NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS

K. L. Kwok
Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
email: kwok@ir.cs.qc.edu

Abstract

We participated in the monolingual Chinese and English-Chinese cross language retrieval track using our PIRCS retrieval system. Employing the query translation approach for crosslingual IR, two methods of translation were tried: MT software, and dictionary lookup followed with disambiguation techniques. Retrieval lists from the two methods were combined to form the final result. Pseudo-relevance feedback was used, but no pre-translation expansion or collection enrichment was employed. Runs for all query lengths were submitted. Short-word with character representation was used for both documents and queries. Using the 'rigid' criteria for evaluation, both VS (very short) and SO (short) queries gave monolingual mean average precision at over 0.6. LO (long) query type surprisingly was worse by about 5% at over 0.57. TI queries (title only of a few words) returned a good 0.46 mean average precision. Cross language retrievals perform at between 77 to 83% of monolingual for the longer query types, and only at 55% for the TI queries.

A prominent factor in cross language retrieval failure is unknown word (such as proper noun) translation. This is particularly acute with TI queries of a few words. We show that it may be overcome to some extent by using longer queries that can provide more redundancy and better context for translations to hedge for errors. Crosslingual IR can also be efficiently performed, often with improved results, by mixting both translations as one single query.

Keywords: *Query Translation and Disambiguation; Bilingual Wordlists; MT software; CLIR*

1 Introduction

Cross language information retrieval (CLIR) is important because it provides the ability for people of one culture to filter and retrieve stored information of another. This is rendered even more useful because of the ease and convenience of access and delivery of foreign documents provided by Internet logistics. Accessing Chinese documents via English is a major sub-problem within CLIR because so many people in the world use these two languages, the fact that Chinese is so different from English, and that the Chinese language is perceived to be quite difficult for foreigners to learn. This paper describes the methods and procedures we took in participating in this NTCIR-2 Chinese and cross language retrieval track. Our PIRCS system was used for retrieval, and we employed the query translation approach for CLIR. Results for all query sizes were submitted, viz.: LO (long - all sections in a topic were used), SO (short - title, query and concepts sections), VS (very short - title and concepts) and TI (title section only). Section 2 describes our procedures for translating the English queries into Chinese with disambiguation; Section 3 summarizes our PIRCS system and retrieval methodology; Section 4 discusses our submitted results; Section 5 describes some additional experiments with other query types and methods, and Section 6 contains our conclusions.

2 Query Translation and Disambiguation

Although several approaches are available to tackle CLIR [1,2] we believe, as many others do, that query translation is often efficient and easier to implement, and may be just as or more effective. Within the query translation approach, one could use more IR-oriented techniques such as [3] (which emphasizes on frequency and term-weighting even though translation model and HMM approach for retrieval are introduced), or more linguistics-oriented techniques such as [4] (where parsing and phrase identification and translation seems to play a larger role). We investigated a combination approach where our highly effective PIRCS IR system is employed with two methods of translating English queries into Chinese, viz.: MT software and bilingual wordlist [5]. Even though we did not implement the MT software, we assume that a certain amount of linguistics sophistication has been employed. Translation is inherently ambiguous; we apply a common-sense observation that 'two heads are better than one'. The 'Transperfect' package (now known as 'TransWiz', produced by a Taiwanese company Otek http://www.otek.com.tw) seems ideal since it should produce Chinese translations that would agree well dialectically with the collection, which also has its source from Taiwan. We had experience with this

package before [6]. It can translate via a u-mode (meaning that every English word/phrase would produce a unique Chinese term), or an m-mode which can output at most three Chinese terms to hedge for translation errors. We employed only the u-mode because we intend to rely on our second method, dictionary translation, for hedging purposes.

For dictionary approach, we used the Chinese-to-English bilingual wordlist from LDC (Linguistic Data Consortium http://www.morph.ldc.edu/Projects/Chinese, which we call ldc2ce) as our main translation dictionary. It has about 120K entries. In addition, about 6000 pairs of word/phrase translations were obtained from the Hong Kong Law parallel text collection that is also available from LDC. After dictionary look-up, an English word usually has many translations. We employed four successive procedures to try to disambiguate and isolate the more correct outputs. These four methods were discussed in [7]:

- Dictionary structure-based: ldc2ce format is employed to select the more correct ones among word translations;
- Phrase-based: ld2ce can be regarded as a phrase dictionary by matching query strings with English explanations of Chinese terms, giving much more accurate phrase translations;
- Corpus frequency-based: select translation words with higher occurrence frequency in the target collection, which usually have higher probability of being correct; and
- Weight-based: a Chinese term set translated for one English term can be weighted as a synonym set that is more effective for retrieval [8].

Thus, MT software and dictionary translation procedures provide us with two independent sets of queries. These were used separately for retrieval. An example of translation is query #50 of the SO type (topic, question, concepts) shown below. The Chinese word set enclosed between two ^'s indicate the mappings for each English word and the numerics indicate the number of terms in a synonym set.

Query #50: Original English
Hot springs resorts in Taiwan .
****the related information about the hot springs resorts in Taiwan.
spa, hot springs, travel, spa site, hot springs resorts, soak in hot springs, spa therapy, spa soaking, spring soaking, spa pool, water quality, healing effect, skin, facility, consume, special product (of a place), inn, hotel, accommodation, transportation, itinerary plan.

Query #50: Translation by Transperfect
在台灣的溫泉度假地。
****有關的關於在台灣的那些溫泉度假地的信息。
礦泉， 溫泉， 旅行， 礦泉場所， 溫泉度假地，吸入溫泉，礦泉治療， 礦泉浸泡， 加熱的彈簧，礦泉池，水質，治療的效應，皮設備消耗特別產品(地方之中)小灑館飯店膳宿供應， 運輸旅行路線計劃.

Query #50: Translation by Wordlist
^1.0 溫泉城^ ^1.0 娛樂場^ IN ^1.0 台灣^.
**** THE
^0.25 敘述 0.25 有關系 0.25 講述 0.25 有關聯 ^^0.25 消息 0.25 信息 0.25 見聞 0.25 訊 ^
ABOUT THE
^1.0 溫泉城^ ^1.0 娛樂場^ IN ^1.0 台灣^.
^1.00 溫泉浴場 ^ ^1.0 溫泉城^^0.25 旅遊 0.25 旅行 0.25 遊歷 0.25 跋 ^^1.00 溫泉浴場 ^^0.33 地點 0.33 所在 0.33 網址 ^ ^1.0 溫泉城^ ^1.0 娛樂場^^0.33 浸透 0.33 泡 0.33 漬 ^ IN
^1.0 溫泉城^^1.00 溫泉浴場 ^ ^1.0 療法^^1.00 溫泉浴場 ^ ^1.0 使人濕透^^0.33 春天 0.33 彈簧 0.33 泉 ^ ^1.0 使人濕透^^1.00 溫泉浴場 ^^0.25 水池 0.25 池塘 0.25 池沼 0.25 蕩 ^^0.33 海域 0.33 水域 0.33 噗 ^^0.25 才能 0.25 品質 0.25 特質 0.25 優質 ^ ^1.0 治癒^^0.25 效果 0.25 效應 0.25 作用 0.25 成效 ^^0.50 皮膚 0.50 臚 ^^0.33 傳輸設備 0.33 核設施 0.33 生產設施 ^ ^1.0 消^^0.25 特別 0.25 特殊 0.25 特定 0.25 專用 ^^0.33 產品 0.33 製品 0.33 產物 ^
OF PLACE
^1.0 店^ ^1.0 旅館^^0.25 提供 0.25 設備 0.25 貸款 0.25 適應 ^ ^1.0 轉運^ ^1.0 日程^^0.25 計畫 0.25 規劃 0.25 方案 0.25 設計 ^.

Table 2.1 records the number of unique index terms averaged over 50 queries (after stopword removal) for different query types in our system. The Chinese representations (under heading 'Orig. Chinese') are longer than the English (under 'Orig. English) because of the use of single characters with short words. 'Dictionary' translation introduces multiple mappings for each English term and therefore produces even longer queries. TI, TQ, TQN denote queries using 'title', plus 'question', and additionally the 'narrative' sections of a topic. The latter two types are discussed in Section 5. VS, SO

	Before Translation		After Translation		
Type	**English**	**Chinese**	**MT**	**Dictionary**	**Mix Both**
TI	3.9	5.1	6.0	12.2	14.6
TQ	7.5	13.1	13.0	26.9	32.9
TQN	20.4	40.3	37.3	81.3	100.5
VS	23.8	27.3	35.0	79.6	94.0
SO	25.9	33.5	40.1	83.0	100.7
LO	35.1	55.2	58.3	124.2	151.3

Table 2.1: Sizes of Different Query Types

LO types correspond to TI, TQ and TQN queries respectively but include the 'concepts' section of a topic as well. This section introduces between 15 to 20 new English terms compared to not using it. The 'narrative' section used in the TQN and LO types also adds ten or more unique terms.

3 PIRCS Retrieval System and Collection Processing

Current approaches to text retrieval usually assign a retrieval status value (RSV) to each document d in a collection based on the properties of a given query q and those of the document and collection. Different systems employ different retrieval models to come up with useful RSV's. After this is done, the documents are ranked (i.e. sorted according to the RSV's) and presented to the user in an ordered fashion. Our PIRCS, acronym for Probabilistic Indexing and Retrieval – Components – System, employs a combination of two retrieval algorithms producing a document-focused RSVd and a query-focused RSVq for each document, and combined with a mixing parameter α. It is an extension of the probabilistic retrieval model and viewed as an activation spreading process in a network with learning (see [9] for greater details). Thus:

$$RSV(q,d) = \alpha * RSVd + (1-\alpha) * RSVq \quad (1)$$

with

$$RSVd = \Sigma_k \, S(qtf_k/L_q) * w_{dk} \quad (2a)$$

$$w_{dk} = \log [tf_k/(L_d - tf_k) * (Nw - L_d - F_k + tf_k)/(F_k - tf_k)] \quad (2b)$$

$$RSVq = \Sigma_k \, S(tf_k/L_d) * w_{qk} \quad (3a)$$

$$w_{qk} = \log [qtf_k/(L_q - qtf_k) * (Nw - F_k)/F_k] \quad (3b)$$

where tf_k, qtf_k are the frequency of term k in d and q respectively, $L_d = \Sigma_k \, tf_k$, $L_d = \Sigma_k \, qtf_k$ are the lengths of d and q, S is a sigmoid-like function, $F_k = \Sigma_{all\ doc} \, tf_k$ is the collection frequency of term k, and $Nw = \Sigma_k \, F_k$ is the number of tokens used in the collection.

Our approach considers every term of a document (or query) as a conceptual component self-relevant to the document (query) itself, and we work in a universe consisting of document components rather than documents. Because of the self-relevance assumption, every query (document) therefore has a relevant and irrelevant set even when no relevant judgment has been made, and we are able to bootstrap and provide probabilistic weights to our terms at the initial retrieval stage. Because we work with conceptual components, repeat term usage and query/document lengths are accounted for, enabling us to remove the binary assumption restriction in traditional probabilistic retrieval model [10]. The weight formula of Eqn. 3b is the familiar probabilistic query term weights but using components instead of whole document. Eqn. 2b is for document-focused retrieval and the form of the weighting, after taking the approximation Nw >> all other frequencies, turns out to be very similar to those used by [11] via a language model approach, but with a different smoothing coefficient.

Thus, our PIRCS system may also be viewed as a combination of the probabilistic retrieval model and a simple language model. It has been employed to do large scale IR experiments such as those run by TREC (see e.g. [12]) with consistently superior results. After translating queries from English to Chinese, crosslingual retrieval becomes monolingual Chinese retrieval if we do not need to worry about translating retrieved documents into English. This is true in these experiments.

Previous experience has shown that several methodologies are available to enhance CLIR such as pre-translation query expansion, two-stage retrieval with post-translation query expansion and collection enrichment. Pre-translation query expansion means expanding a given English query with highly associated terms based on the top retrieved documents from an appropriate English collection. We did not use this technique because the English collections that we have (from TREC) may not be appropriate with this new target collection and queries. Post-translation query expansion means performing a pseudo-relevance feedback using two-stage retrieval with the target Chinese collection and the translated Chinese queries. The first stage retrieval defines the top documents and best terms to be used to expand the Chinese queries. This is done as a default in our PIRCS system. The parameters used were n=40 top documents and m=100 best terms. Collection enrichment entails using an appropriate external Chinese collection during first stage retrieval. The idea is to improve the chance of getting more good documents into the top retrieved, and this might lead to better terms in those chosen as best. Again, we did not use this technique because the target collection is new and unfamiliar, and we are afraid that the available TREC Chinese collections may not be appropriate for enrichment purposes.

Chinese text needs to be segmented for retrieval purposes. We used short-words with characters as our default document and query representation method, since we have experience that it is both effective and efficient [13]. However, we differed from our previous work by using the translation dictionary (discussed in Section 2) as our segmentation dictionary. In that wordlist, we essentially keep all the Chinese terms that are four or less characters in length.

After text processing and indexing of the collection, the resultant dictionary generated has a size of 86,848. After setting Zipf thresholds of 3 and 20,000 to define statistical stopwords, the unique indexing terms remaining were 55,550.

A default mode of our system is to break long documents into approximately 550 character sub-documents ending on a paragraph boundary. We ended up with 176,924 sub-documents. We did not process the last batch of 'chi' documents that were not available when the task began, because we were pressed for time. We believe this batch would not influence results materially since its statistics is insignificant compared to the rest of the collection. Also it turns out there is only 1 relevant document in this batch for query #23 using relax assessment.

Monolingual retrieval was done as previously described. Crosslingual retrieval makes use of both the MT-software translated queries as well as the dictionary-mapped queries, each providing its own retrieval list. The RSV's of these two lists are combined using a ratio of 6:4 in favor of MT-software retrieval to produce the final result. This is done for all query types.

4 Results and Discussion

4.1 Monolingual Retrieval

We first look at our monolingual runs, which will form the basis from which crosslingual results will be measured. Table 4.1a shows our submitted results using the 'rigid' assessment method. The values in the '%' columns are percentage increases calculated using the TI values as basis. It is seen that even the short 'title only' queries provide very respectable results at a mean average precision (MAP) of 0.4653. The query types with the 'concepts' and other sections added: VS, SO LO all give much better MAP values of ~0.6. Interestingly, these MAP values are at similar levels to those done for the TREC 5&6 Chinese experiments [13].

	MAP	%	RR	%	P@10	%	P@20	%
TI	.4653		651		.4680		.3420	
VS	.6139	+32	652	+0	.5740	+23	.4100	+20
SO	.6037	+30	652	+0	.5660	+21	.4170	+22
LO	.5726	+23	652	+0	.5420	+16	.4000	+20

a) Rigid Assessment

	MAP	%	RR	%	P@10	%	P@20	%
TI	.5860		1561		.6820		.5930	
VS	.6998	+19	1568	+0	.7940	+16	.6780	+14
SO	.6936	+18	1568	+0	.7940	+16	.6790	+15
LO	.6806	+16	1567	+0	.7860	+15	.6640	+12

b) Relax Assessment

Table 4.1: Monolingual Retrieval Results (% increase is calculated using 'TI' as basis)

Best result is achieved with the VS version, which we defined as using 'title and concepts' sections only. For example, within 10 top retrieved documents one can expect on average over 5.7 relevant documents, or over 8.2 within the top 20. Usually, longer queries (such as the SO or LO versions) lead to better results, but it is not true here. Perhaps the concept sections have such precise and rich descriptions of the information needs that additional wordings only serve to add noise to the retrieval. The relevants retrieved (RR) at 1000 documents of 652 is 100% of those judged relevant using the 'rigid' assessment protocol. The performance of the four official monolingual runs are also displayed as precision recall curves in Fig. 4.1

For comparison, the overall best submitted monolingual MAP values are respectively: 0.4683, 0.6596, 0.6529 and 0.6486 for the TI, VS, SO and LO query types using 'rigid' assessment. Except for the LO type, we are within 8% of these best values. For the LO type, our MAP value is off the best value by nearly 12%.

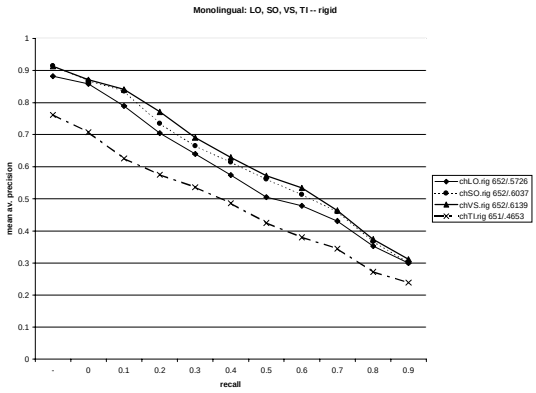

Figure 4.1: Recall-Precison Curve for Monolingual ---- Rigid Assessment

Table 4.1b shows the same runs using the 'relax' assessment. It is even better in all measures except the relevants retrieved which vary between 99% to close to 100% of the 1571 judged relevant. We believe the 'rigid' assessment is probably more realistic. It turns out that the pseudo-relevance feedback parameters for these runs were not set well. Improved monolingual results with better parameters are discussed in Section 5.

4.2 Crosslingual Retrieval

Our automatic crosslingual results are tabulated in Table 4.2. The percentage columns are calculated with respect to the corresponding monolingual values of Table 4.1 for evaluation purposes. Except for the 'title only' queries, the submitted runs are able to achieve ~80% in MAP values, ~95% of relevants retrieved and > 80% of precision at ten and twenty documents when compared to monolingual effectiveness using 'rigid' assessment. 'Title only' queries give precision values between 55 to 60% of monolingual, and 88%of relevants retrieved. Similar values of 'relax' assessment are shown in Table 4.2b.

	MAP	%	RR	%	P@10	%	P@20	%
TI	.2554	55	572	88	.2680	57	.2050	60
VS	.4724	77	627	96	.4580	80	.3450	84
SO	.4774	79	626	96	.4580	81	.3490	84
LO	.4733	83	628	96	.4660	86	.3440	86

a) Rigid Assessment

	MAP	%	RR	%	P@10	%	P@20	%
TI	.3164	54	1329	85	.3380	50	.3380	57
VS	.5483	78	1478	94	.5570	70	.5570	82
SO	.5522	80	1475	94	.5580	70	.5580	82
LO	.5499	81	1459	93	.5490	70	.5490	83

b) Relax Assessment

Table 4.2: Crosslingual Retrieval Results (% are calculated using Table 4.1 corresponding monolingual values as basis)

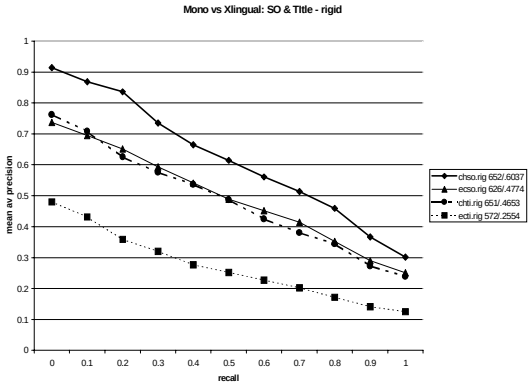

Figure 4.2: Monolingual vs Crosslingual: 'SO' and 'TI' ---- Rigid Assessment

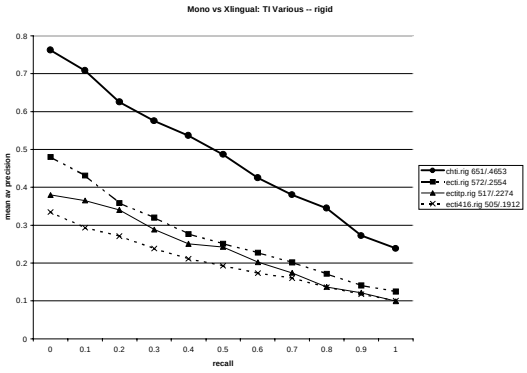

Figure 4.3: MT, Dictionary and Combination P-R curves for 'TI' Queries ---- Rigid Assessment

The percentage of monolingual effectiveness is similar to the 'rigid' assessment, except for the precision at 10 documents where the percentage is 70% vs 80%. These crosslingual MAP are state-of-the-art values.

The best absolute MAP value of 0.4774 is attained with the SO query type, while for monolingual it is the VS type. We conjecture that this is because of the 'question' section in the SO type that provides extra redundancy to hedge for errors during query translation. A MAP of ~0.47 for the longer query types are quite reasonable for this difficult task. The precision at 10 and 20 top documents retrieved show that one can expect on average close to 4.6 and 7 relevant documents respectively. These numbers are somewhat less than those for monolingual, but still quite respectable from a utility point of view. A comparison between monolingual and crosslingual precision-recall curves are shown in Fig.4.2. The two solid curves are for SO and the dotted curves for TI queries. It is observed that crosslingual SO queries achieve much closer performance to its monolingual counterpart than TI queries. We again attribute this to more redundancy and better context for translation with longer query types. LO and VS queries produce precision recall curves almost indistinguishable from this SO curve. Fig.4.3 shows the TI performance broken up into MT software and dictionary contribution. In general, MT software is somewhat better in these experiments.

4.3 Failure Analysis for 'Title' Queries

The results for the 'title only' queries (MAP 0.2554) are not good, achieving only about 55% of monolingual MAP value. These average to less than four English content terms per query (Table 2.1). Unfortunately, it is well known that most users generally prefer to issue short queries (in the web, for

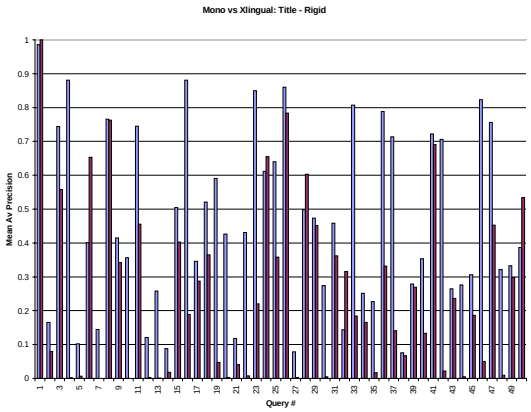

Figure 4.4: Monolingual vs Crosslingual per query: 'TI' ---- Rigid Assessment

example, searching with one or two words is common). These results demonstrate that users will be disappointed issuing short queries for CLIR. 'Relax' assessment provides better absolute effectiveness (MAP 0.3164), but still in the 54% level compared to monolingual. (In a sense, even humans may misunderstand intentions of a few words, not to mention machines.) Fig.4.4 shows individual TI query performance compared with

monolingual. There are 6 topics (#1, 6, 24, 28, 32, 50) better than monolingual, but 44 worse. Two topics (#7, 10) have zero MAP, and 15 have MAP values <0.1 while their monolingual values are >0.1. These are topics #2, 4, 5, 12, 13, 19, 20, 21, 22, 30, 35, 42, 44, 46 and 48. We assume that these queries fail because of translation problems. What follows is an attempt to analyze why they perform badly.

Query 7 failed because 'Carter' was translated into: 运货马车夫 'the worker who carts', while query 10 needs a special Chinese term 库藏股 for 'Treasury stock' and is lacking in our translation approaches. #12 (Michael Jordan), 30 (El Nino), 44 (Hua-shan), 46 (Ma Yo-yo) are related to unknown proper nouns or transliteration. #21 has IC in English that was not present in the original Chinese topic. IC was translated as '积体电路 or 集成电路 (Integrated Circuits)'. This constitutes noise. #2 and 5 have reasonable translations but performed badly. The rest have genuine bad or no translations for some words such as: #4 'commercials' has no or bad translation, 35 'diet product' mapped to '国会产品 ('diet' in the Japanese Diet sense), or 规定饮食产品', #42 'millennium' has a special slang 千禧年 that was missed. These 17 queries account for about 56% of the difference between 'title only' monolingual and crosslingual MAP result.

There are other queries like #3 (White Horror), 6 (Kosovar), 15 (Bai Xiao-yan), 27 (Meinung), 28 (Chilan), 33 (Bai-feng), 34 (Viagra), 38 (Chunghwa, ROCSTAT), 47 (Jin Yong) that involve un-translated proper nouns or transliterations. However, their MAP values are >0.1 and not dismal. In fact, #6 and 28 performs better than monolingual! Queries #23 (Disneyland), 39 (Leonid), 45 (Cloud Gate), and several queries contain (Taiwan). These were properly translated. Two more queries #13 (NBA) and #43 (CIH) have abbreviations that also failed translation. However, the two original Chinese queries also did not use them. Thus, these 50 topics seem to have a realistic proportion of proper nouns. Generally, these terms lead to good results for monolingual retrieval because they are unambiguous, but their translation failures cause the corresponding crosslingual queries to behave poorly -- at only ~55%. However, their longer counterparts VS, SO or LO query types have sufficient redundant descriptions and manage to restore the performance to ~80%. This points to the importance of using longer query descriptions for CLIR.

5 Additional Experiments

5.1 Improving Monolingual Results

As discussed in Section 4.1, our monolingual blind experiments, though good, are less than the top runs by about 8% and more for the LO queries. After results were known, we discovered that the parameters that we set for pseudo-relevance feedback (prf) of 40 top documents and 100 best terms are good for crosslingual but sub-optimal for monolingual. After adjusting to using 10 top documents 100 terms, the result improved as shown in Table 5.1.

	prf: 10d100t				**prf: 40d100t**		
	MAP	%	P@10	P@20	MAP	P@10	P@20
TI	.4853	+4	.4500	.3450	.4653	.4680	.3420
TQ	.5179	+5	.5020	.3760	.4943	.4860	.3750
TQN	.5450	+7	.4840	.3780	.5092	.4760	.3710
VS	.6384	+4	.6020	.4230	.6139	.5740	.4100
SO	.6271	+4	.5720	.4240	.6037	.5660	.4170
LO	.6214	+9	.6000	.4170	.5726	.5420	.4000

Table 5.1: Comparing Monolingual Results – New and Old Parameters (Rigid Assessment)

The values based on new parameters are shown at the left with the MAP percentage improvements over the old values shown on the right. These new values for the official query types are within about +/- 4% of the overall best monolingual results submitted. The VS query type still performs best, but the long LO type improves substantially over our submitted values. The RR values remain as before.

Table 5.1 also shows two new runs for the TQ and TQN queries. As pointed out before, TI, TQ and TQN correspond to VS, SO and LO types minus the 'concepts' section. The use of 'question' and 'narrative' sections in TQ and TQN improves results over using 'title' only, but still far from the effectiveness returned by incorporating the 'concepts' section.

5.2 Cross Language IR with 'Title', 'Question' and 'Narrative' Sections

The 'concepts' section in a topic contains many rich and precise phrases and wordings concerning the information needs. This is very good for both monolingual and crosslingual IR results but is unrealistic in real-life situations. It is difficult for normal users to supply these terms. On the other hand, 'title' only queries perform poorly. We therefore investigated whether adding the 'question' and 'narrative' part of a topic to the 'title' section (forming TQ and TQN queries) can lead to better performance. The 'question' part is a natural language statement of a user's needs, while the 'narrative' can be regarded as further exposition in free text. They should be easier for a user to compose than the 'concepts'. We translate as before and perform retrieval with them. These are shown in Table 5.2.

It is seen that he TQ queries get between 8 to 20% better precision values than 'title' only, while TQN

	MAP	%	RR	%	P@10	%	P@20	%
TI	.2554		572		.2680		.2050	
TQ	.2968	+16	598	+5	.2900	+8	.2470	+20
TQN	.4105	+61	620	+8	.3960	+48	.3080	+50

Table 5.2: Comparing Crosslingual TI with TQ, and TQN Queries – Rigid Assessment

queries lead to between 48 to 61% improvements. The latter has a MAP value of 0.41 and nearly 4 documents out of the top 10 retrieved are relevant. These are good performance. Thus, just adding more related text can bring us much closer to the 0.47 MAP values for queries with 'concepts' section (Table 4.1). Number of relevants retrieved, RR, also improves. This illustrates again that longer descriptions of needs should be encouraged for CLIR. The wordings need not be high precision conceptual terms.

5.3 Mixing Two Translations as One Query

In our submitted experiments, the two translation outputs – MT-software and bilingual wordlist lookup with disambiguation – were employed individually for retrieval, and then their retrieval lists are combined. An alternative is to combine both outputs into one single longer query, and perform retrieval with it. This makes sense in that if one translation method is faulty (missing or wrong), the other method may have the correct mappings and thus remedy the query to a certain extent. If both methods confirm the same translation for an English term, the Chinese output will attain double weight compared to other translations that do not agree. Moreover, one needs to do a single retrieval only and is more efficient. We call this process 'mixing translations' into one query. The result is shown in Table 5.3. Except for TI and TQN types, mixing translations can lead to 2-5% better results for all other query types when compared to Tables 4.2a and 5.2 where combination of retrieval lists is employed. TI has MAP value decrease slightly by half a percent, but TQN decreases substantially by 7%. It needs further investigation to see why mixing translations for TQN queries is not good.

	MAP	**%**	**RR**	**%**	**P@10**	**%**	**P@20**	**%**
TI	.2544	52	578	89	.2580	57	.2060	60
TQ	.3107	60	602	92	.3160	63	.2490	66
TQN	.3814	70	617	95	.3820	79	.2850	75
VS	.4872	76	646	99	.4720	78	.3610	85
SO	.4999	80	645	99	.4920	86	.3700	87
LO	.4835	79	642	98	.4920	82	.3560	85

Table 5.3: CLIR Results using Mixing of Translations in One Query ---- Rigid Assessment (% calculated using Table 5.1 corresponding monolingual values as basis)

Comparing with the improved monolingual results of Table 5.1, 'title' only queries still operate at a disappointing 52% of monolingual. Precisions at 10, 20 documents however are better at ~60%. On average >2.5 documents among the top 10, or 4 among top 20 retrieved are relevant. Lengthening the queries with the 'question' section (TQ) improves monolingual comparison to 60%, and to 70% when 'narrative' section is also used (TQN). There would be on average over 3 and nearly 5 relevant documents among the top 10 and 20 retrieved respectively if one employs the 'title' and 'question' sections as queries (TQ). Queries with 'concepts' operate at 76 to 80% of monolingual. Plots of crosslingual performance at different (English) query sizes are shown in Fig. 5.1.

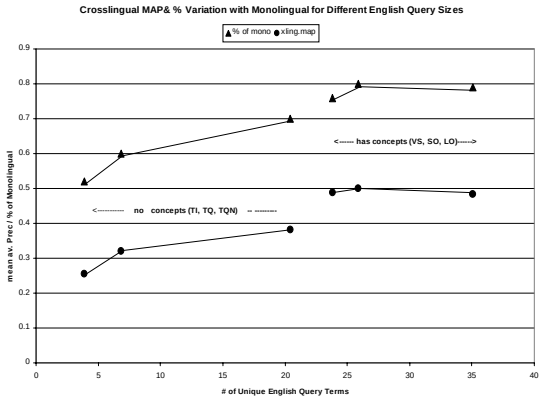

Figure 5.1. Crosslingual Performance at Different English Query sizes

Essentially the plots summarize what we observed: when queries contain the 'concepts' section (right side of Fig. 5.1), adding the 'question' section (SO) helps, but further adding the 'narrative' section (LO) leads to worse performance. For queries without 'concepts' (left side), performance improves monotonically with more free texts (TQ and TQN).

6 Conclusion and Future Work

This NTCIR-2 Chinese and cross language experiments make use of ~200MB of texts and 50 topics. All topics have at least four relevant answer documents in the target collection. A reasonable number of the topics also include proper nouns, transliterations that are difficult to translate. The monolingual mean average precision effectiveness was high: at over 0.6 for long queries, and at about 0.47 even for queries involving only a few words of the title section of a topic.

Cross language IR leads to about 0.52-0.55% of monolingual MAP value for 'title' only queries. We attribute this low effectiveness to queries involving

proper nouns and transliterations. These accentuate the gap because their un-ambiguity gives good monolingual results while their translation failure leads to poor crosslingual values. Lengthening the queries with various amount of related free texts can enhance precision to 0.38-0.4 or 70% of monolingual. Adding highly precise and related concept terms improve crosslingual MAP results further to nearly 0.5 or 76 to 80% of monolingual. Long query descriptions are recommended for English-Chinese CLIR.

MT-software and dictionary lookup translation with disambiguation techniques seem to be able to complement each other. Our approach of linearly combining their retrieval lists appears to work well. A more efficient strategy is to mix translations into one query. It is shown that except for the TQN query type, it can lead to about equal or improved results compared to linear combination of retrieval lists.

One cannot do crosslingual retrieval without properly translating the queries (or the documents). Proper noun and transliteration translation are therefore crucial for CLIR. Expanding coverage of the LDC bilingual wordlist or improving the weighting of translation terms may also lead to better CLIR performance. These are some of the topics we intend to investigate in the future.

Acknowledgments

This work was partially supported by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-1-8912. Norbert Dinstl helped formatting the queries and collections.

References

[1] G. Grefenstette. Cross language Information Retrieval. Kluwer, 1998.

[2] D. Oard & A. Diekema. Cross-language information retrieval. In: Annual Review of Information Science and Technology, 33: 223-256, 1998.

[3] J. Xu & R. Weischedel. TREC-9 cross-lingual retrieval at BBN. Preliminary paper at TREC-9 Conference, Gaithersburg, MD, Nov, 2000.

[4] J. Gao, J-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou & C. Huang. TREC-9 CLIR experiments at MSRCN. Preliminary paper at TREC-9 Conference, Gaithersburg, MD, Nov, 2000.

[5] K.L. Kwok, L. Grunfeld, N. Dinstl & M. Chan. TREC-9 cross-lingual, web and question-answering track experiments using PIRCS (Draft). Preliminary paper at TREC-9 Conference, Gaithersburg, MD, Nov, 2000.

[6] K.L. Kwok. English-Chinese cross language retrieval based on a translation package. *In: MT Summit VII Workshop: Machine Translation for Cross Language IR.*, 8-13, 1999.

[7] K.L. Kwok. Exploiting a Chinese-English biligual wordlist for English-Chinese cross language information retrieval. Proc. of 5th Intl.Workshop on Information Retrieval with Asian Languages IRAL2000, 173-179, 2000.

[8] A. Pirkola, H. Keskustalo & K. Jarvelin. The effects of conjunction, facet structure and dictionary combinations concept-based cross-language retrieval. *Information Retrieval*, 1(3): 217-250, 1999.

[9] K.L. Kwok. Improving English and Chinese ad-hoc retrieval: a Tipster Text Phase 3 project report. *Information Retrieval*, 3(4): 313-338, 2000.

[10] S.E. Robertson & K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27: 129-146, 1976.

[11] D. Hiemstra & W. Kraaj. Twenty-One at TREC-7:ad-hoc and cross language track. In: *Information Technology: The Seventh Text Retrieval Conference (TREC-7).* E.M.Voorhees & D.K. Harman, (eds.), NIST Special Publication 500-242, GPO: Washington, D.C, 227-238, 1999

[12] E.M. Voorhees, & D.K. Harman. Overview of the Eigth Text REtrieval Conference (TREC-8). In: *Information Technology: The Eigth Text REtrieval Conference (TREC-8).* E.M.Voorhees & D.K. Harman, (eds.), NIST SP 500-240, pp.1-24. GPO: Washington, D.C.

[13] K.L. Kwok. Employing multiple representations for Chinese information retrieval. *Journal of the American Society for Information Science*, 50(8): 709-723, 1999.