Combining Multiple Sources for Short Query Translation in Chinese-English Cross-Language Information Retrieval

Aitao Chen, Hailing Jiang and Fredric Gey
School of Information Management and Systems
University of California at Berkeley, CA 94720-4600, USA
{aitao,hjiang1}@sims.berkeley.edu, gey@ucdata.berkeley.edu

Abstract

In this paper, we examine various factors that affect the retrieval performance of Chinese-English cross-language retrieval. The factors include segmentation dictionary coverage, segmentation algorithm, transfer dictionary coverage, transfer dictionary quality, and translation disambiguation. The paper introduces an idea of recovering the original English names for the transliterated Chinese words, mainly the proper names, using search engine. We used two transfer dictionaries and a Chinese search engine to translate short Chinese queries into English. The majority of the Chinese words were translated into English, but the overall precision of the Chinese to English cross-language retrieval is only about 56% of the overall precision for the monolingual retrieval. [1]

1 Introduction

Information retrieval is the task of finding the documents in document collections that are likely to satisfy the users' information needs expressed in queries. When the natural language in which documents are written differs from the language in which queries are posed, the task of retrieval is called cross-language information retrieval (CLIR). A common approach to cross-language retrieval is to translate queries into the document language using a bilingual dictionary and then perform monolingual retrieval. Monolingual information retrieval is itself a difficult task, and the language difference in CLIR adds an additional dimension to the complexity of information retrieval task. Translating users' queries into document language requires bilingual resources such as machine translation systems, bilingual dictionaries, parallel bilingual corpora, and so on. The effectiveness of CLIR, usually measured in recall and precision, depends on many factors, such as the effectiveness of the monolingual retrieval engine, the coverage of the bilingual dictionaries used to translate queries, the quality of the translations, the translation disambiguation strategy, and other factors.

We present in this paper experimental results from combining multiple sources for translating short queries from English into Chinese and examine in some detail the issues in Chinese to English cross-language retrieval. Cross-language information retrieval involving Chinese has one more factor, *word segmentation*, that may affect the effectiveness of retrieval compared to CLIR dealing with a pair of European languages.

Section 2 describes the test sets for our CLIR experiments. Section 3 presents the monolingual retrieval experiment results. Section 4 describes the Chinese to English cross-language retrieval approach, the sources used to translated short English queries and the experiment results. In section 5 we examine in some detail the factors that may have degraded the performance of the Chinese to English cross-language retrieval. Section 6 briefly notes related work on Chinese to English retrieval, and section 7 provides our conclusions.

2 Test Sets

The TREC topics 301-450 used in the ad hoc tasks in TREC-6 [11], TREC-7 [13] and TREC-8 [12] were manually translated from English into Chinese by two Chinese native speakers. The original English topics contain three parts: title, description and narrative. The titles typically consist of one or two concept terms. The descriptions are about one sentence long, stating the users' information need. The narrative of a topic gives additional information about what documents should be considered relevant and/or what documents should not be considered relevant. The three parts are explicitly marked. The description and narrative are absent in typical user queries posed to retrieval systems such as online library catalog systems and Internet search engines.

Only the title and description fields of the TREC topics 301-450 were manually translated from English into Chinese. In the experiments, only the title translations were used as queries. And the performance on English monolingual and on Chinese to English cross-language retrieval were based on the title queries only.

The average length of the 150 titles is 2.41 terms, close

topic number	original English title	Chinese translation
313	Magnetic Levitation-Maglev	电磁悬浮 - 磁悬浮
335	Adoptive Biological Parents	收养以及亲生父母
349	Metabolism	新陈代谢
357	territorial waters dispute	海域争端
367	piracy	海盗行为
378	euro opposition	反对欧元
417	creativity	创造力
434	Estonia, economy	爱沙尼亚, 经济
436	railway accidents	火车事故
446	tourists, violence	游客, 暴力

Table 1: Ten randomly selected topics.

recall level	MONO	CLIR1 LDC	CLIR2 LDC+ Kingsoft	CLIR3 LDC+ KingSoft+ Yahoo
at 0.00	0.6658	0.3868	0.4369	0.4578
at 0.10	0.4533	0.2192	0.2492	0.2706
at 0.20	0.3639	0.1628	0.1828	0.2021
at 0.30	0.2929	0.1243	0.1432	0.1617
at 0.40	0.2268	0.0934	0.1082	0.1246
at 0.50	0.1850	0.0717	0.0794	0.0912
at 0.60	0.1415	0.0538	0.0592	0.0681
at 0.70	0.1025	0.0431	0.0466	0.0519
at 0.80	0.0701	0.0264	0.0279	0.0320
at 0.90	0.0387	0.0192	0.0196	0.0213
at 1.00	0.0190	0.0117	0.0122	0.0129
average precision	0.2108	0.0939	0.1069	0.1183
percent of monolingual		44.5%	50.7%	56.1%
relevant retrieved	7019	4496	4888	4973

Table 2: English Monolingual and Chinese-English cross-language retrieval performance.

to the average query length found in the transaction logs of 51,473 queries posed by 18,113 users of Excite [4].

The 150 TREC topics cover a variety of subjects such as agriculture, law, technology, medicine, international relations, policy, philosophy, tourism, weather, and so on. Ten randomly selected topics with Chinese translations are presented in table 1.

Our test document set used in the experiments consists of the TREC disks 4&5 without the *Congressional Record* collection. The test collection contains 528,155 documents with an average length of 497.9 words [14]. The document collections used in the experiments include articles in the *Financial Times* published from 1991 to 1994, the *Federal Register* in 1994, the *Foreign Broadcast Information Services* (FBIS) and, the *Los Angeles Times* between 1989 and 1990. The text in the four collections is close to two gigabytes. A total of 13,692 documents in the test document set are known to be relevant to the 150 TREC topics.

We did one English monolingual retrieval run using the titles in the original topics and the test document set. The retrieval algorithm used in our experiments (both monolingual and cross-language) is a logistic regression based document ranking algorithm developed at Berkeley [3].

The indexing procedure used in processing topic titles and documents consists of tokenization, stopword removal, and stemming. A token or term is defined in our experiments as consisting of alphabetical letters only and all capital letters are changed to lower case. Our stoplist consists of about 600 words. A Porter stemmer was used to reduce words to stems [10].

For Chinese to English cross-language runs, the manually translated Chinese queries were automatically translated back into English using multiple sources to be described below.

3 Monolingual Performance

The retrieval performance of the English monolingual run using title only is presented in the MONO column in table 2. The overall precision is about 0.21 and the overall recall is about 0.51. This will be the baseline for comparing Chinese to English cross-language retrieval to monolingual retrieval. In our experiments reported in this paper, we did not employ the pseudo relevance feedback technique. The performance for all retrieval runs are based on the results of the initial search without query expansion.

Beside the document ranking algorithm used in the retrieval, the shortness of the titles and the non-specificity of some titles could also degrade the overall performance of the monolingual retrieval. With no context, it is difficult to judge the real information needs of the users who posed queries like *piracy*, *creativity*, *Women in Parliaments*, *World Court*, *mainstreaming*, *alternative medicine*, *tourism*, and *suicides* which are among the 150 topics used in this study. For example, *piracy* could mean the act of robbery on the high seas or the unauthorized use of copyrighted products such as software. The topic *alternative medicine* is vague when examined out of context.

4 Cross-Language IR

There are a number of ways to perform the task of cross-language information retrieval in which a query posed in one language is searched against a collection of documents written in a different language. Oard and Diekema provide a recent survey on cross-language information retrieval in [8]. A retrieval method based on matching a query in one language against documents in a different language would fail when there are no cognates between this language pair (e.g., Chinese and English). For matching-based retrieval algorithms to work, both the documents and queries need to be expressed in the same language or conceptual space, as in latent semantic indexing. A common approach to cross-language information retrieval is to couple translation with monolingual information retrieval. A user's query can be translated into the document

language or the document collection can be translated into the language with which the users feel most comfortable. When a translation system is available, it can be used to translate either query or documents. When no such resource is available, bilingual dictionaries or wordlists, if available, can be used to translate queries in word-by-word or in phrases. Alternatively, parallel or comparable bilingual corpora can be used to mine a bilingual wordlist for query translation.

The steps involved in our Chinese to English cross-language retrieval include segmentation of Chinese queries, translation of the Chinese query terms, selection of translation words when multiple translations exist for a Chinese term, monolingual retrieval of the translations against the English test collection described above. The Chinese queries are the manually translated English titles as mentioned above.

We applied a dictionary-based longest matching method to split a Chinese query into terms. We will refer to the dictionary used in word segmentation as the segmentation dictionary, and a bilingual dictionary used to translate Chinese into English as the transfer dictionary. Our segmentation dictionary contains about 240,000 entries, including characters, words, phrases, and proper names. Two main contributing sources to the segmentation dictionary are the Chinese collection for the TREC-5 and TREC-6 Chinese retrieval track and the Chinese words found in the LDC transfer dictionary described below.

One of the bilingual dictionaries we used to translate Chinese queries into English is the Chinese-to-English wordlist (version 2.0) compiled by Linguistic Data Consortium[2]. The wordlist consists of a list of Chinese words, each paired with a set of English words. It has about 128,000 entries. We will refer to this bilingual wordlist as the LDC transfer dictionary.

The second bilingual dictionary used in our experiments is the Kingsoft online dictionary[3]. We will call this dictionary the Kingsoft transfer dictionary. It consists of a general dictionary and a set of 23 specialized dictionaries, such as ships, electricity, telecommunication, law, broadcasting, environment, chemistry, economy and trade, computer, medicine, and so on. The general dictionary contains about four million entries and the specialized dictionaries together contain about two million entries[5]. In our experiments, only the Chinese query terms missing in the LDC transfer dictionary were looked up in the online version of the KingSoft dictionary. Again, the missegmented query terms were not manually corrected before translation.

The third source used in our translation is the Yahoo-China search engine available at http://cn.yahoo.com/. When a Chinese term is not found in both the LDC and KingSoft dictionaries, it is searched in Yahoo-China. The search results returned from Yahoo-China search engine contains a title and the sentences where the search term occurs. when the original English term is included, it usually appears right after or close to the Chinese term. The English words appear immediately after or close to the Chinese term are extracted from the search results. The extracted English words are taken to be the translations of the Chinese term.

Three Chinese to English cross-language retrieval runs were carried out. The Chinese queries were first segmented into words using the segmentation dictionary. The average length of the segmented Chinese queries (including the ones inappropriately segmented) is 2.49 terms.

[2] The bilingual wordlist can be downloaded from http://morph.ldc.upenn.edu/Projects/Chinese/

[3] accessible from http://ciba.kingsoft.net/online/

No. of Chinese terms	No. of English Translations
41	0
87	1
61	2
41	3
49	4
19	5
18	6
9	7
15	8
5	9
8	10
7	11
4	12
3	13
1	14
1	16
1	20
2	21
1	23

Table 3: Distribution of the number of English translations for the Chinese query terms in the LDC transfer dictionary.

The first run used only the LDC transfer dictionary to translate the Chinese queries into English by dictionary lookup. The missegmented query terms were not manually corrected before dictionary lookup.

Table 3 presents the distribution of the number of English translations for the Chinese query terms produced after word segmentation. Column 2 gives the number of English translations, column 1 is the number of Chinese query terms that have the same number of English translations. For example, the third row in the table means 61 Chinese query terms have two English translations found in the LDC transfer dictionary. The table shows 41 Chinese query terms have no English translations because they are missing in the LDC transfer dictionary. It also shows nearly half of the Chinese query terms have three or more English translations in this transfer dictionary.

A total of 373 terms were generated by segmentation, including the ones generated because of inappropriate segmentation. Examples of improper word segmentation will be presented below. Among all query terms, the Chinese query term ”混合” has the largest number of translations, which is 23. The translated English terms are listed in the following table:

```
to mix/to blend/admix/admixtsure/admixture/
battering/blending/brewing/ commingle/commix/
commixture/compositing/concoction/confect/
immixture/ingraft/interblend/interflow/
interfusion/intermixing/intermixture/meld/
sophistication/
```

Table 3 clearly suggests the need to disambiguate the translations of the Chinese terms. We adopted a simple method to choose translations if there are multiple ones for a Chinese query term. Our method is to retain the two English translations that occur most frequently in the English test document set. The translated terms and the words in the English test collection are normalized before the occurrences of words are counted. The normalization process changes plural nouns to their singular form, and verbs to their base form.

We decided to choose two as the number of translations to keep because sometimes the translation occurring most frequently in the test document set may not the best choice. So keeping two increases the chances of including the most appropriate translation if it is among the translations. It is also partly because some Chinese terms may have two or more equally appropriate translations. The second term may be a related term. In these cases, one would want to include both of them. For example, the Chinese term "石油" can be translated as either *oil* or *petroleum*. The top two translations for "汽车" are *car* and *auto*. And the top two translations for "犬" are *dog* and *canine*. A case where this selection strategy failed involves the Chinese term "气" in "电气", meaning *electricity and gas* in English. The top two translations by occurrence count in test document set are *air* and *gas*.

In the second run, the Chinese queries were translated into English using the LDC transfer dictionary just as was done in the first run. The Chinese words that are missing in the LDC transfer dictionary were looked up in the Kingsoft transfer dictionary. Since the words missing in the LDC transfer dictionary are most likely uncommon terms or proper nouns, one would expect there may be only one or a few translations for the missing words. For those words, we did not disambiguate the translations if there are more than one. All translations found in the Kingsoft transfer dictionary for the missing words in the LDC dictionary were retained. When a word is found in the general dictionary, only the translations from the general dictionary were retained; otherwise the translations found in all specialized dictionaries were retained.

In the third run, the Chinese queries were translated into English as was done in the second run except the words missing in both transfer dictionaries were looked up using the Yahoo-China search engine. Some of the query words were not recognized in segmenting the Chinese queries largely due to the limited coverage of the segmentation dictionary. For this run, we manually corrected the segmentation errors and then looked up those words in Yahoo-China search engine as well. So segmentation is not an issue for the third cross-language retrieval run.

The use of Yahoo-China search engine in finding translations for Chinese words is based on the observation that when an English term, particularly a new term, is used in written text, the English term sometimes occurs in parentheses right after its Chinese translations. For example, the word "欧元" which is the Chinese translation for *euro* was posed as a query to Yahoo-China search engine, one segment from the search results is "何谓欧元（ EURO)？ ". In this case, the original English term "euro" appears immediately following its Chinese translation. More queries and segments of their search results are presented in table 4.

Table 2 presents the evaluation results for the English monolingual retrieval run and three Chinese to English cross-language runs. The column labeled "MONO" shows the evaluation result for the English monolingual retrieval run. The columns labeled "CLIR1", "CLIR2", and "CLIR3" present the evaluation results for the first, second and third cross-language retrieval runs. The results show the overall precision of the first cross-language run using the LDC transfer dictionary alone in translating queries is only 44.5% of the overall precision for the monolingual run. When the Kingsoft transfer dictionary is used to translate words missing in the LDC dictionary, the overall precision for the second cross-language run increased from 0.0939 to 0.1069. The overall precision for the third cross-language run is further increased to 0.1183 when the words not found in both transfer dictionaries were looked up in Yahoo-China search engine and the automatic segmentation errors were manually corrected. The performance for the third CLIR run represents 56.1% of the overall precision for the monolingual run.

The poor performance of Chinese-English retrieval is the result of a number of factors: the coverage of the segmentation dictionary, the segmentation algorithm used, the coverage of the transfer dictionary, the quality of the tranfer dictionary, ambiguities in translation.

5 Failure Analysis

In this section, we examine several factors that may directly or indirectly have degraded the retrieval effectiveness of Chinese to English cross-language retrieval. The factors include segmentation and translation, particularly the transfer dictionary coverage, translation disambiguation, and inappropriate translations.

5.1 Segmentation

Since the queries are short, consisting of one or two concept terms in general, it is critical to translate them all into English appropriately. Because words are not explicitly marked in Chinese writing, it is necessary to segment the Chinese queries into words before dictionary lookup can take place. As mentioned above, a dictionary-based longest matching was applied to segment the Chinese queries. Overall the segmentation results are good, partly because the queries are very short. Nevertheless, several Chinese queries were incorrectly segmented because the words in these queries are missing in the segmentation dictionary. The Chinese words that were incorrectly segmented are presented in table 5. All of the Chinese words shown in the second column are missing in our segmentation dictionary. As a result, the average precision for these eight queries are almost zero. Another example of inappropriate segmentation is the fragment of text "残疾儿童教育 (teaching disabled children)" which was split into "残疾 (disabled)" and "儿童教育 (child education)". The appropriate segmentation would be "残疾 / 儿童 / 教育 " or "残疾儿童 / 教育 ".

5.2 Translations

First we will examine the transfer dictionary coverage. Table 3 shows 41 out of 373 Chinese query words are not found in the LDC transfer dictionary. A number of them are proper nouns. The words missing in the transfer dictionary are listed in table 6. The list of missing words also shows the variety of the topics used in this study. The performance for all the queries with words not translated are very poor since the queries are only one or two words long.

In the second cross-language experiment, we looked up the terms, missing in the LDC dictionary, in the Kingsoft online dictionary. About half of the terms are found in the Kingsoft dictionary even though some of the translations may not be the same as the original English terms. Such an example is

Chinese word	Segments of Yahoo-China search results containing original English terms
厄尔尼诺	... 大范围持续性增温现象称为厄尔尼诺事件 (El Nino)。厄尔尼诺是. . . ... 龙卷风与厄尔尼诺现象 (ELNINO) ...
欧元	... 何谓欧元（ EURO ）? 和欧洲货币单位（ ECU ）有何差异?
莱姆病	... 莱姆病简介（ Lyme disease ）　　莱姆病（ Lyme disease ）是一种人畜共通传染病, ... 莱姆疏螺旋体病 (Lyme borreliosis) 又称莱姆病 (Lyme disease),
申根	... 欧洲地区】申根 (Schengen) 德国 (Deutchland) ...
哈勃	哈勃 Hubble, Edwin Powell(1889.11.20 - 1953.9.28) 美国天文学家, . . .

Table 4: Segments of Yahoo-China search results that contain the original English terms. The first column shows the words posed as search queries to Yahoo-China search engine. The second column presents segments in the search results that contain the original English terms.

topic number	words	inappropriately segmented
309	饶舌乐 (Rap)	饶舌 / 乐
351	福克兰 (Falkland)	福 / 克兰
378	欧元 (euro)	欧 / 元
406	帕金森氏症 (Parkinsons' Disease)	帕 / 金 / 森 / 氏 / 症
410	申根 (Schengen)	申 / 根
418	绗缝 (quilts)	绗 / 缝
423	米里亚娜马尔科维奇 (Mirjana Markovic)	米里 / 亚 / 娜马尔 / 科维奇
441	莱姆病 (Lyme disease)	莱 / 姆 / 病

Table 5: Words that were inappropriately segmented.

home schooling which was manually translated into " 家教 " in Chinese, then the Chinese word was translated into *family education* by looking up the Kingsoft dictionary.

Some of the query terms, notably the proper nouns, are missing in both dictionaries. The missing terms include " 哈勃 (Hubble)", " 厄尔尼诺 (El Nino)", " 阿尔茨海默氏 (Alzheimer's), " 泛美 (Pan Am)", " 米洛舍维奇 (Milosevic)", and " 侯赛因 (Hussein)". Searching Yahoo-China found translations for an additional eleven Chinese words, which further improved the overall precision for Chinese to English retrieval. Table 7 lists the Chinese words and their translations found in the search results of Yahoo-China. Our procedure to extract translations from Yahoo-China search results is simple. The English words in parentheses following the Chinese term are extracted as the translation of the Chinese term. And in the case of no parentheses following the Chinese term, we are looking for English words occurring right after the Chinese term.

5.3 Inappropriate translations

For several queries, the translations from Chinese back into English by transfer dictionary lookup are not the same as the original English words. Two such examples are "hydroelectric" and "suicide". The term *hydroelectric* is manually translated into " 水力发电 " in Chinese. However when the Chinese term is translated back into English through the LDC transfer dictionary, the English translations are *hydropower* and *waterpower*. The second example is topic 424 which has just one term *suicide*. It was translated into " 自杀 ", however when it was translated back into English, the translations are *self-slaughter* and *self-destruction*. The LDC transfer dictionary lists four translations for this Chinese term (the infinitive preceding the verb is removed): *kill oneself*, *commit suicide*, *self-slaughter* and *self-destruction*. When the translations are counted in the test document set, the phrase translations are counted as a single term. Since the index terms in the test document do not include any phrases, the occurrence count of the phrase *commit suicide* is zero, resulting in not be selected as one of the translations of the Chinese term. One more example of inappropriate translation is topic 428 whose title is *Legionnaires' disease*. Its Chinese translation is " 军团病 ". The sole English translation according to the LDC transfer dictionary is *Legionnelosis*, a word occurred in no document in the test set.

The title of topic 353 is "Antarctica exploration". It was manually translated into Chinese as " 南极考察 ". When the Chinese term was automatically translated back into English by the LDC transfer dictionary lookup, the English translation becomes "south pole expedition". The Chinese term " 南极 " was translated into *south pole* instead of *Antarctica*. The LDC transfer dictionary contains the pair " 南极洲 /Antarctica". If the original Enlgish query was translated into Chinese as " 南极洲考察 " in the first place, the translated English query would contain the term *Antarctica*. Both Chinese translations of the phrase "Antarctica exploration" are appropriate, how-

Chinese words	Translations extracted from Yahoo search results	Original English
莱姆病	Lyme disease/LD/	Lyme disease
饶舌乐	Gangsta rap	rap
福克兰	FALKLAND ISLANDS	Falkland
欧元	EURO	euro
申根	Schengen	Schengen
哈勃	Hubble, Edwin Powell/TNT2/I/	Hubble
厄尔尼诺	ELNINO/El Nino/SST/GIANT	El Nino
米洛舍维奇	slobodan milosevic, 1941.8-	Milosevic
泛美	PanAmSat Corp	Pan Am
雪茄	Cuphea/Cuphea ignea A. DC.	cigar
马尔科维奇	Being John Malkovich/BEING JOHN/John	Malkovic

Table 7: The English words extracted from the search results of Yahoo-China search engine.

ever the shorter version, " 南极考察 ", is more commonly used.

The English term *bee* can be translated into Chinese as " 蜂 " or " 蜜蜂 ". The dictionary contains the English translation for the Chinese term " 蜂 " but not the other Chinese term " 蜜蜂 ". The English term *cigar* has two Chinese translations " 雪茄烟 " or simply " 雪茄 ". The LDC transfer dictionary contains the English translation for " 雪茄烟 ", but not for " 雪茄 ".

The topic 377, "cigar smoking", was translated into " 吸雪茄 " in Chinese. The English translation of " 吸雪茄 " by the LDC transfer dictionary lookup became "absorb/breathe". Here " 雪茄 " was not translated because it is missing in the transfer dictionary and " 吸 " was translated into "absorb/breathe".

6 Related work

Chen and Nie [2] described a mining system capable of automatically finding parallel texts. They used the parallel text to construct a Chinese-English bilingual dictionary that was used to translate queries. The parallel text complements existing bilingual dictionaries. Kwok [6] used a machine translation system and a small bilingual wordlist to translate short queries and found the bilingual wordlist complement the machine translation, resulting in an improvement in retrieval performance. Bian and Chen [1] used a bilingual dictionary of about 125,000 entries to translate queries. Four query translation selection strategies were tested. Oard and Wang [9] examined the effects of term segmentation on Chinese/English CLIR. And Levow and Oard [7] studied the lexicon coverage for CLIR.

7 Conclusions

We have presented the results for three Chinese to English cross-language retrieval runs using 150 short queries against a large English test document set. We have used two large transfer dictionaries and the Yahoo Chinese search engine to translate Chinese queries into English. although the topics we used vary over a wide variety of subjects, when we combined the two transfer dictionaries with Yahoo Chinese search engine, we were able to translate about 94% of the Chinese query words. The best performance of the Chinese to English cross-language retrieval is only about 54% of that achieved in English monolingual retrieval in spite of most of the query words being translated. The relatively poor performance of the cross-language retrieval can be attributed to a number factors such as the coverage of the segmentation dictionary, coverage of the transfer dictionaries, the quality of the transfer dictionaries, and the existence of multiple translations for many of the Chinese query words. Another factor is the shortness of the queries. Our failure analysis has shown several of the Chinese query words were not recognized as words because they are missing in the segmentation dictionary. Our analysis also has shown many of the Chinese query words have three or more translations which makes the selection of the most appropriate translation even more difficult. Our strategy of retaining the top two translations works well for some Chinese words, but become less effective for other words, and sometimes this strategy chooses the wrong translations. The idea of using the Yahoo Chinese search engine to discover translations for Chinese words works well for proper nouns. Our experiments have shown resources available for translation and resolving the disambiguity of translations can have large impact on the effectiveness of Chinese to English cross-language retrieval.

8 Acknowledgements

We wish to thank Michael Buckland and the anonymous referees for constructive comments. This research was supported by DARPA grant “Translingual Information Management Using Domain Ontologies” N66001-00-1-8911.

References

[1] Guo-Wei Bian and Hsin-Hsi Chen. Cross-language information access to multilingual collections on the in-

missing Chinese words	English translations
饶舌乐	(Rap) (Y)
福克兰	(Falkland) (Y)
欧元	(euro) (Y)
帕金森氏症	(Parkinsons' Disease)
申根	(Schengen) (Y)
绗缝	(quilts)
米里亚娜马尔科维奇	(Mirjana Markovic) (Y)
莱姆病	(Lyme disease)
小儿麻痹	polio (K)
哈勃	Hubble (Y)
濒危	endangered (K)
无线电波	radio wave (K)
悬浮	Levitation (K)
邪教	cult
偷税	income tax evasion (K)
亲生父母	biological parents (K)
黑熊	black bear (K)
病毒性	viral (K)
阿斯匹林	Aspirin (K)
阿尔茨海默氏	Alzheimer's
英吉利海峡	English Channel (K)
制衣厂	clothing sweatshops
厄尔尼诺	El Nino (Y)
厌食症	anorexia (K)
雪茄	cigar (Y)
残疾	disabled (K)
肥胖症	obesity (K)
有机	organic (K)
研发	R&D
家教	home schooling (K)
旅游业	tourism (K)
雨林	rain forest (K)
偷猎	poaching (K) (poach)
泛美	Pan Am (Y)
珍宝	treasure (K)
打捞	salvaging (K)
钢铁	steel (K)
金三角	Golden Triangle (K)
一氧化碳	carbon monoxide (K)
艺术品	art (K)
米洛舍维奇	Milosevic (Y)
童工	child labor (K)
侯赛因	Hussein

Table 6: Chinese query words not translated by LDC transfer dictionary lookup. The symbol "(K)" means the Chinese word is found in the Kingsoft online dictionary and the symbol "(Y)" means the Chinese word is found in the search results from Yahoo-China search engine but not in both transfer dictionaries.

ternet. *Journal of the American Society for Information Science*, 51:281–296, 2000.

[2] Jiang Chen and Jian-Yun Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, Washinton, USA, April 29-May 4, 2000*, pages 21–28, 2000.

[3] W. S. Cooper, A. Chen, and F. C. Gey. Full text retrieval based on probabilistic equations with coefficie nts fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[4] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227, 2000.

[5] KingSoft. http://ciba.kingsoft.net/ciba2000/cidian.htm.

[6] K.L. Kwok. English-chinese cross language retrieval based on a translation package. In *Machine Translation Summit VII workshop on Machine Translation for Cross Language Information Retrieval*, Kent Ridge Digital Laboratories, Singapore, 1999.

[7] Gina-Anne Levow and Douglas W. Oard. Evaluating Lexicon Coverage for Cross-Language Information Retrieval . In *Workshop on Multilingual Information Processing and Asian Language Processing*, 1999.

[8] Douglas W. Oard and Anne R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*. American Society for Information Science, 1998.

[9] Douglas W. Oard and Jianqiang Wang. Effects of Term Segmentation on Chinese/English Cross-Language Information Retrieval . In *Symposium on String Processing and Information Retrieval (SPIRE)*, 1999.

[10] M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1985.

[11] E. M. Voorhees and D. K. Harman, editors. *The Fifth Text Retrieval Conference (TREC-5)*, National Institute of Standards and Technology, Gaithersburg, MD, 1997.

[12] E. M. Voorhees and D. K. Harman, editors. *The Eighth Text Retrieval Conference (TREC-8)*, National Institute of Standards and Technology, Gaithersburg, MD, 1999.

[13] E. M. Voorhees and D. K. Harman, editors. *The Seventh Text Retrieval Conference (TREC-7)*, National Institute of Standards and Technology, Gaithersburg, MD, 1999.

[14] Ellen M. Voorhees and Donna Harman. Overview of the Seventh Text REtrieval Conference (TRC-7). In Ellen M. Voorhees and Donna Harman, editors, *The Seventh Text REtrieval Conference (TREC-7)*, pages 1–23, July 1999.