect a sufficient number of such instances for any given word and applying statistical techniques, we are then able to infer the appropriate translation with reasonable confidence. The idea of using the Web to search for translations is not new [5, 11], however our technique is segmentation-free and consequently can extract translations that were previously undetected, or only detected by manual intervention to provide correct segmentation.

It is common to find a small amount of English text in Chinese web documents, but extremely rare to find Chinese text in English web documents. We therefore rely on Chinese web documents to extract translations in both directions.

Segmentation causes difficulty in both directions, however different approaches are needed in each case. When looking for English translations of Chinese OOV terms, the Chinese OOV terms need to be appropriately detected as the first step. Normally a segmenter is used to determine Chinese word boundaries, and this information would be used to assist in the identification of the Chinese OOV term. The problem is that the Chinese OOV term we are looking for is currently unknown, and thus we have no information about how it should be segmented. In previous work [5], this problem was overcome by manual intervention to provide appropriate segmentation. Looking for a Chinese translation of an English OOV term is also not straightforward since a number of candidate Chinese character strings are normally found and must then be segmented. Previous automatic procedures [11, 4] have not been particularly successful.

Our segmentation free technique overcomes the difficulties experienced previously by researchers in this area. The details of each procedure will be explained in the following sections.

3. ENGLISH TRANSLATION EXTRACTION IN CHINESE-ENGLISH CLIR

In this section, we discuss the task of automatically extracting the English translations of Chinese OOV terms through the mining of web text. We use a four stage process to extract the English translations: Chinese OOV term detection, web text extraction, collection of co-occurrence statistics, and translation selection.

3.1 Chinese OOV Term Detection

First, we detect the Chinese OOV terms. This process builds on a translation disambiguation technique we developed previously [15]. We used a HMM to select the most appropriate translations for a sequence of query key terms. A byproduct of this technique is that the probability estimate can be used to indicate the likely occurrence of OOV terms. When a Chinese OOV term occurs, it will be incorrectly segmented into individual characters, which will then be separately translated into English terms. For example, a Japanese personnel name was segmented and translated as 北 *(north)* 野 *(limit)* 武 *(military)*. Since the correlation between these English terms is weak, the language model probability value (P_{value}) given by the HMM will be very low. When P_{value} falls below P_{min}, we conclude that the query contains OOV terms.

3.2 Extraction of Web Text

Second, we extract strings that contain the Chinese query terms and some English text from the Web.

1. When a Chinese query term is missing from the dictionary or the probability value does not reach the P_{min} [15], we run a script file that uses *Google* to fetch the top 100 *Chinese* documents using the whole Chinese query and save them into a local file using the following command:

   ```
   lynx -source "http://www.google.com.au/
   search?q=Chinese-Query&num=100&
   lr=lang_zh-TW&cr=countryTW&hl=zh-TW&
   ie=UTF-8&oe=UTF-8" > local_file
   ```

   It should be noted that a side effect of using a search engine is that only higher quality web text is returned.

This reduces the likelihood of noisy translations being collected.

2. For each returned document, only the title and the query-biased summary are extracted and saved into a local file.

3. The file is then filtered to remove HTML tags and metadata, leaving only the web text.

   For example, suppose we have a query Q which is composed of a sequence of Chinese terms $(c_1, c_2, c_3, c_4, c_5)$, and we have retrieved a series of titles and query-biased summaries of web text that contain both Chinese query substrings $C \in Q$ and English terms e, as shown in Figure 1.

.....c_2c_3**e1**.....$c_1c_2c_3c_4c_5$**e2**.....
...c_2c_3**e1**.....$c_1c_2c_3c_4c_5$**e3**.....
.....c_2c_3**e1**.....c_2c_3**e4**.....
...$c_1c_2c_3c_4c_5$**e3**.....c_2c_3**e1**.....
...c_1c_2**e2**.....c_3c_4**e1**.....

Figure 1: *Web text retrieved*

3.3 Collection of Co-occurrence Statistics

We then collect co-occurrence information from the data we obtained, in the following mannner:

1. Scan for the occurrence of English text. Where English text occurs, check the immediately proceeding Chinese text to see if it is a substring of the original Chinese query.

2. We collect the frequency of co-occurrence of the English text and all Chinese query substrings that appear immediately before the English text.

For each English term e_i with frequency $f(e_i)$ we obtained a group of associated Chinese query substrings C_{ij} with length $|C_{ij}|$ and co-occurrence frequency $f(e_i, C_{ij})$. Extending the example from in Figure 1, this information is summarized in Table 1.

e_i	$f(e_i)$	C_{ij}	$\|C_{ij}\|$	$f(e_i, C_{ij})$
e_1	5	c_2c_3	2	4
		c_3c_4	2	1
e_2	2	$c_1c_2c_3c_4c_5$	5	1
		c_1c_2	2	1
e_3	2	$c_1c_2c_3c_4c_5$	5	2
e_4	1	c_2c_3	2	1

Table 1: *The frequency of co-occurrence of English terms and Chinese query substrings*

3.4 Translation Selection

We then select the most appropriate translation as follows:

1. Firstly search for *longest* Chinese substring C_t:

   (a) Search for the Chinese query substrings $C_{targets}$, where $|C_{targets}| = max(|C_{ij}|)$.

   (b) Extract the English term e_t and the Chinese query substring C_t, where $f(e_t, C_t) = max(f(e_i, C_{targets}))$.

   (c) Add (C_t, e_t) into the translation dictionary.

2. Then search for the English term $e_{t'}$ with the *highest frequency*:

   (a) Search for the English terms $e_{targets}$, where $f(e_{targets}) = max(f(e_i))$.

   (b) Extract the English term $e_{t'}$ and the Chinese query substring $C_{t'}$, where $f(e_{t'}, C_{t'}) = max(f(e_{targets}, C_{ij}))$.

   (c) if $C_{t'} \neq C_t$ and $e_{t'} \neq e_t$, add $(C_{t'}, e_{t'})$ into the translation dictionary.

In the example in Table 1 above, two translation pairs (c_2c_3, e_1) and ($c_1c_2c_3c_4c_5$, e_3) are extracted and added into the translation dictionary. We have extracted at most two translation pairs, which proved to be ample for short queries; and in fact, in most cases, only one translation pair was extracted.

4. CHINESE TRANSLATION EXTRACTION IN ENGLISH-CHINESE CLIR

Our work builds on previous work in this area, in particular that of Chen and Gey [4]. In their translation extraction process, each English OOV term is submitted as a query to Yahoo!Chinese in traditional Chinese (Big5 encoding). The top 200 result entries are then segmented into words using a dictionary-based longest matching method. For each line containing the English query word or phrase, they consider the five Chinese "words" immediately before and after the English word or phrase, and use a weighting scheme to select the top m of these as the best translation, where m is the number of English terms.

Since the Chinese word being searched for is currently unknown, there is no information as to how it should be segmented, and thus segmentation errors may occur, leading to an incorrect translation extraction. When this occurs the retrieval effectiveness is inevitably substantially degraded. In an example provided [4], 7 out of 17 extracted translations exhibited this problem. Another problem is that sometimes the correct Chinese translation may occur some distance from the English OOV term, for example " ... 南沙群岛，我方称为 the Nansha Islands，而西方则称为 the Spratley Islands ...". In contrast to Chen and Gey [4], our technique uses a larger window size and does not discriminate against terms that occur some distance from the English OOV term. It does not rely on a prior segmentation and is based on the consideration of every possible Chinese substring occurring adjacent to the English OOV term. In the experiments we have conducted we have found this procedure to be free from segmentation error in translation extraction. We describe below our process to automatically extract the Chinese translations of English OOV terms from the Web.

4.1 Extraction of Web Text

First, we extract strings that contain the English OOV term and some Chinese text from the Web.

1. We use *Google* to fetch the top 100 *Chinese* documents with the English OOV term e_{oov} as the query. For each returned document, only the title and the query-biased summary are extracted and saved into a local file using the following command:

```
lynx -source "http://www.google.com.au/
search?q=English-OOV-Term&num=100&
lr=lang_zh-CN&cr=countryCN&hl=zh-CN"
> local_file
```

2. The file is then filtered to remove HTML tags and metadata, leaving only the web text.

4.2 Collection of Co-occurrence Statistics

We then collect co-occurrence information from the data.

1. We scan for the occurrence of e_{oov} and accumulate the frequency f_{oov}. Where e_{oov} occurs, we collect *twenty* Chinese characters immediately before as S_{left} and *twenty* Chinese characters immediately after as S_{right}.

2. Since we want to use a process that does not rely on segmentation, we start by considering all substrings in S_{left} and S_{right}, and collecting the frequency f_n and the length $|s_n|$ of each Chinese substring.

3. We then rank the substrings based on the likelihood of being the correct translation. We use the ranking function r to select only the top ten strings for further consideration. Generally we prefer substrings that occur more frequently over those that occur less frequently, and prefer longer substrings over shorter ones. However the natural distribution is that shorter strings occur more frequently than longer ones. The ranking function we have developed is

$$r = \alpha \times \frac{|s_n|}{L} + (1 - \alpha) \times \frac{f_n}{f_{oov}}$$

Where L is the maximum length of the substring. From our experiments, we determined that $\alpha = 0.25$ provides the best combination of frequency and length, a value that proved to be robust across our experiments. An example is given in Table 2, which we refer to in the remaining steps.

s_n	$\|s_n\|$	$f(s_n)$	r
s_1	4	13	0.598529
s_2	4	11	0.510294
s_3	8	9	0.447059
s_4	6	9	0.434559
s_5	6	9	0.434559
s_6	4	9	0.422059
s_7	4	9	0.422059
s_8	4	7	0.333824
s_9	4	7	0.333824
s_{10}	16	5	0.320588

Table 2: *The frequency and length of Chinese substrings*

4.3 Translation Selection

From these top ten substrings in Table 2, we select the most appropriate translations in the following manner:

1. We select the two substrings s_{10} and s_3 from Table 2 with the *longest length*. In the event of a tie we use frequency to discriminate. Provided that one is not a substring of the other, both s_{10} and s_3 are added into the translation candidate set T as shown in Table 3.

2. We select the two substrings s_1 and s_2 from Table 2 with the *highest frequency*. In the event of a tie we use length to discriminate. Provided that one is not a substring of the other, both s_1 and s_2 are added into the T as shown in Table 3.

s_n	$\|s_n\|$	$f(s_n)$
s_{10}	16	5
s_3	8	9
s_1	4	13
s_2	4	11

Table 3: *Translation candidate set*

3. From the translation candidate set T, we exclude any substring that is already in the translation dictionary or does not occur in the document collection.

   If we have added more than one translation to the translation dictionary, we use our disambiguation technique to select the most appropriate alternative in the given context.

5. EXPERIMENTS AND RESULTS

We wish to explore issues related to querying in each direction, and therefore we have conducted CLIR experiments on both Chinese-English and English-Chinese query translation.

5.1 Chinese-English CLIR

In this section, we describe the experimental setup for retrieving English documents using Chinese queries.

Document Collection and Queries
We used the English document collection from the NTCIR-4 [1] CLIR task and the associated 50 Chinese training topics. A Chinese topic contains four parts: *title*, *description*, *narrative* and *key words* relevant to whole topic. The *titles* of the Chinese topics were translated and used as queries to retrieve the documents from the English document collection. The average length of the titles is 3.3 terms which approximates the average length of short web queries.

Chinese-English Dictionaries
We used two dictionaries in our experiments: ce3 from Linguistic Data Consortium[2] and CEDICT Chinese-English dictionary[3] to translate Chinese queries into English. Using these two dictionaries, we were able to find at least one English translation for each term in a given Chinese query for 74% of the queries. However, 54% of translated queries contain inappropriate or wrong translations.

Pre-processing
English stop words were removed from the English document collection, and the Porter stemmer [13] was used to reduce words to stems. To obtain optimal segmentation accuracy, we combined two segmenters. The first one is compiled by Erik Peterson[4]; the second, Autotag, provided by Chinese Knowledge Information Processing Group (CKIP) from Taiwan. The Chinese stop list was manually selected from

[1] http://research.nii.ac.jp/ntcir-ws4/
[2] http:// www.ldc.upenn.edu/
[3] http://www.mandarintools.com/cedict.html
[4] http://www.mandarintools.com/segmenter.html

the statistical results we obtained from the given Chinese topic file. Each Chinese query was segmented into words using the segmenters as described above, the Chinese stop words were then removed from each Chinese query. Our CLIR experiments used the *Lucy* search engine developed by the Search Engine Group[5] at RMIT University.

Experiment Design
The following three runs were performed in our Chinese to English CLIR experiments:

1. RUN1: To provide a baseline for our CLIR results, we used BableFish to "manually" translate each Chinese query. Kraaij [8] showed successful use of the widely used BableFish [6] translation service based on Systran.

2. RUN2: Chinese queries were translated using a dictionary look-up and the disambiguation technique previously developed [15].

3. RUN3: Chinese queries were translated using the disambiguation technique combined with the English translation extraction technique described in Section 3.

The latter two experiments allow us to isolate the improvement contributed by each of the techniques.

Experimental Results and Discussion
As the relevance judgements for this collection are as yet unavailable, we were only able to evaluate the translation quality. We define successful translation as each term in the given query being correctly translated. An inappropriate translation occurs when a translation of a query term has been found that relates to the original meaning, but is inappropriate in the given context. For example, "train" was selected in Query011, since the more appropriate term "railway" was not in the dictionary. A wrong translation occurs when a query term has been translated into a term which has no relation to the original meaning. Such wrong translations never return relevant documents. Figure 2 shows the comparison of translation quality of the three runs we described earlier. When the disambiguation technique [15] was applied, the number of successful translations increased to 23. When we combined our English translation extraction technique with the disambiguation technique, 30 queries were successfully translated.

Our English translation extraction technique yielded a number of translations of previously untranslated terms, as well as correct translation where previously only inappropriate ones existed (see Table 5). For example, we were able to replace "north limit military director film" with "Takeshi kitano director film" in Query017 and "non national boundary doctor" with "Medecins Sans Frontieres" in Query024. Interestingly, our techniques sometimes produced translations that might be considered more correct than the provided translation. For example, our system extracted the English term "*La Nina*", which is arguably more correct than the given translation "*anti-El Nino*". Of the 50 queries from the training topics 8 were found to contain Chinese OOV terms. The translations that our technique found are shown in Table 5, it can be seen that 7 out of 8 are correct.

Although this result is quite encouraging, the final data set is very small. In order to test the ability of our system

[5]http://www.seg.rmit.edu.au/
[6]http://world.altavista.com/

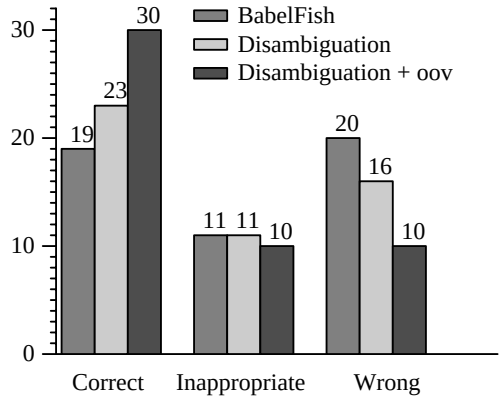

Figure 2: *Translation Quality Comparison*

to extract correct translations of Chinese OOV terms, we decided to use the NTCIR-4 formal run topics to test the robustness of our technique. The *title* of each topic is presented a list of comma separated query key terms. From the 60 titles, we found 25 Chinese OOV terms. Of these, our technique was able to extract English translations for 18 terms (72%). Each of these was the same as or equivalent to the provided English translation, as shown in Table 6. While not quite as good as the results we obtained for the NTCIR-3 collection, it clearly demonstrates that in the majority of cases, we are able to automatically extract the Chinese OOV terms and corresponding English translations without having to rely on manual segmentation.

5.2 English-Chinese CLIR

In this section, we describe the experimental setup for retrieving Chinese documents using English queries. The aim of our work is to find appropriate Chinese translations of English OOV terms.

Document Collection and Queries
The test collection used in this task is the TREC-5 and TREC-6 Chinese collection. We used dictionary-based segmentation with greedy-parsing to segment the document collection. There are 54 English topics of TREC-5 and TREC-6 Chinese track. Each topic consists of three sections: *title, description, and narrative.* Not every query contains English OOV terms and such queries are obviously not effected by the OOV problem. As we are particularly interested in OOV problem, we have only selected the queries containing English OOV terms. In order to mimic typical web queries, we decided to use only the *title* of topics as queries. To maximize the number of test queries, we augmented some otherwise OOV free topics with English OOV terms from the *description* and *narrative* sections (see Table 7). This provided a total of 14 queries.

English-Chinese Dictionaries
We compiled a translation dictionary for our experiments using three dictionaries: ec2 and ec2 from Linguistic Data Consortium, and CEDICT Chinese-English dictionary. Our translation dictionary contains 128, 527 entries including 19, 081 multi word phrases that were used for phrase identification and translation.

Experiment Design
The goal of our experiments was to measure the ability of our

technique to find appropriate Chinese translations of English OOV terms. Since the relevance judgements for this collection are available, rather than directly judging translation quality we instead measure the difference in retrieval effectiveness of the translated queries. In our first run we used the given Chinese queries without any of the Chinese equivalents of the English OOV terms (*C-C*), and used this to compare the performance of the translated English queries without any English OOV terms (*E-C*). This allowed us to test the basic effectiveness of our system using the translation disambiguation technique, without regard to the OOV problem. We then manually added the Chinese equivalents of the English OOV terms to the Chinese queries (*CO-C*), and used this as a further baseline to test the ability of our technique to automatically find the appropriate Chinese translations of the English OOV terms. This was done by adding the English OOV terms to the English queries and using our system to translate and then retrieve Chinese documents (*EO-C*). Our English-Chinese CLIR experiments used the *MG* [14] search engine.

Experimental Results and Discussion

Table 4 shows a comparison of the recall precision values for the English-Chinese CLIR experimental results. Without any English OOV terms, our translated queries achieved 86.7% of the monolingual result. The underlying performance of our query translation system was effected by the following two factors: first, some of the English query translations provided by the TREC organizers did not precisely parallel the original Chinese queries, for example: in CH47, "菲律宾(Philippines)" is missing from the English query; in CH21, there is no exact English equivalent of "回归中国(return to China)" given in the English query; second, some translations provided by the translation dictionary are inappropriate in the given context, for example, in CH28, the English term "*cellular phone*" is translated into "汽车电话", where the given Chinese equivalent is "移动电话"; additionally, in CH47, "*impact*" is translated into " 冲击", where the given Chinese equivalents is "后果". When the translated English OOV terms were added, we achieved 77.1% of the monolingual result. Besides the two factors we discussed above, we were not able to automatically find the most appropriate Chinese translations for every English OOV term. We failed to find the translation of "*Sino -Vietnamese*" in CH46, and thus did not obtain any improvement in retrieval effectiveness. In CH48, the correct translation of "*Kuwaiti*" was lost because we did not consider any translation that is already in the translation dictionary. This is a shortcoming of our extraction algorithm, which assumes that no English OOV term shares the same Chinese translation with any English term in the translation dictionary. We are presently working to overcome this problem. In CH49, we extracted an inappropriate translation of "*START treaty*", because there is currently no single dominant accepted Chinese translation for this term; "削减战略武器条约", "削减进攻性战略武器条约" and "限制战略武器条约" being used interchangeably.

It can be seen in Table 4 that successful translation of English OOV terms results in a 80% improvement in average recall precision. Obviously, this is because we have specifically selected the queries known to contain English OOV terms. While the improvement in a heterogenous set of queries would be more modest, the results show that (a) not being able to translate OOV terms leads to significant loss in retrieval performance for such queries; and (b) our technique is effective in automatically finding appropriate translations of English OOV terms. Although we were only able to automatically find appropriate translations for 88% (22 out of 25) of English OOV terms, this is a considerable improvement on previous work in this area [5], where 72% (8 out 11) translations required manual segmentation correction in order to be processed correctly.

Recall	*C-C*	*E-C*	*CO-C*	*EO-C*
0.00	0.4850	0.4639	0.6261	0.5698
0.10	0.2839	0.2361	0.4992	0.3928
0.20	0.2208	0.1953	0.4160	0.3561
0.30	0.1828	0.1694	0.3431	0.3030
0.40	0.1411	0.1369	0.2985	0.2561
0.50	0.1104	0.1055	0.2625	0.2121
0.60	0.0824	0.0718	0.2255	0.1797
0.70	0.0624	0.0409	0.1933	0.1291
0.80	0.0339	0.0276	0.1293	0.0614
0.90	0.0089	0.0077	0.0828	0.0280
1.00	0.0000	0.0000	0.0082	0.0057
Avg.P	0.1284	0.1113	0.2610	0.2013
% Mono	-	86.68	-	77.13

Table 4: *English-Chinese CLIR results*

As was the case with our Chinese-English experiments, the final sample size was somewhat small. To further test the robustness of our technique, we collected 50 English OOV terms from news web sites and applied our Chinese translation extraction technique. However, since our technique requires a corpus to provide disambiguation, we were not able to carry out the final step of our translation extraction procedure, namely using the corpus to select the most appropriate translation from a set of candidate translations. Table 8 shows set of candidate translations for each English OOV term. It can be seen that 10 of the English OOV terms have a single correct Chinese translation. A further 35 English OOV terms have at least one correct translation in the candidate set, while our technique failed to find correct translations in 5 instances. In our experiments using the TREC Chinese data, we arrived at similar situation at the penultimate stage, but with the aid of corpus-based disambiguation were able to select the most appropriate translation in the final phase.

We were concerned that the English OOV terms extracted from Chinese web pages might only pertain specifically to Chinese news. However, we note that we were able to find Chinese translations for 96% of English OOV terms from TREC-5 and TREC-6. This gives us some confidence that the technique should at least have general applicability to news based queries. We are currently investigating how well the techniques performs in other vocabulary domains.

6. CONCLUSION

We have looked in detail at the OOV problem as it applies to Chinese-English and English-Chinese CLIR. We have described a new technique to detect potential Chinese OOV terms based on a HMM and term co-occurrence. We have also described improved ways to extract the translation of OOV terms from the Web in a way that does not rely on prior segmentation. We have tested these techniques on sev-

eral collections and a set of terms from news articles and found them to be robust and provide a substantial improvement in OOV term translation quality. Interestingly, although the Web is constantly changing, we were able to find most OOV terms, many of which related to news events up to 10 years ago.

7. REFERENCES

[1] N. AbdulJaleel and L. Larkey. Statistical transliteration for english-arabic cross language information retrieval. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 139–146, 2003.

[2] L. Ballesteros and W. B. Croft. Dictionary-based Methds for Cross-Lingual Information Retrieval. In *Proc. International Conference on Database and Expert Systems Applications*, pages 791–801, 1996.

[3] L. Ballesteros and W. B. Croft. Resolving Ambiguity for Cross-Language Retrieval. In *Research and Development in Information Retrieval*, pages 64–71, 1998.

[4] A. Chen and F. Gey. Experiments on Cross-language and Patent Retrieval at NTCIR-3 Workshop. In *Proceedings of the 3rd NTCIR Workshop*, Japan, 2003.

[5] A. Chen, H. Jiang, and F. Gey. 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*, pages 17–23, 2000.

[6] M. Federico and N. Bertoldi. 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*, pages 167–174, Tampere, Finland, 2002.

[7] J. Gao, M. Zhou, J. Nie, H. He, and W. Chen. 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*, pages 183–190, Tampere, Finland, 2002.

[8] W. Kraaij. TNO at CLEF-2001: Comparing Translation Resources. In *Proceedings of the CLEF-2001*, pages 79–83, 2001.

[9] S. Kuai, Q. Lu, and L. Yang, editors. *A New English-Chinese Dictionary*. Joint Publishing Company, HongKong, 1st edition, 1984.

[10] K. L. Kwok. Exploiting a Chinese-English Bilingual Wordlist for English-Chinese Cross Language Information Retrieval. In *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, pages 173–179, 2000.

[11] H. Meng, B. Chen, S. Khudanpur, G. Levow, W. Lo, D. Oard, P. Schone, K. Tang, H. Wang, and J. Wang. Mandarin-English Information (MEI): investigating translingual speech retrieval. In *Computer Speech and Language*, 2003.

[12] A. Mirna. Using Statistical Term Similarity for Sense Disambiguation in Cross -language Information Retrieval. *Information Retrieval*, 2(1):67–68, 2000.

[13] M. F. Porter. An Algorithm for Suffix Stripping. *Program(Automated Library and Information Systems)*, 14(3):130–137, 1980.

[14] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, 2nd edition, 1999.

[15] Y. Zhang and P. Vinese. Improved Cross-Language Information Retrieval via Disambiguation and Vocabulary Discovery. In *Proceedings of the 8th Australasian Document Computing Symposium*, pages 3–7, 2003.

	English OOV terms	*Chinese Translation Candidate Set*
1	Pervez Musharraf	穆沙拉夫/ 总统佩尔韦兹穆沙拉夫/
2	Shiite Muslim	在联合国安理会讲话全/ 联合国安理会讲话全文/
3	Ali al-Sistani	高称号赛义德阿里阿尔/ 最高称号赛义德阿里阿/
4	Paul Bremer	拉克/新浪教育新浪网/
5	Avian flu	禽流感/ 对抗香港禽流感的疫苗/
6	Hambali	汉巴里/ 在东南亚地区组织恐怖/ 已经被捕他们与另一名/
7	Mad Cow	疯牛/牛病/疯牛病/ 榜因为当年在英国出现/
8	Nintendo	任天堂/
9	Carlsberg	嘉士伯/ 节点类型节点类型例子/ 的节点类型节点类型例/
10	Credit Lyonnais	信贷银行/里昂信贷银行/
11	Osama bin Laden	拉登/本拉登/
12	John Howard	华德/澳大利亚总理/
13	Saddam Hussein	萨达姆/萨达姆被/
14	Kofi Annan	合国/联合国秘书长/ 联合国秘书/
15	Hezbollah	真主党/ 及黎巴嫩武装恐怖组织/ 以及黎巴嫩武装恐怖组/
16	NASD	证券交易商协会/
17	PricewaterhouseCoopers	永道/普华永道/华永道
18	SARS	非典/
19	Matrix Reloaded	客帝/客帝国/重装上阵
20	Martha Stewart	尔特/玛莎斯图尔特
21	Tour de France	环法/环法自行车赛/
22	NMD	美国国家导弹防御系统/
23	Mars rover	火星探测器
24	Blaster	冲击波/病毒专杀工具/
25	Forest Gump	阿甘/欧美电影音乐/ 阿甘正传
26	DJIA	道琼斯工业/
27	NAS	网络存储器/
28	Land Rover	陆虎/与陆虎/
29	Kim Clijsters	克里/克里斯特尔斯/ 克里斯特尔/
30	Likud Party	利库德/利库德集团/
31	Lord of the Rings	魔戒/ 魔戒首部曲魔戒现身/
32	Starbucks	星巴克/ 星巴克咖啡/ 还是紫藤庐/
33	Enron	安然公司/公司破产/
34	Abdullah Gul	拉居尔/ 外长阿卜杜拉居尔/ 总理兼外长阿卜杜拉居/
35	Olympus	相机/奥林巴斯/
36	Cappuccino	亚轩/卡布奇诺/萧亚轩/
37	Espresso	意大利特浓咖啡用小杯/ 大利特浓咖啡用小杯品/
38	Mohammad Khatami	哈塔米/穆罕默德/
39	Finding Nemo	海底总动员/
40	Arnold Schwarzenegger	瓦辛格/阿诺德施瓦辛格/
41	Rupert Murdoch	默多克/特默多克/新闻集团/
42	Lancome	兰蔻/
43	TAFE	技术与继续教育学院/
44	Logitech	罗技/鼠标/简体中文版/
45	PDP	等离/等离子/等离子显示器/
46	Aopen	建基/主板/准系统/
47	ViewSonic	优派/ 优派系列显示器/ 派系列显示器/
48	Ariel Sharon	以色列总理沙龙/
49	Donald H. Rumsfeld	拉姆斯/拉姆斯菲尔/
59	N-Gage	诺基亚/游戏手机/

Table 8: *Extracted Chinese translations of English OOV terms*

Query ID	*Chinese query*	*Chinese OOV Terms*	*Extracted English Translations*	*Given English Translations*
003	大学学术追求卓越 发展计划	大学学术追求卓越 发展计划	Program for Promoting Academic Excellence of Universities	Program for Promoting Academic Excellence of Universities
007	华航失事	华航	CI611	China Airlines
009	中新一号卫星	中新一号卫星	ST1	ST1
010	反圣婴现象	反圣婴现象 圣婴	La Nina El Nino	Anti-El Nino El Nino
016	佐佐木主浩将 加入西雅图水手队	佐佐木主浩 西雅图水手队	Kazuhiro Sasaki Seattle Mariners	Kazuhiro Sasaki Seattle Mariners
017	北野武导演的电影	北野武	Takeshi Kitano	Takeshi Kitano
020	日产与雷诺汽车公司 资本结合	日产 雷诺	NISSAN RENAULT	Nissan Mortor Company Renault
024	无国界医生	无国界医生	MEDECINS SANS FRONTIERES	Medecins Sans Frontieres

Table 5: *NTCIR3: Extracted English translations of Chinese OOV terms*

Query ID	*Chinese query*	*Chinese OOV Terms*	*Extracted English Translations*	*Given English Translations*
001	秋门		—	Chiutou
002	约翰走路	约翰走路	Johnnie Walker	Johnnie Walker
003	胚胎干细胞	胚胎干细胞	Embryonic Stem Cell	Embryonic Stem Cells
004	葛瑞菲斯 乔纳 花蝴蝶	葛瑞菲斯	Griffith — —	Griffith Joyner Flojo
005	戴奥辛	戴奥辛	Dioxin	Dioxin
006	麦可乔丹	麦可乔丹	Michael Jordan	Michael Jordan
007	巴拿马运河 卡杜条约	巴拿马运河	Panama Canal —	Panama Canal Torrijos-Carter Treaty
008	威尔钢	威尔钢	Viagra	Viagra
012	黑泽明	黑泽明	Akira Kurosawa	Akira Kurosawa
013	小渊惠三		—	Keizo Obuchi
014	环境荷尔蒙	环境荷尔蒙	environmental hormone	Environmental Hormone
021	电子商务交易	电子商务	Electronic Commerce	Electronic Commercial Transaction
022	起亚汽车	起亚汽车	Kia Motors Corp	Kia Motors
030	动物复制技术	复制	clone	Cloning
034	东京都知事		—	Tokyo provincial governor
038	奈米技术	奈米技术	Nanotechnology	Nanotechnology
046	基因治疗	基因	gene	Genetic Treatment
048	国际太空站	国际太空站	ISS	International Space Station
051	隐形战斗机	隐形战斗机 战斗机	stealth fighter F117	Stealth Fighter —-
048	非接触式智慧卡	非接触式智慧卡	Contactless Smart Cards CSC	Contactless SMART Card
052	皇太子妃 雅子		— —	Crown Princess Masako

Table 6: *NTCIR4: Extracted English translations of Chinese OOV terms*

	Topic Number	*English OOV terms*	*Extracted Chinese Translations*	*Given Chinese Translations*
1	*2*	reunification	和平统一	统一
2	*2*	cross-strait	海峡两岸关系	两岸
3	*3*	Daya Bay	大亚湾	大亚湾
4	*3*	Qinshan	秦山	秦山
5	*7*	Dongsha Islands	东沙群岛	东沙群岛
6	*7*	Xisha Islands	西沙群岛	西沙群岛
7	*7*	Spratly Islands	南沙群岛	南沙群岛
8	*8*	Richter	里氏	芮氏
9	*11*	Peace-keeping	维和部队	维和
10	*14*	HIV	艾滋病毒	No translation
11	*21*	Peng Dingkang	彭定康	彭定康
12	*21*	Reunification	和平统一	统一
13	*28*	PSDN	分组交换	分组交换网
14	*31*	Castro	卡斯特罗	卡斯特罗
15	*42*	Liaohe River	辽河	辽河
16	*42*	Haihe River	海河	海河
17	*42*	Huaihe River	淮河	淮河
18	*42*	Songhua River	松花江	松花江
19	*42*	Pearl River	珠江	珠江
20	*46*	Sino-Vietnamese	Not Found	中越
21	*47*	Pinatubo	皮纳图博火山	皮纳图博
22	*47*	Subic Bay	苏比克湾	苏比克湾
23	*48*	Kuwaiti	科威特第纳尔	科威特
24	*49*	Non-Proliferation Treaty	不扩散核武器条约	不扩散核武器条约
25	*49*	START	战略武器条约	消减战略武器条约

Table 7: *TREC-5 and TREC-6: Extracted Chinese translations of English OOV terms*