English-Chinese Cross-Language Retrieval based on a Translation Package

K.L. Kwok

Computer Science Dept., Queens College, City University of New York
Flushing, NY 11367, USA
email: kwok@ir.cs.qc.edu URL: http://ir.cs.qc.edu

Abstract An inexpensive COTS translation package, augmented with a downloadable bilingual dictionary, was employed for a study of English-Chinese cross-language information retrieval (CLIR) using the query translation approach. The experimental setting involved the 170 MB Chinese collections and 54 queries of TREC and their relevance judgment, and our PIRCS bi-lingual retrieval system. With some standard retrieval techniques such as pre-translation query expansion and combination of retrieval lists, we were able to achieve over 70% of monolingual results for both long and short queries. Insufficient context of short queries appears not a problem for machine translation for English-Chinese CLIR.

1 Introduction

CLIR has gained importance in recent years (Oard & Dorr 1996, Grefenstette 1998, Schauble & Sheridan 1998) because accessing foreign web sites and text searching have become popular and convenient. Many language pairs need to be considered, but one can fairly say that automatic English-Chinese cross language IR would become increasingly important because of the growing significance of China in business, politics, science & technology, etc. as well as the sheer size of the Chinese population. English also is practically the de facto world language. Thus, the ability to do effective retrieval of Chinese language collections using English language queries without incurring professional translation costs would be of great convenience and utility to those users who need to search or monitor Chinese information.

CLIR appears to be a 'good application of crummy translation' (Church & Hovy 1993) because IR is mainly content term based, and output of 'word salad' is usable by an IR system. Style and readability of target language generation are not necessary, simplifying matter. Approaches to CLIR include translating documents to match the language of the query (Oard & Hackett 1998), or converting both documents and queries into an intermediate representation such as the LSI method (Dumais et.al. 1997). But, by the far the simplest approach is to convert queries to the document language, do monolingual ranked retrieval, then report and translate documents back to the query language. Many CLIR using this approach (Davies 1997, Pirkola 1998, Ballesteros & Croft 1998, Eichmann et.al. 1998), have been based on large dictionary look-up and some ambiguity resolution methods. They are done between English and European languages, where cognates may be helpful. English and Asian languages are radically different, resources limited, and CLIR investigations are not as prevalent. Recently, however, the confluence of three factors have made English Chinese CLIR a possibility, viz. availability of TREC Chinese collection and evaluation data, COTS machine translation software, and a reasonable size, freely available bilingual dictionary. With these resources, we have been able to throw some light to the following questions concerning CLIR (Kwok 1997): what would be the effectiveness when English queries are used to retrieve on the Chinese collection? How does it compare with monolingual results? And what techniques could one employ for improvements?

2 Three Timely Resources

2.1 Evaluated Collections

In the past few years, the annual TREC (Text REtrieval Conference) large-scale blind experiments sponsored by NIST and DARPA have given immense impetus to IR research. During TREC 5&6, monolingual Chinese retrieval was investigated with a fairly large GB-encoded Chinese collection of 170 MB in size and 54 queries (Voorhees & Harman 1997, 1998). Accompanying each Chinese query is an English counterpart. Although these are not exact translations, they carry the same meaning pretty closely, and we will regard them as a standard English query set to start with. Each query also has a set of Chinese answer documents that have been judged relevant to the query or not. These have been obtained manually from retrieval lists submitted from multiple search engines that participate in the Chinese track. The relevant answer documents are not exhaustive, but an approximation to it. With them, one can evaluate the standard recall and precision curve of ranked retrieval lists, and make comparisons among them.

2.2 MT Software

A low-cost English to Chinese machine translation software called Transperfect (http://www.otek.com.tw) has been available for PCs with claims of 80% accuracy. The software comes with a proprietary 100K English-Chinese basic dictionary and English sentence parser. It can translate a word, sentence or whole document. For document translation, it can work in unique (u) or multiple (m) translation mode. (Other modes of output are also available such as alternative sentence outputs that list different candidate sentence structure, or English Chinese interleaved outputs that are useful for manual editing. But they do not seem useful for IR purposes.) In u-mode, all English source words have only one translation, and some words may get their erroneous senses. In m-mode, some words in the output contain multiple (candidate) translations. This way, the translation result gets a higher probability of having the correct sense of each word covered at the risk of more noisy output. Words that the software cannot handle are left untouched in the output in the original English. These are mostly acronyms and abbreviations like WTO (World Trade Organization), APEC (Asian Pacific Economic Commission), or place and person names like Bosnia, Xinjiang or Mandela, etc. However, the package does list separately other domain-specific dictionaries such as business and commerce, computer, etc. that we did not purchase. In addition it also allows the user to maintain manually a separate user dictionary to capture whatever new words the user desires.

Use of a translation package is convenient for CLIR since it takes care of problems like word morphology, parsing, etc. On the other hand, its inner workings and dictionaries are proprietary, and one can only treat it as a black box and has to accept the output as is with little possibility of influencing them. Also, it appears that no application interface is available. The package is developed in Taiwan where Big5 coding is used. But it comes with a convenient utility for conversion between Big5 and GB coding. The TREC Chinese collections consist of Mainland China news stories. Since Taiwan or any local region may use a dialect or language peculiarities of its own, the resultant translated queries may have mis-matches with the documents, and it may not be ideal. Nevertheless, we have employed the package for translating the TREC 5&6 English queries (of the Chinese track). The translation appears reasonable; and only about 5% of the words are left un-translated. An example is displayed below using the m-mode translation:

Query #29

Title in English:
Building the Information Super Highway

Title in Original Chinese:
信息高速公路 的 建设

Title by Machine Translation:
建筑 [建立] 消息 [知识/报告] 上等的 表面的] 公路 [大道/直截的途径]

Description in English:
Information Super Highway, building

Description in Original Chinese:
信息高速公路，建设

Description by Machine Translation:
消息 [知识/报告] 上等的 表面的] 公路 [大道/直截的途径] , 建筑 [建立]

Narrative in English:
building the Information Super Highway, including any technical problems, problems with the information infrastructure, or plans for use of the Internet by developed or developing countries.

Narrative in Original Chinese:
信息高速公路的建设，包括任何技术上的，
»蛲胄畔11. 冂枋┼泄氏奈侍猓币约坝泄
胤 4. 锎 一蚍⑽怪泄 叶怨 释 ç
的应用计划 .

Narrative by Machine Translation:
建筑 [建立] 消息 [知识/报告] 上等的 表面的] 公路 [大道/直截的途径] , 包括任何的技术上的 [专门的/工业的] 问题 难题] ，消息 [知识/报告] 下部构造下部组织] 的问题 [难题] , »蛭[a] 因为/至于] 网际网路的使用 [习惯/价值] 的计划 [策略/方法] 被 [
经由/在～之前] 发展 [发达/进步] 或发展中的国家 [乡村/山地] 。

It can be seen that `information` was translated into 消息 (news) 知识 (knowledge)报告 (report) which are not as accurate in this context as 信息, . Similarly 'Super' is wrongly translated as 上等的 [表面的] (upper class, superficial). 'Internet' however is reasonably correct. All in all, it does seem to indicate what the query is talking about.

2.3 Downloadable Bi-lingual Dictionary

As discussed before, about 5% of the query words failed to get translated with the basic dictionary in the MT software. These are mostly acronyms and names that are highly specific and valuable for retrieval. No translation package can cover all terminology over time. Hence, using multiple resources may reduce this out-of-vocabulary problem and enhance CLIR effectiveness. Unfortunately most English Chinese dictionaries are commercialized. They can be used for consultation but not for stand-alone applications. Recently however, Paul Denisowski has accumulated a reasonable size (~20K) Chinese-English dictionary on the web, and it is freely available for research purposes (http://www.mindspring.com/paul_denisowski). It does contain some of the abbreviation/acronyms that the basic Transperfect package does not have. We have employed this dictionary to diminish our OOV problem on the translation output and to see its effect on CLIR. Example entries of the dictionary are given below:

东南亚国家联盟 [dong1 nan2 ya4 guo2 jia1 lian2 meng2] /ASEAN (Association of Southeast Asian Nations)/

大麻 [da4 ma2] /hemp/marijuana/

新疆 [xin1 jiang1] / Xinjiang (Uygur Autonomous Region)/

乌鲁木齐 [wu1 lu3 mu4 qi2] /Urumqi (capital of Xinjiang autonomous region)/

The dictionary is general, simple and contains no POS information of either the Chinese or the English. Each Chinese entry is followed first by the pronunciation in brackets, then the translation(s) separated by slashes. Within slashes, a translation may be followed by further explanation that is enclosed in parenthesis. We have indexed the dictionary by each English word (after some stopword removal and stemming trailing s) to obtain a reverse English Chinese. Then, for each un-translated query word we performed a dictionary look-up and output all Chinese entries that were indexed under this word. These are then filtered by starting to match the longest phrase in the query pattern first (and later reducing word by word) and according to the following order: 1) exact or lead string match - meaning that either the English query pattern (word or phrase) exactly matches a dictionary entry that is not parenthesized, or it matches the lead string portion; 2) embedded match - meaning that the query pattern is embedded in a non-parenthesized dictionary entry; 3) query pattern matches or embeds in the parenthesized text. This provides phrase translation before single words, attempts to identify more important entries first, as well as reduces the number of translations, and is our method of disambiguating the dictionary look-up process. In this experiment, we did not reduce un-translated phrases to single word matching, but look for full phrase matching in the dictionary.

In Fig.1, we display for each query the words that are found in this dictionary (marked with an asterisk *), together with their frequencies. With this procedure, acronyms (like: WTO, NATO) and proper nouns (like: Spratly, Bosnia) were picked up correctly, some with multiple translations. In the example dictionary entries above, for 'Xinjiang' of query #CH10, our procedure will pick up the first translation but not the second because matching in parenthesized wordings are avoided when the better exact match occurs. The dictionary lookup makes additional 48 translations and affects some 16 queries.

We did not attempt to perform dictionary look-up first then followed by MT because the 20K size is probably too small to use the dictionary as the primary translation tool.

3 CLIR Results and Discussion

3.1 Retrieval System and Evaluation Measures

The retrieval engine employed for experimentation is our PIRCS (Probabilistic Indexing & Retrieval -Components - System) which supports both English and Chinese retrieval. It has been used for TREC experiments with highly successful results (Kwok et.al 1998). For Chinese text, we employ our approximate short-word segmentation and use these short-words and characters as indexing terms. Given a query, PIRCS ranks documents based on a probabilistic algorithm that takes account of global and within-document frequencies of terms in a natural way. Retrieval is done in a 2-stage fashion: an initial retrieval with a given raw query returns some d top-ranked documents. These documents are assumed to be either relevant or highly related to the query topic domain and mined for query-associated terms. These are then weighted and added to the raw query. Retrieval with the expanded query usually provides better final results. In the following, unless otherwise stated, results using this 2-stage process are reported.

Many effectiveness measures have been used for IR experiments and they are averaged over all queries. These include: Pn (precision of retrieval at n documents retrieved) which is the percent of n documents that are relevant and is highly user-oriented; RR1k (number of retrieved relevants ranked within the top 1000) which is useful for people needing high recall; and AvP (average over all precision values at positions where a relevant documents is ranked) which provides a measure that may characterize a system's effectiveness over all recall range. TREC traditionally has used the last one as the official measure for run comparisons, and we will follow the convention.

3.2 Long Query Results

All 54 topics of TREC 5&6 topics were processed as queries to retrieve against the 170MB Chinese collection. In this sub-section, we report on the long query results. As shown in Fig.1, TREC topic wordings are separated into title, description and narrative sections. We convert topics into a stream of running text and long queries use all sections. They are paragraph(s) in size averaging over 20 English content and 36 original Chinese unique terms on average.

We first establish a monolingual basis by which we can measure cross language results. The original Chinese queries were used for retrieval against the Chinese collection and results are tabulated in Table 1 Column 1. It can be seen that P10, P20 are respectively 0.735 and 0.684. Thus over 7 out of the top 10, and nearly 14 out of the top 20 ranked documents are relevant. RR1k is 4849, which is 94% of the possible number of judged relevants. AvP is 0.536. These numbers represent very good retrieval.

The same queries in English were then processed through the Transperfect package. Two translation outputs were obtained: one in u-mode and another in m-mode. With u-mode, the queries have an average of 33.7 unique terms while with m-mode the average value is about twice. Un-translated English words were left as is. These outputs were used untouched for retrieval. Results

are tabulated as Columns 2 & 3, Fig.1. It can be seen that effectiveness diminished substantially: AvP of about 0.3 is only 56% of monolingual 0.536. The precision at 10 and 20 documents P10, P20 are 57-58% of monolingual results. The difference between the two modes of translation is small, with a slight edge for u-mode. M-mode translation, because of more variety in wordings, helps in reducing missing relevants but can also bring in irrelevants. Apparently, our PIRCS system can tolerate these noisy terms. The overall result compares favorably with the traditional 50% monolingual results obtained by simple bilingual dictionary lookup of words (Oard & Dorr 1996).

As discussed earlier, many of the un-translated words are person and place names or acronyms of organizations. They are highly specific and their absence may adversely impact retrieval. To reduce this problem, we employ Denisowski's downloadable bilingual dictionary to resolve some of these OOV terminology. Altogether, some 16 of the 54 queries were affected. Results after using this dictionary are tabulated in Columns 4 & 5, Table 1. This process brings improvements of 7-10% better precision than not using the dictionary. Again, the two modes behave similarly, and with AvP of about 0.33, we have achieved close to 62% of monolingual effectiveness.

Pre-translation query expansion has been known to improve CLIR in an English-Spanish setting (Ballesteros & Croft 1997). This means using a separate English collection for pre-translation retrieval in order to expand the English query with highly associated English terms. These terms may help focus on the query topic and bring more translated terms that together are useful for disambiguating the translation effect. We have also experimented with this procedure using the Foreign Broadcasting collections of TREC and employing various levels of query expansion. Columns 6 & 7 show the two modes with 45-term expansion. This brings about over 10% of additional improvements compared to not using it. Effectiveness is now about 68-70% compared to monolingual. For example, the u-mode AvP value of 0.373 is about 70% of the basis 0.536, while the m-mode AvP is 0.365 or 68%. The precision at n documents retrieved is also quite reasonable. For example, the m-mode P10 and P20 values are .543 and .517 respectively. This means one can expect on average over 5 good documents at 10 retrieved and over 10 at 20 retrieved, and represents 73-75% of monolingual results. In real-life, such retrievals are quite usable.

Since the u-mode translation is unique, we assume their translations are more probable and accurate than the additional alternatives provided via the m-mode. Also, pre-translation query expansion does not always lead to better results. This suggests to us to hedge by combining the retrievals of u-mode without pre-translation expansion retrieval list (Column 4) with that of m-mode with pre-translation expansion (Column 7) using a ratio of 3:7. Results are shown in Column 8. The AvP value is now 0.379 or 71% of monolingual. P10, P20 are respectively 0.567 and 0.522, which are over 75% of monolingual. Moreover, the RR1k of 4079 summed over all 50 queries is 84% of the monolingual 4852 value. Thus, for these long queries, using a basic commercial machine translation package, a downloadable bilingual dictionary, TREC collections and evaluated results, and some standard IR techniques, we are able to achieve quite reasonable and usable English Chinese cross language results.

3.2 Short Query Results

Since long, paragraph size queries are unrealistic as most users would not issue them, we perform another set of experiments using only the title section of each topic. The 54 queries average to about 4 English content terms per query and the original Chinese average to about 6.5 terms. Using translated queries, the average number of index terms is 6.3. The same results are tabulated in Table 2 below. Long queries have been shown to perform better than short queries for both languages in

Table 1: English Chinese CLIR Results - Long Queries

Col	1	2	3	4	5	6	7	8
	base	**u**	**m**	**u + dict**	**m + dict**	**u + dict+ exp**	**m + dict+ exp**	**cmb col 4&7**
RR 1k	4849	3363	3332	3574	3707	4091	4035	4079
%	100	69	69	74	76	84	83	84
AvP	.536	.301	.296	.332	.325	.373	.365	.379
%	100	56	55	62	61	70	68	71
P10	.735	.426	.413	.467	.454	.526	.543	.567
%	100	58	56	64	62	72	74	77
P20	.684	.392	.407	.434	.447	.488	.517	.522
%	100	57	60	63	65	71	76	76

Table 2: English Chinese CLIR Results - Short Queries

Col	1	2	3	4	5	6	7	8
	base	**u**	**m**	**u + dict**	**m + dict**	**u + dict+ exp**	**m + dict+ exp**	**cmb col 4&7**
RR 1k	4447	2650	2427	2946	2726	3610	3515	3793
%	100	60	55	66	61	81	79	85
AvP	.449	.251	.212	.277	.237	.290	.290	.354
%	100	56	47	62	53	65	65	79
P10	.624	.335	.311	.372	.348	.411	.420	.509
%	100	54	50	60	56	66	67	82
P20	.572	.321	.284	.354	.319	.389	.411	.497
%	100	56	50	62	56	68	72	87

monolingual retrieval (Voorhees & Harman 1998, Kwok 1999), and this is no exception in a CLIR setting.

As expected, the basis of the monolingual Chinese retrieval for short queries is 16% worse than for long queries (AvP of .449 vs .535). When we use only MT for CLIR, the unique mode is much better than the m-mode (AvP .251 vs .212, Table 2 Columns 2 and 3). Short queries are more sensitive to added noise words (Kwok 1999), and the alternative senses of words included for the m-mode translation appear to be the cause. Both modes improve about 10% when some remaining OOV words are removed by bilingual dictionary translation (Columns 4 and 5). When pre-translation English query expansion were performed using an expansion level of 15, the m-mode result improves substantially to 0.29 AvP and equals that of the u-mode result. If the u-mode without pre-translation expansion (Column 4) is combined with the m-mode with pre-translation of Column 7, a further boost is obtained achieving an AvP of .354, which is over 78% of the monolingual result of 0.449. On a percentage basis, it is surprising that short query CLIR does better than long queries. On absolute values, short query AvP at .354 is 7% worse than for long. We definitely need more variety of queries than these 54 to confirm this result.

It has been reported before that short queries may not provide sufficient context for a MT software to translate properly. This is apparently not true for the English Chinese pair. It could also be that Transperfect actually uses some simple strategies that do not rely much on context, so that the effect of context is small. Thus, at least with these 54 queries and the TREC Chinese collections, CLIR is quite successful for both long and short query types by employing a MT tool with an available medium size bilingual dictionary. We believe that similar or better results might be obtained if additional domain-specific dictionaries (either from Transperfect or elsewhere) were used. It also seems that the u-mode performs better than the m-mode in general, but combination of them gives the best results.

3.3 Further Analysis

The retrieval process used for the previous experiments is quite sophisticated and requires query expansions (English) before and (Chinese) after translation. Expansion brings in associated terms that are dependent on the top ranked documents of an initial retrieval. The process is statistical in nature, involving many terms and not easy to analyze. In an effort to analyze how good the MT software works for CLIR, we display in Fig.2 a comparison of the simplest retrieval results for the 54 short queries. These are the initial retrieval results for the monolingual (AvP .361) and the CLIR (AvP .193) that do not involve any query expansion. An initial retrieval only uses the terms that are available with a raw query. Generally, but not always, ineffective initial retrieval would lead to ineffective results with query expansion, because bad initial retrieval usually means the top documents are not relevant and they would not be able to bring good and related terms for expansion.

Based on Fig.2 we classify queries into three types:

a) 9 queries that have both monolingual and CLIR results of AvP <= 0.1. They are #2, 5, 7, 13, 17, 18, 33, 36, and 51. We assume that these are ineffective, not because of translation, but because the query topics are difficult for IR.

b) 34 queries with monolingual AvP > CLIR AvP. We further attempt to determine the cause for the deficit qualitatively by manually scanning the index terms for each query. 19 were due to bad translation. Examples are: #1 concerning '.most-favored nation', there is something like an abbreviation, acronym or slang for this entity in Chinese (最惠国), and it was translated wrong. Similar reason may be given to queries #12 'World Conference on Women', #19 'Project Hope'. Some non-compositional like phrases such as 'peace-keeping force' in #11, 'missing-in-action' in #20 have special terminology in Chinese and are also not picked up. In #9 and #32 concerning 'drug problems in China' and 'drug traffickers in Latin America', the word 'drug' was translated to the pharmaceutics sense (药). Other cases include #28 'cellular phones' was given the biological cell sense and #53 'auto industry', 'auto' was translated as 'automatic'. There were also 2 queries in this group of 19 where the Chinese version was more detailed than the English: #21 'The role of the Governor of Hong Kong in the Reunification with the PRC', and #47: 'The Impact of the 1991 Mount Pinatubo Volcano'. In #21, the Chinese version has the name of the governor as well, but even for that, the translation query probably would not catch up because 'governor' was not translated right, and 'reunification', 'PRC' were not in the dictionaries. In #47, 'Philippines' was in the Chinese query, but 'Mount Pinatubo' was an OOV.

In the rest, 14 actually have understandable translation but they were not good for retrieval. Some are due to the wordings that do not match well with the documents. An example is: #3 '核电站 ' which is the original Chinese for 'nuclear power plants' was translated as '核子发电厂' which is too detailed. One final case is due to picking up too many alternatives when looking up the external dictionary for the word 'Xinjiang' for query #10.

c) queries with monolingual AvP <= CLIR. There are 11 of them and 8 were determined to be reasonable translations. These together with the 14 in b) may be seen as 22 cases and randomly some get better and others get worse results. For the rest three queries, one that we considered more appropriate translation than the original is #22 where 'infection' was translated to the more common 传染 rather than the original Chinese 感染. In #35, the original basis was penalized because the entity South Africa (南非) was segmented wrong and got split up. The last case is #8 'Numeric indicators of earthquake severity in Japan'. The translation for 'numeric indicators' 数值指示器 was bad. But the original Chinese query includes words like 'deaths and injuries'

which were lacking in the English version and therefore not translated. These terms may function as noise words for the basis.

4 Conclusion

We have employed an inexpensive commercial translation package augmented with a free, downloadable bilingual dictionary to work with our PIRCS retrieval system for English-Chinese cross- language retrieval experiments. We have taken the query translation approach. In general, we recognize that these tools are not sufficiently sophisticated for translating these diverse, domain-unrestricted queries for large-scale CLIR. On the other hand we are surprised that we can achieve as much as over 70% of monolingual accuracy. The results leave us optimistic that, with more professional and better language-matching translation software, and larger bilingual dictionaries, we should be able to achieve results that are close to that of monolingual.

References

Ballesteros, L & Croft, W.B. (1997). Phrasal Translation and Query Expansion Techniques for Cross- Language Information Retrieval. In: Proc. of 20th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.84-91, ACM Press: NY.

[BaCr98] Ballesteros, L & Croft, W.B. (1998). Resolving ambiguities for cross-language retrieval. In: Proc. of 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.64-71, ACM Press: NY.

Church, K.W. & Hovy, E.H. (1993). Good applications for crummy machine tranlation. Machine translation 8 pp.239-259.

Davies, M. (1997). New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In: Information Technology: The Fifth Text REtrieval Conference (TREC-5), E.M.Voorhees & D.K. Harman, (eds.), NIST SP 500-238, pp.447-453. GPO: Washington, D.C.

Dumais, S.T, Letsche, T.A, Littman, M.L & Landauer, T.K (1997). Automatic cross-language retrieval using latent semantic indexing. AAAI-97 Spring Symposium: Cross-language Text and Speech Retrieval; pp.15-21.

Eichmann. D., Ruiz, M.E. & Srinivasan, P. (1998). Cross-language information retrieval with the UMLS metathesaurus. In: Proc. of 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.72-80, ACM Press: NY.

G. Grefenstette (ed.) Cross language Information Retrieval. Kluwer, 1998.

Kwok, K.L, Grunfeld, L & Xu, J.H. (1998) "TREC-6 Chinese and English retrieval experiments using PIRCS". In: The Sixth Text REtrieval Conference (TREC-6), D.K. Harman, ed. NIST Special Publication 500-240, Gaithersburg, MD 20899. pp.207-214.

Kwok, K.L. (1999). Employing multiple representations for Chinese information retrieval. J.of ASIS, 50:8 pp.709-723.

Kwok, K.L. (1997) Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment. AAAI-97 Symposium on Cross Language Text & Speech Retrieval. pp.133-137.

Oard, D.W. & Dorr, B.J. (1996) A Survey of Multilingual Text Retrieval. CS-TR-3615, Univ. of Maryland, Institute for Advanced Computer Studies.

Oard, D.W & Hackett, P (1998). Document translation for cross-language text retrieval in the University of Maryland. In: Information Technology: The Sixth Text REtrieval Conference (TREC-6), E.M.Voorhees & D.K. Harman, (eds.), NIST SP 500-240, pp.687-696. GPO: Washington, D.C.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proc. of 21th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.55-63, ACM Press: NY.

Schauble, P & Sheridan, P (1998). Cross-language information retrieval (CLIR) track overview. In: Information Technology: The Sixth Text REtrieval Conference (TREC-6), E.M. Voorhees & D.K. Harman, eds. National Institute of Science Technology SP 500-240, pp.31-44.

Voorhees, E.M. & Harman, D.K (1997). Overview of the Fifth Text REtrieval Conference (TREC-5). In: Information Technology: The Fifth Text REtrieval Conference (TREC-5), E.M.Voorhees & D.K. Harman, (eds.), NIST SP 500-238, pp.1-28. GPO: Washington, D.C.

Voorhees, E.M. & Harman, D.K (1998). Overview of the Sixth Text REtrieval Conference (TREC-6). In: Information Technology: The Sixth Text REtrieval Conference (TREC-6), E.M.Voorhees & D.K. Harman, (eds.), NIST SP 500-240, pp.1-24. GPO: Washington, D.C.

Query	Words	Query	Words	Query	Words	Query	Words
#2	reunification (4)	#12	4th world conference		herzogovenia		haihe
#3	daya		(2)	#25	ecoprotection		liaohe
	qinshan	#14	yunnan*	#27	robotic		songhua
#6	wto (3)*		hiv	#28	psdn	#43	lama (5)*
#7	prc (2)	#15	un (4)*	#30	betweeen+		indepedenc
	spratly (3)*		multination		1983-1993	#44	resettlemer
	dongsha	#16	un (7)*	#31	castro	#46	sino (2)
	xisha	#17	apec (3)*	#32	traffikers (3)+		vietnamese
	asean*		wto (3)*		cali		nongovernr
#8	richter		signatory		medina		campuchea
#9	cocaine	#19	benefited		traffiers+	#47	pinatubo
	marijuana*	#20	mia's	#33	hijackings		minatubo
	trafficking		vietnam (4)*	#34	arrid+		subic
#10	xinjiang*		vietnamese*		acerage+		clark
	uigur	#21	prc (3)	#35	mandela*	#51	formaulatec
	trading*		reunification (2)	#38	survelliance+	#53	sino (3)
#11	un (2)*		dingkang (2)	#39	assasination+		f-16 (3)
	bosnia (2)*	#23	un*	#41	kowloon (3)		
	nato*	#24	bosnian*	#42	yangtze*		
	bosniaun		bosnia*		huaihe		

Fig.1: Dictionary translated (*), Misspelled (+) and Un-translated Words in Queries