NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments Using PIRCS

Kui-Lam Kwok, Norbert Dinstl and Sora Choi

Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
kwok@ir.cs.qc.edu, emc21@earthlink.net, sorac@hotmail.com

Abstract

In NTCIR-4 we participated in Korean, Chinese, English monolingual, Chinese-English, English-Korean bilingual, and Chinese-Korean cross language (using English as pivot) retrieval tasks based on our PIRCS retrieval system. The query translation approach was employed for CLIR. We combined two MT translations for Chinese-English, and two for English-Korean. For the latter, a web-based entity-oriented translation procedure was also used to translate un-translated OOV terms. Concatenation of MT output was found to lead to better CLIR effectiveness than single MT, while entity translation brings further improvements of about 15%. The English-Korean and Chinese-English CLIR achieve between 71% and 88% of monolingual, and appear to be best among submissions. For Korean-Korean and English-Korean retrievals, bigram indexing performs better than word indexing, and combination of the two provides better results, in most cases. Chinese-Korean retrieval via English as pivot language provide results with mean average precision between 56% and 66% of Korean monolingual. All submissions are automatic runs.

Keywords: *direct bilingual CLIR; pivot CLIR; translation concatenation.*

1 Introduction

Participants in CLIR tasks need to experiment with more than two languages in NTCIR-4. We took this opportunity to add Korean (K) as the third language to our PIRCS [1] retrieval system's usual English (E) and Chinese (C) capability. Usage of these three languages is diagrammed in Fig.1 below to show the tasks that we have done and submitted. Our convention is to denote the query language via the notation Q^{ABC}: meaning that the final language is C and it has been derived through query translation from source language A via a pivot language B. Direct query translation is denoted as Q^{AB}, for example. The last superscript language character always indicates what collection language this query would operate on.

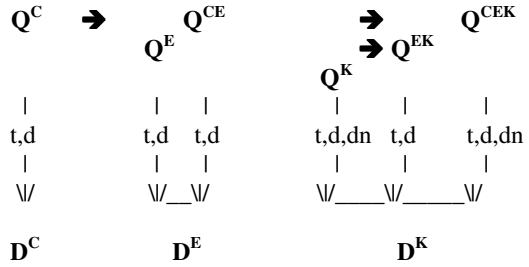

Fig.1: Diagram of Submitted Runs (Q = query; D = collection; superscript = language; ➔ = translation; \|/ = retrieval)

There were a total of fourteen runs. These include retrievals with Chinese target collections named as:

pircs-C-C-T-01 pircs-C-C-D-02;

retrievals with English target collections named as:

pircs-E-E-T-01 pircs-E-E-D-02
pircs-C-E-T-01 pircs-C-E-D-02;

and retrievals with Korean target collections named as:

pircs-K-K-T-01 pircs-K-K-D-02
pircs-K-K-DN-03 pircs-E-K-T-01
pircs-E-K-D-02 pircs-C-K-T-01
pircs-C-K-D-02 pircs-C-K-DN-03

The pircs-C-K-x-yy CLIR experiments employ the Q^{CEK} queries with transitive translations. All retrievals include using the title or the description sections of the topics provided, and a couple also use the description plus narrative section. In addition to extending our support for Korean retrieval, our goal is also to see how well English can play as a pivot language between Asian languages. Sections 2 and 3 describe our query translation and Korean processing resources. Sections 4, 5 and 6 discuss our results for retrieval with Korean, English and Chinese collections respectively using Chinese, English and Korean queries. Section 7 contains our conclusions.

2 Translation Resources

The most important tools for cross language tasks are translation resources. We continue to employ the

efficient query translation approach. Resources are needed to translate from Chinese to English, English to Korean and Chinese to Korean. The latter however seems not available easily (in the U.S.). Both Chinese to English and English to Korean translation are new to us. These are considered major languages. While English/Chinese dictionaries are available in the U.S. such as from the LDC (although their time currency is questionable), English/Korean dictionary is not easy to obtain. We decide to use commercially available MT software for this purpose. We assume that they will provide reasonable translation for general English, but may not be sufficient for entity or terminology words. We augment the result with an entity/terminology-oriented web-based translation methodology that was being developing. Our goal in NTCIR-4 is to see whether concatenating multiple MT outputs for query translation works better than single MT.

2.1 Chinese-English MT Software

For Chinese-English translation, Systran [2] and Loto [3] software packages were used. Systran has a long history of C-E translation. Loto is a product newly marketed in America; it evolved from the HuaJian English-Chinese MT software in China. A license to Loto allows one to have a stand-alone MT package on a PC, as well as web access to their company's central translation software. The latter is advertised to get updated regularly to provide better translation than stand-alone versions. We used the online translation facility for these experiments.

Our hypothesis is that combination of MT translation can bring more robust results. Given a query, two separate translations are performed and the results are concatenated together. If a source word/phrase leads to the same (duplicate) target translations, they may be regarded as 'confirmed' correct and are automatically weighted heavier. When translations differ, there is also possibility that they provide different wordings for the same source concept and therefore may hedge against insufficient coverage. In the case of English-Korean, one MT may provide semantic translation while the other may output transliterations for example. The trade off is that when both were wrong, we end up with twice as much noise.

The following shows example output of typical Chinese-English translations of the description section of a topic for readers to judge their quality. Included at the end are six additional terms that are obtained from pre-translation expansion processing (see Section 4.3). In general, there are both translation successes and failures. Except for entity names, the output appears acceptable for CLIR, both from the view of segmentation and translation.

qry#55 Original Chinese:
亞洲各國對北韓發射大浦洞（Daepodong）飛彈反應。
試射　彈道　和南　射程　防衛廳　飛越

qry#55 English Translation via Systran:
The Asian various countries launches the big water's edge hole to North Korea (Daepodong) the guided missile response.
Test fire　Trajectory　With south　Firing distance　Japanese Defense Agency　Leap

qry#55 English Translation via Loto:
Asian various countries launch the big Pu hole to Korea (Daepodong) The stray bullet reacts.
Trial fire　Trajectory　With the south　Range　Defence office　Fly over

2.2 English-Korean MT Software

In the U.S., resources for Korean language are not as common as other major languages. For English-Korean, we employed the English to CJK capability of Systran. Another package called English Guide (EnGuide) [4] from LniSoft was also acquired from Korea. The latter has user interface in Korean only, and is therefore not suitable for users who do not understand Korean. It also has difficulty handling sentences having words with capitalized first letter in the middle of a sentence (which is the case with the title section of the queries in English). We overcome this problem by producing two versions of the title section: one with the title all in lower case, and the other with the case information retained but put each word into a separate line.

Some examples of English-Korean translations are shown below. Included are un-translated English words that are picked up by our entity-oriented translation procedure (Section 2.3).

qry#2 Original English:
Find out who joined the Jonnie Walker Charity Golf Tournament in Taiwan in 1999 and the related activities.

qry#2 Korean Translation via Systran:
Jonnie 보행자 자선 골프 경기 대회를 1999 년에 대만과 관계있는 활동 결합했다 까 누구안에 알아내십시요.
Jonnie　조니 46 워커 16 로스 19

qry#2 Korean Translation via LniSoft:
누가 1999 년과 관계 있는 활동에 대만에 Jonnie Walker Charity 골프 Tournament 와 합류하였던지를 아십시오.
Tournament　토너먼트 30 대회 82 .

In addition, the Chinese-English translations in Section 2.1 are fed directly into the English-Korean MT software to provide four mappings between Chinese-Korean via transitive translation. The output for query #55 used in Section 2.1 are shown below:

qry#55 Systran English pivot, Korean Translation via Systran:
아시아 각종 나라는 북한 Daepodong 에 큰 물
가장자리 구멍을 유도 미사일 응답 발사한다.
시험 불
탄도
남쪽에
발포 거리
일본 방위 기관
뛈

qry#55 Systran English pivot, Korean Translation via LniSoft:
아시아 여러가지 국가들은 유도탄 대답인
북한 Daepodong 으로 큰 물의 경계 곤경을
날립니다.
테스트는 불이 붙습니다.
탄도 .
남쪽과 .
불이 붙고 있는 거리 .
일본 Defense 정부 기관.
Defense 방어 121 체제 55
뛰십시오.

qry#55 Loto English pivot, Korean Translation via Systran:
아시아 각종 나라 발사 한국에 무거운 구멍 Pu
Daepodong The 무숙자 탄알은 반작용한다 .
Pu 폴리우레탄 41 접착제 27
예심 불
탄도
남쪽에
범위
방위 사무실
넘어서의 비행거리

qry#55 Loto English pivot, Korean Translation via LniSoft:
아시아 여러가지 국가들은 한국으로 무거운 구멍
Pu 을 내보냅니다 Daepodong The 길잃은 가축
총알은 반작용합니다.
Pu 폴리우레탄 41 접착제 27
시험은 불이 붙습니다.
탄도 .
남쪽과 .
줄지으십시오.
방어 사무실 .
비행하십시오.

2.3 Web-based Entity Translation

Our assumption is that MT software can provide reasonably translation for general language expressions but may not be sufficient for entities such as names of person or places, etc. We implemented a web-based translation from English to Korean (and Chinese) that is oriented to entity names and terminology [5]. It is based on the normal convention of writers to express translations in bilingual document fragments in the following form: .. kkkkk (eeeee).. or ..eeeee (kkkkk).., where kkkkk and eeeee are Korean and English strings respectively. When either of such patterns is encountered, it is quite likely that kkkkk will contain some kind of translation of eeeee, or vice versa. We search the web using an English term as key and request output snippets in Korean. These snippets are searched for the pattern above, and candidates for translation are isolated after some text processing and noise filtering.

This procedure was employed in E-K CLIR to translate any OOV English terms that remain after the two MT software operations. Examples of these translations are also shown in Section 2.2. Consider qry#2 via Systran: translations for 'Jonnie' were picked up with the indicated occurrence frequency in the returned web snippets. This was not performed for qry#2 via LniSoft because additional English words are adjacent to 'Jonnie'. Our procedure regards such a word sequence as an indivisible phrase to gain precision, and try to locate its translation on the web. Apparently it failed. In qry#55, the translation for 'Daepodong' was also not found by our procedure.

3 Korean Text Processing and Indexing

Korean text is written with blank space as delimiter, but the string in between can be words, compounds or phrases [6]. For all tasks involving Korean, we employed a simple strategy of overlapping bigram indexing on the original texts without stemming or stopword removal as a default. In addition, we used a program called HAM version 6.0.0 [7] for the E-K and C-K retrieval tasks. HAM is an acronym for Hangul Analysis Module (or Model). It is a Korean lexical analyzer for Hangul text. It supports an 'index' program which removes suffixes and stopwords and extracts simple words from compounds. For our indexing purposes, the simple nouns, the original compounds and the stemmed verbs, etc. are kept. Compounds are retained because we can have some phrase indexing and also like to hedge concerning the outcome of segmentation. We call this HAM indexing.

4 Retrieval with Korean Collections:

4.1 K-K Monolingual Retrieval

Eight submissions using the Korean collection as retrieval target were submitted. Three were monolingual using title **(T)**, description **(D)** and description with narrative **(DN)** sections of a topic to form Korean queries. These serve as basis for evaluating other cross language retrievals with the Korean documents. Table 1 shows their results for the measures: R% (percent of average recall after 1000 retrieved), MAP (mean average precision), P10, P20 (average precision-at-10 and -20 documents retrieved) and R.Pre (average precision at the exact number of relevant documents for a query). Values of **submitted runs are bolded** in all tables. Rows with a * denote un-submitted runs. A 'b', 'w' or 'bw' following a run id denotes bigram, HAM indexing, or combination of these two retrieval lists

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
K-K-T-01 b	**90**	**.4934**	**.6298**	**.5842**	**.4953**
*K-K-T-01 w	85	.4419	.6018	.5500	.4601
*K-K-T-01 bw	91	.4860	.6263	.5921	.4916
Description Queries					
K-K-D-02 b	**83**	**.4049**	**.5561**	**.5044**	**.4225**
*K-K-D-02 w	79	.3828	5281	.4833	.3879
*K-K-D-02 bw	84	.4187	5667	.5149	.4298
Description + Narrative Queries					
K-K-DN-03 b	**92**	**.5161**	**.6807**	**.6184**	**.5012**

a) Relax Assessment (number relevant = 3917)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
K-K-T-01 b	**92**	**.4588**	**.5386**	**.5044**	**.4678**
*K-K-T-01 w	87	.4112	.5140	.4754	.4377
*K-K-T-01 bw	93	.4515	.5404	.5044	.4617
Description Queries					
K-K-D-02 b	**85**	**.3777**	**.4877**	**.4421**	**.3925**
*K-K-D-02 w	81	.3548	.4439	.4123	.3622
*K-K-D-02 bw	86	.3904	.4860	.4439	.3955
Description + Narrative Queries					
K-K-DN-03 b	**94**	**.4848**	**.6070**	**.5456**	**.4743**

b) Rigid Assessment (number relevant = 3131)

Table 1a,b: Monolingual Korean Results for 57 Query Types T, D, DN.

Table 1 shows that monolingual Korean results have good MAP values (> 0.4) except in the case of D queries using rigid assessment (.3777). These queries are probably comparatively easy for the target collection. Average precision-at-10 for relax judgment range from 0.5561 to 0.6801. Queries of long (DN) type have better performance followed by short title (T) queries. D queries surprisingly perform some 18% worse than T type (MAP .4049 vs .4934, and the improvement is significant at the 5% level using sign test). One general observation for the majority of our submitted runs is that D queries have worse MAP values than T queries for both Korean and English collections. This may be due to the fact that the short titles (of topic) have specific words and phrases only (e.g. qry#24: Illegal Tapping, Violation, Privacy), while the descriptions (of topics) are grammatical sentences often with only functional words added (qry#24: searching for documents dealing with the violation of people's privacy due to illegal tapping.)

Our submitted monolingual Korean retrieval makes use of bigram representation only. Table 1 shows also post-relevance-judgment runs using HAM indexing (stemming and stop-word removal) listed as: *pircs-K-K-T-01w and *pircs-K-K-D-02w. They are inferior to simple bigram indexing. Adding the bigram and word retrieval lists result in runs indicated by tag bw. They are not much different from the bigram only results. Our official description query MAP (rigid) value, though ranked 4th, is nearly 20% less than the best value submitted.

4.2 E-K Crosslingual Retrieval

Table 2 shows results of English-Korean cross language retrieval. As discussed in Section 2, an English query was translated to Korean by both Systran and EnGuide. No pre-translation query expansion was employed, unlike [8]. Output from both was concatenated into a single query. This further went through our web-based translation to minimize the number of un-translated English terms. The resultant queries were indexed in two ways: directly via bigrams (b); and via stems produced by HAM (w). This would allow us to compare HAM indexing with bigrams. Our submissions pircs-E-K-T-01 and pircs-E-K-D-02 are however combination of retrieval lists from the two indexing schemes.

Table 2 shows that for E-K the MAP difference between T and D queries are small (.3598 vs .3566 relax) unlike K-K monolingual. Worth noting is that the precisions at 10 and 20 for D queries are about 10% better than for T (e.g. .5123 vs .4614). Apparently, translation of the longer English D queries behaves on average similarly to T queries, but can lead to translations more suitable for low-recall retrieval.

The E-K MAP values appear to be the best achieved among all submissions. Compared to K-K monolingual retrieval, these crosslingual precision values attained 73% (T: .3598 vs .4934) and 88% (D: .3566 vs .4049) of relax effectiveness. The same comparisons for rigid assessment give: 73% (.3331 vs. .4588) and 86% (.3249 vs. .3777) respectively. These percentages are high because our K-K MAP values are relatively low.

The un-submitted D run tagged 'no web' in Table 2 means no web entity translation was performed and can be compared with the submitted D run. This process has led to over 15% improvement (0.3064 vs. 0.3566 relax, significant at 5% level using sign test).

The * rows tagged 'b' and 'w' in Table 2 show

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-K-T-01 bw	**79**	**.3598**	**.4614**	**.4342**	**.3752**
***E-K-T-01 b**	78	.3578	.4474	.4386	.3760
***E-K-T-01 w**	73	.3342	.4140	.4000	.3461
Description Queries					
E-K-D-02 bw	**79**	**.3566**	**.5123**	**.4737**	**.3762**
***E-K-D-02 bw no web**	76	.3064	.4772	.4333	.3278
***E-K-D-02 b**	76	.3388	.4737	.4588	.3687
***E-K-D-02 w**	76	.3154	.4526	.4211	.3398
***E-K-D-b-sys**	69	.2958	.4246	.3825	.3243
***E-K-D-b-gui**	67	.2581	.3702	.35	.2794

a) Relax Assessment (number relevant = 3917)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-K-T-01 bw	**80**	**.3331**	**.3982**	**.3781**	**.3497**
***E-K-T-01 b**	80	.3357	.4018	.3825	.3490
***E-K-T-01 w**	73	.3085	.3474	.3430	.3174
Description Queries					
E-K-D-02 bw	**80**	**.3249**	**.4456**	**.4035**	**.3507**
***E-K-D-02 bw no web**	77	.2756	.4105	.3605	.3000
***E-K-D-02 b**	77	.3118	.4105	.3939	.3387
***E-K-D-02 w**	77	.2891	.3842	.3526	.3131
***E-K-D-b-sys**	69	.2755	.3719	.3316	.2927
***E-K-D-b-gui**	68	.2347	.3193	.2956	.2579

b) Rigid Assessment (number relevant = 3131)

Table 2a,b: E-K Crosslingual Results for 57 Query Types T, D.

un-submitted results of using bigram and HAM indexing scheme alone. The latter returns slightly worse MAP values than pure bigram: e.g. 0.3342 vs 0.3578 (relax) for T queries, and 0.3118 vs 0.2891 (rigid) for D queries. Combining the two retrieval lists (our submitted results) improves over both individually more often, unlike K-K runs.

In Table 2, we also show two bigram D runs that use either Systran (pircs-E-K-D-b-sys) or EnGuide (pircs-E-K-D-b-gui) translations only. These results are inferior (e.g., MAP rigid for Systran is 0.2755, for EnGuide is 0.2347, compared to 0.3118 for pircs-E-K-D-02b where both translations were concatenated. Sign tests are significant at the 5% level for these improvements). This appears to support our assumption that MT combination leads to better effectiveness compared to using them singly.

We investigated why EnGuide results are inferior to Systran for description queries using bigram indexing. Part of the reason seems to be that entity names (like query #2: 'Jonnie Walker Charity Golf Tournament') in English queries are capitalized, and EnGuide has problem with them. Systran however is more flexible in regard to Ascii case and often provides the correct translation.

When title queries from our submitted E-K run are compared with the submitted K-K run, only 12 have better MAP (rigid) values, an overwhelming 45 perform worse. Of these 45, 12 have MAP differences > 0.3 and they add up to ~80% of the total difference 0.1256 (i.e. 0.4588-0.3331). The faults of these 12 range from reasonable translation but synonyms of the Korean query words (e.g. #37 Starvation (기근) 기아, 아사; #2 Charity (자선) 자비;), ambiguity or transliteration (e.g #17,36 AIDS (에이즈) 후천성 면역 결핍증; #34 Tokyo (도쿄) 도오쿄; #43 Derivative (파생상품) 미분, 함수, 제어, 시간), bad translation and/or noise (e.g. #24 Illegal tapping, Violation, Privacy (불법감청, 불법침해, 사생활) 불법에게 두드림의 위반, 기밀, #28 law (법) 법과대학).

4.3 C-K Crosslingual Retrieval via English as Pivot

Results of our C-K retrieval using Q^{CEK} transitive translation queries are tabulated in Table 3. Here, the Chinese queries (T and D only) first underwent a pre-translation expansion using the Chinese collections. (For DN queries, we assume they are sufficiently long that pre-translation would not have much effect.) We employ the top 10 documents of an initial retrieval and added conservatively only 6 terms to each query. The queries were translated two ways into English using Systran and Loto packages as discussed in Section 2. The English output were further translated into Korean by Systran and EnGuide, resulting in four Korean mappings for each query. Any English terms left un-translated were processed by our web-based translation. The final queries were then indexed two ways bigram (b) and HAM indexing (w) as in E-K. The submitted results use combination of retrieval lists from (b) and (w) runs.

An error was later discovered in the PRF process for the description pircs-C-K-D-02 run, which is tagged with 'e'. (The number of feedback documents was random for this run.) The next row *C-K-D-02 bw without error tag 'e' tabulates the corrected values. It is about 5-6% better.

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-K-T-01 bw	**76**	**.2783**	**.4228**	**.3728**	**.3022**
***C-K-T-01 b**	66	.2448	.3526	.3263	.2690
***C-K-T-01 w**	75	.2722	.4105	.3737	.2956
***C-K-T-01 bsys**	72	.2706	.3825	.3386	.2953
Description Queries					
C-K-D-02 bw e	**69**	**.2601**	**.3895**	**.3518**	**.2855**
****C-K-D-02 bw***	*71*	*.2784*	*.3965*	*.3658*	*.2923*
***C-K-D-02 b**	62	.2402	.3123	.2965	.2640
***C-K-D-02 w**	73	.2718	.3930	.3561	.2908
***C-K-D-02 bsys**	69	.2681	.3807	.3447	.2905
Description + Narrative Queries					
C-K-DN-03 bw	**70**	**.3076**	**.4281**	**.3737**	**.3181**

a) Relax Assessment (number relevant = 3917)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-K-T-01 bw	**78**	**.2590**	**.3702**	**.3237**	**.2792**
***C-K-T-01 b**	69	.2290	.3175	.2886	.2519
***C-K-T-01 w**	77	.2520	.3614	.3246	.2760
***C-K-T-01 bsys**	74	.2528	.3351	.2921	.2647
Description Queries					
C--K-D-02 bw e	**70**	**.2471**	**.3526**	**.3202**	**.2706**
****C-K-D-02 bw***	*73*	*.2632*	*.3596*	*.3298*	*.2782*
***C-K-D-02 b**	63	.2260	.2807	.2632	.2513
***C-K-D-02 w**	74	.2555	.3509	.3219	.2734
***C-K-D-02 bsys**	70	.2518	.3386	.3061	.2778
Description + Narrative Queries					
C-K-DN-03 bw	**71**	**.2956**	**.3965**	**.3377**	**.3118**

b) Rigid Assessment (number relevant = 3131)

Table 3a,b: C-K Cross Language Results for Query Types T, D, DN.

The C-K rigid assessment MAP values range between 0.2590 T queries to 0.2956 for DN queries. They represent 56% (T queries: .2590/.4588), 70% (D: .2632/.3777) and 61% (DN: .2956/.4848) of monolingual K-K retrieval. Relax assessment have similar ratios. Short title queries have worst comparison to K-K monolingual.

Table 3 also shows results using bigram indexing or word indexing alone. Here, bigram indexing returns results much worse than HAM indexing. It seems that going through four translations lead to proportionately more suffixes than content. Meaningless bigrams proliferate and becomes a factor. With HAM processing, stems and stopwords are removed resulting with much less noise. The two rows tagged with 'bsys' are bigram runs with queries that concatenate only two Systran English-Korean translations (without EnGuide). Their results improve to close the gap with HAM indexing results.

Comparison with E-K retrieval shows that C-K accomplishes between 73-83% of direct E-K CLIR. 22 of these C-K queries perform better in MAP (with 3 having differences > 0.2) and 35 worse (with 11 having differences > 0.2). These 3+11=14 query differences add up to 72% of the total MAP difference of 0.0741 (i.e. E-K MAP 0.3331 – C-K MAP 0.2590).

The C-K retrieval makes use of pivot translation to English as an intermediate step. If Chinese-English translation leads to a bad query q^{CE}, it would have poor MAP value for C-E retrieval (Section 5.2) compared to the corresponding E-E retrieval. When this bad query q^{CE} is further translated to Korean forming q^{CEK}, one would naturally expect it also has bad MAP value for C-K retrieval compared to E-K. However, poor C-E MAP comparison does not necessarily mean bad C-K MAP comparison. Fig.2 shows a scatter plot of the difference MAP(EE)-MAP(CE) vs MAP(EK)-MAP(CK); one discerns little correlation: r = 0.1926. A translated query performing poorly at the pivot language stage need not mean that it will continue to perform badly at the target language stage. A term that is not useful (e.g. frequency consideration) at the pivot language retrieval may become useful when further translated to the target language. In our pre-translation expanded queries, for example #25 the not too specific term 'unemployment rate' gets translated into specific words (실업율, 실업률) for Korean retrieval. In #33, 'Albuminuria' does not exist in the English collection, but translated into a specific word in Korean (단백뇨증). In #57, an important word 'ecological' gets represented into three forms (생태, 생태학, 생태학적) after HAM processing.

We would like to have direct Q^{CK} bilingual retrieval results for comparison purposes but did not find resources for the direct C-K translation.

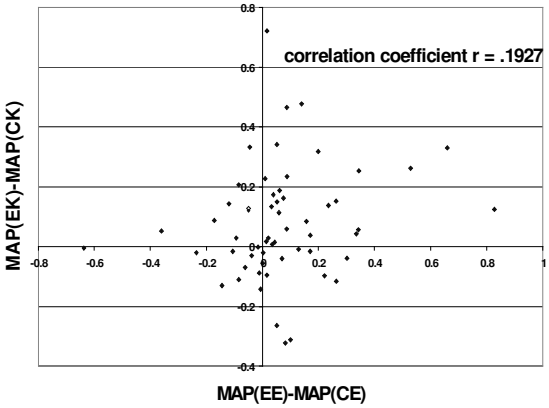

Fig.2: Scatter Plot – MAP(EK)-MAP(CK) vs MAP(EE)-MAP(CE)

5 Retrieval with English Collections:

5.1 E-E Monolingual Retrieval

English monolingual retrieval was performed to provide a basis for evaluating our Chinese-English crosslingual retrieval. Results are tabulated in Table 4. Porter's stemming, stopword removal, 2-word phrases are employed for indexing, as well as PRF procedures in our retrievals.

We also discovered later that our submitted runs actually used a command file that did not make use of the full functionality of our PIRCS retrieval system. When the correct command files were used with some parameter adjustments, the rows tagged as

pircs-	R%	MAP	P10	P20	R.Pre
E-E-T-01	**70**	**.4042**	**.6310**	**.5879**	**.4116**
E-E-D-02	**71**	**.3876**	**.5845**	**.5638**	**.3998**
***E-E-T-01 co**	77	.4468	.6793	.6388	.4544
***E-E-D-02 co**	77	.4390	.6931	.6388	.4510

a) Relax Assessment (number relevant = 11056)

pircs-	R%	MAP	P10	P20	R.Pre
E-E-T-01	**72**	**.3175**	**.4603**	**.4086**	**.3328**
E-E-D-02	**76**	**.3055**	**.4138**	**.4000**	**.3184**
***E-E-T-01 co**	79	.3517	.5069	.4638	.3562
***E-E-D-02 co**	81	.3452	.5155	.4698	.3603

b) Rigid Assessment (number relevant = 5866)

Table 4a,b: E-E Monolingual Results for 58 Query Types T, D.

'co' (for corrected) show the results, which improve by more than 10%.

5.2 C-E Crosslingual Retrieval

Chinese-English bilingual retrieval was done as an intermediate step to our goal of C-K pivot retrieval. The process has been discussed in Section 4.3. These results are tabulated in Table 5. The rigid assessment MAP values of 0.2380 for T and 0.2238 for D queries are the top results among participants. These represent 75% (T) and 73% (D) compared to our E-E monolingual retrieval, and 71% and 73% for relax assessment (Table 5). These Q^{CE} queries do not have the assistance from web-assisted translation. The retrieval result supports our observation that the MT software are reasonably adequate for CLIR purposes.

Table 5 also shows two un-submitted runs that do not include pre-translation expansion (x-). It is seen that pre-translation expansion hurts both title and description queries. Another two un-submitted D-query runs show results of using MT software individually: tagged as 'sys' and 'lot'. Systran translation is better than Loto. Just as in E-K, the

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-E-T-01	**62**	**.2879**	**.5017**	**.4888**	**.3319**
***C-E-T-01 x-**	62	.3235	.5069	.4888	.3494
Description Queries					
C-E-D-02	**61**	**.2829**	**.4845**	**.4629**	**.3267**
***C-E-D-02 x-**	60	.2930	.4879	.4552	.3212
***C-E-D-02 sys**	59	.2736	.4483	.4241	.2943
***C-E-D-02 lot**	53	.2446	.4034	.3707	.2760

a) Relax Assessment (number relevant = 11056)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-E-T-01	**68**	**.2380**	**.3862**	**.3707**	**.2746**
***C-E-T-01 x-**	63	.2471	.3586	.3293	.2692
Description Queries					
C-E-D-02	**66**	**.2238**	**.3552**	**.3310**	**.2628**
***C-E-D-02 x-**	64	.2286	.3483	.3241	.2536
***C-E-D-02 sys**	63	.2159	.3276	.3069	.2361
***C-E-D-02 lot**	56	.1875	.2948	.2552	.2133

b) Rigid Assessment (number relevant = 5866)

Table 5a,b: C-E Crosslingual Results for 58 Query Types T, D.

concatenated translation results for pircs-C-E-D-02x- are better than these that use translations singly. Here however, the improvements are not significant according to the sign test at the 5% level.

There are 12 C-E title queries with rigid MAP values at least 0.2 worse, and 3 better, compared to their E-E counterpart. They amount to 74% of the total difference MAP of 0.0795 (i.e. 0.3175-0.2380). Several are due to bad entity/terminology translations (such as: #2 約翰走路 (John Walker) John walks; #4 葛瑞菲絲．喬納(Griffith Joyner) Ge Juifei silk. Jonah; #50 地底核武試爆 (Underground Nuclear Test) Core tries exploding militarily, protest at the ground bottom). Some are further weakened by related but noise terms from pre-translation expansion (such as #30 動物複製技術(cloning) duplication, transplant, organ; #19 國際海上意外事件 (International incidents at Sea) International marine accident; warships, nuclear, security check; #33 研究，蛋白質(Research, Protein) Study, the protein; AIDS). An example success for C-E is #54 奧林匹克，賄賂，傳聞 (Olympic, Bribe, Suspicion) Olympic, bribes, the rumor; Salt Lake City. The last phrase was added by pre-translation expansion.

6 Chinese C-C Monolingual Retrieval

Chinese monolingual retrieval was performed as before [9]: based on combination of retrieval lists using bigram + 1-gram, and short word + character indexing. Results are shown in Table 6; they provide a basis for CLIR involving Chinese collections. The description result with its rigid MAP value of 0.2150, is the second best among automatic C-C submissions.

pircs-	R%	MAP	P10	P20	R.Pre
C-C-T-01	**84**	**.2673**	**.3373**	**.2864**	**.2725**
C-C-D-02	**86**	**.2761**	**.3542**	**.2941**	**.2810**

a) Relax Assessment (number relevant = 2085)

pircs-	R%	MAP	P10	P20	R.Pre
C-C-T-01	**83**	**.2097**	**.2356**	**.1958**	**.2059**
C-C-D-02	**85**	**.2150**	**.2475**	**.1975**	**.2010**

b) Rigid Assessment (number relevant = 1318)

Table 6a,b: C-C Monolingual Results for 59 Query Types T, D.

7 Conclusion and Discussion

We tested several MT packages for direct and pivot cross language retrieval purposes: Chinese to English (Systran, Loto), English to Korean (Systran, EnGuide). These are augmented with our web-based entity/terminology-oriented translation procedure. Experiments show that concatenation of two translations performs better than using them singly

for direct C-E and E-K CLIR. Individually, Systran translation provides better retrieval outcome for C-E compared to Loto, and for E-K compared to EnGuide.

Our web-based entity/terminology-oriented translation is found effective, and can provide some 15% improvement in mean average precision for E-K CLIR.

In Korean retrieval, bigram provides better effectiveness than word indexing except in C-K runs where a query has a combination of 4 translations and random bigram noise may become an adverse factor. In general, combination of their retrieval lists provides better effectiveness except for K-K title run.

The MT software can also be chained to provide transitive translation via English as the pivot language. Results show that pivot Chinese-English-Korean bilingual retrieval can provide about 55% to 65% of monolingual effectiveness. Performance at the pivot language retrieval stage does not have much correlation with the performance at the target retrieval. Direct C-E and E-K CLIR runs provide 71% to 88% of monolingual effectiveness.

Table 7 summarizes our results compared to the official Maximum and Median of all submitted runs from participants. (=> denotes the basis values from which the right-hand columns have percentages calculated).

MAP (relax)	**Q^{CC}**	**Q^{EE}**	**Q^{CE}**	**Q^{KK}**	**Q^{EK}**	**Q^{CEK}**
Title Queries						
Max	.3799	.4512	.2879	.5361	.3598	.4343
pircs	**.2673**	**.4042**	**.2879**	**.4934**	**.3598**	**.2783**
%mono		**=>**	**71%**	**=>**	**73%**	**56%**
Median	.2356	.3954	.2420	.4934	.2429	.4199
Description Queries						
Max	.3880	.4368	.2829	.5097	.3566	.4314
pircs	**.2761**	**.3876**	**.2829**	**.4049**	**.3566**	**.2601**
%mono		**=>**	**73%**	**=>**	**88%**	**64%**
Median	.2219	.3859	.2255	.3992	.2313	.3458
Description + Narrative Queries						
Max	.3103	.4962	.2294	.6212	.0849	.5138
pircs				**.5161**		**.3076**
%mono				**=>**		60%
Median	.2915	.4423	.1147	.5004	.0730	.4572

a) Relax Assessment

MAP (rigid)	**Q^{CC}**	**Q^{EE}**	**Q^{CE}**	**Q^{KK}**	**Q^{EK}**	**Q^{CEK}**
Title Queries						
Max	.3146	.3576	.2380	.5078	.3331	.4726
pircs	**.2097**	**.3175**	**.2380**	**.4588**	**.3331**	**.2590**
%mono		**=>**	**75%**	**=>**	**73%**	**56%**
Median	.1881	.3245	.1860	.4588	.2244	.3870
Description Queries						
Max	.3255	.3469	.2238	.4685	.3249	.3973
pircs	**.2150**	**.3055**	**.2238**	**.3777**	**.3249**	**.2471**
%mono		**=>**	**73%**	**=>**	**86%**	**65%**
Median	.1741	.3026	.1819	.3727	.2115	.3222
Description + Narrative Queries						
Max	.2556	.4000	.1746	.5825	.0750	.4726
pircs				**.4848**		**.2956**
%mono				**=>**		61%
Median	.2363	.3573	.0796	.4694	.0647	.4196

a) Rigid Assessment

Table 7a,b: Comparison with Max and Median Results (values above => means monolingual basis)

Acknowledgment

This work was partially supported by a U.S. Govt. DST/ATP contract 2003*H532600*000. Peter Deng provided the program for our entity-oriented web translation.

References

[1] K.L. Kwok. Improving English & Chinese Ad-Hoc Retrieval: A Tipster Text Phase 3 Project Report. Information Retrieval, 3:313-338, 2000.

[2] Systran MT software: http://systransoft.com

[3] Loto MT software: http://www.lotousa.com

[4] Enguide MT software: http://www.lnisoft.co.kr

[5] Deng, P & K.L. Kwok. A cross language name finding system. IJCNLP-04 Companion Volume of Proceedings (Interactive Posters/ Demos) pp.9-12, 2004.

[6] Lee, J.H & Ahn, J.S. Proc. 19th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.216-224 (1996).

[7] Hangul Analysis Module: http://nlp.kookmin.ac.kr/HAM/kor/download.html

[8] Seo, H-C, Kim S-B, Kim B-I, Rim H-C & Lee, S-Z. KUNLP system for NTCIR-3 English-Korean Cross-Language Information Retrieval. NTCIR Workshop 3 Meeting: CLIR. Pp.73-78, 2002.

[9] K.L. Kwok. NTCIR-2 Chinese and cross language experiments using PIRCS. In: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. Tokyo:NII. pp.111-118, 2001.