NTCIR-5 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS

Kui-Lam Kwok, Sora Choi, Norbert Dinstl and Peter Deng

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

Abstract

In NTCIR-5 our focus is to see if web-assisted query expansion is useful, and to test an English-Korean bilingual dictionary. We participated in Chinese, Japanese, Korean and English monolingual retrieval using also web expansion for Chinese and English. We also performed Chinese-English, English-Chinese, English-Korean bilingual, and Chinese-Korean pivot bilingual CLIR. The query translation approach was employed. MT translation was combined with our web-based entity-oriented translation which translates named entities extracted from the original query as well as web-based expansion terms. For English-Korean, a dictionary translation method was also used.

In general, monolingual retrieval results are about median, while cross-lingual runs (except Chinese-English) appear to be among the top results. For English-Korean, our bilingual dictionary translation by itself is not competitive, but when concatenated with web translation came to within 5% of MT with web translation. Using the web to expand a query before retrieval was not successful except for monolingual English and pivot bilingual Chinese-Korean retrieval.

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

1 Introduction

We continue to use the query translation approach for CLIR. The different types of experiments done for NTCIR-5 can be summarized in Fig.1. Here, a column denotes a target collection and its language (in superscript) such as: D^K, meaning the Korean collections. Each row denotes a query set and its source language. These queries can be used for monolingual retrieval as denoted by the diagonal cells of the matrix such as Q^J:D^J. They may be translated from their source language to other target languages for retrieval, such as Q^{CEK} (cell row Q^C under column D^K) where the original Chinese query set was translated to English, then transitively to Korean, for retrieval with the Korean collection. Characteristics of the document collection and the queries are given in [1].

Doc→ Query ↓	D^C	D^J	D^K	D^E
Q^C	Q^C:D^C		Q^{CEK}:D^K	Q^{CE}:D^E
Q^J		Q^J:D^J		
Q^K			Q^K:D^K	
Q^E	Q^{EC}:D^C		Q^{EK}:D^K	Q^E:D^E

Fig.1: Different Types of Experiments

There were a total of 33 runs. These include retrievals with the Chinese collections with run IDs:

pircs-C-C-T-01 pircs-C-C-D-02
pircs-C-C-DN-03 pircs-C-C-T-04
pircs-C-C-D-05
pircs-E-C-T-01 pircs-E-C-T-02
pircs-E-C-D-03 pircs-E-C-D-04
pircs-E-C-T-05;

Retrievals with the Japanese collections named as:

pircs-J-J-T-01 pircs-J-J-D-02
pircs-J-J-DN-03;

Retrievals with the Korean collections named as:

pircs-K-K-T-01 pircs-K-K-D-02
pircs-E-K-T-01 pircs-E-K-T-02
pircs-E-K-D-03 pircs-E-K-D-04
pircs-C-K-T-01 pircs-C-K-T-02
pircs-C-K-D-03 pircs-C-K-D-04
pircs-C-K-D-05

Retrievals with the English collections named as:

pircs-E-E-D-01 pircs-E-E-T-02
pircs-E-E-T-03 pircs-E-E-D-04
pircs-C-E-D-01 pircs-C-E-T-02
pircs-C-E-T-03 pircs-C-E-D-04
pircs-C-E_D-05.

This year is our first attempt on Japanese (monolingual) retrieval with n-gram processing. Korean processing was improved from last year. We tested query expansion via the web for English and Chinese queries. In addition various web-assisted query translation processing was employed for our cross-lingual retrievals. Sections 2 describes our query translation and Korean processing resources. Sections 3 to 6 discuss our results for retrieval with Chinese, Japanese, Korean and English collections. Section 7 has our conclusion.

2 Tools for Different Languages

Our tools for Korean translation and processing are slightly different from NTCIR-4. In addition to Systran and Lni MT software, we also added an E-K dictionary for English-Korean translation. This dictionary was compiled from several sources on the web. First is X-Korean phonetic list from The National Academy of the Korean Language (http://www.korean.go.kr/search) of about 41K entries that consist of about 14% place and 41% person names, and 45% general terms. Another source is a general E-K dictionary of about 200K downloaded from (http://www.kecl.ntt.co.jp/icl/mtg/resources/engdic/engdic-doc.html). These two lists were consolidated with the Korean texts further processed by the KLT software (described below) for root words. Another translation resource is our web-assisted terminology and entity-oriented translation algorithm [2]. One of our goals in NTCIR-5 is to see if dictionary translation can be comparable to MT for CLIR.

Bigram processing of Korean text was improved by simple stemming first before bigram formation. Morphological analysis was performed using the KLT 2.0 index program to produce root words. This is an updated version of HAM 6.0 used last year.

Tools for Chinese remain similar to last year's except that only Systran MT was employed together with our web-based translation. Some experiments employ the web to expand Chinese or English queries before retrieval.

When a common English term is used as input to our web-based translation for English-Chinese, the result may sometimes be noisy with unintended outputs. To avoid this problem, we employed an entity extraction software IdentiFinder from BBN [3] (that can handle both English and Chinese texts) to extract entity names from queries, or from expansion terms for input to web translation.

For English-Korean web translation however, our experience is that there are much less noisy output even with common terms. Here, we employ MINIPAR parser (http://www.cs.ualberta.ca/~lindek/minipar.htm) to extract adjectives, verbs, nouns, noun-noun phrases as well as MINIPAR-defined phrases as input for web translation.

Japanese processing was kept simple using n-grams only. A small stopword/stop-character list was used for processing Hiragana before bigram formation.

3 Retrieval with the Chinese Collections

3.1 Monolingual (C-C) Chinese Retrieval

We followed our previous procedures of creating two indices for the Chinese collection: bigram with 1-gram, and short-word with single character, both frequency-thresholded. Retrieval lists from the two indices were combined to form the final output.

For the first three submissions: pircs-C-C-xx-01 to 03, the Chinese string from the respective section of each topic (title, description or description plus narrative) was used directly as query. Certain introductory or stop phrases were removed in the description section, as done for English retrieval. Results are shown in Table 1 attaining rigid MAP values of .3958 (title), .3897 (description) and .4276 (description+narrative). Title and description results are close. The last set of long queries performs best.

For the other two submissions: pircs-C-C-T-04 and –D-05, the Chinese query was first used as probes on the web via Google search using a 'window rotation' method [4]. A maximum of 3 consecutive words from the query were used as probe on the web. This 3-word window rotates through the whole query generating n web-snippet lists for a query of n words. These snippet lists are then filtered by voting, and the final web snippet list content serve to expand the original query with 15 terms based on occurrence frequency. Thus, the web is exploited as an all-domain thesaurus that brings in associated terms to enrich the original query representation before our usual PIRCS retrieval. However, the web can be noisy with respect to a probe and can cause topic shift. This may be because the web probe is not sufficiently precise. To guard against such bad cases, a simple filter for such drifting is used to screen each

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-C-T-01	87	**.3958**	.4880	.3990	.3897
C-C-T-04	88	.3493	.4120	.3550	.3508
!C-C-T-04f	*88*	*.3467*	*.4060*	*.3540*	*.3416*
Description Queries					
C-C-D-02	94	**.3897**	.4780	.4090	.3727
C-C-D-05	91	.3589	.4400	.3830	.3422
!C-C-D-05f	*90*	*.3422*	*.4240*	*.3620*	*.3294*
Title+Description Queries					
C-C-DN-03	96	**.4276**	.5260	.4490	.4180

a) Rigid Assessment (number relevant = 1885)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-C-T-01	86	**.4651**	.6080	.5410	.4485
C-C-T-04	87	.4084	.5620	.5060	.4041
!C-C-T-04f	*86*	*.4080*	*.5480*	*.4990*	*.4035*
Description Queries					
C-C-D-02	90	**.4625**	.6200	.5510	.4400
C-C-D-05	88	.4265	.5800	.5180	.4057
!C-C-D-05f	*87*	*.4112*	*.5600*	*.4910*	*.3969*
Title+Description Queries					
C-C-DN-03	94	**.4982**	.6580	.5990	.4783

b) Relax Assessment (number relevant = 3052)

Table 1a,b: C-C Monolingual Results for 50 Queries of Types T, D, DN.

of the original query and the expanded query as follows using the bigram with 1-gram representation of the query:

if (v2 < 5) & (v2/o2 + 0.1 v1/o1) < .5)
then drift is detected.

o2 and o1 are counts of bigram and 1-gram of the original query, and v2 and v1 are counts of the overlap between the two query types. The 1-gram factor is added to relax a bit the bigram overlap requirement for no drift (if only bigrams are considered). For those queries that do not pass the filter, the original query result from pircs-C-C-T-01 is used.

It is seen that our web expansion attempt leads to a loss of about 10% or more (rigid MAP .3493 (vs. .3958) for pircs-C-C-T-04, .3589 (vs. .3897) for –D-05). Web pages can be noisy, and our filter, which screens out 12 queries, seems not sufficiently sensitive. If we had allowed all the expanded queries without filtering, the results shown in the italicized rows (un-submitted runs !C-C-T-04f and !-D-05f) are obtained. They are slightly worse -- the filter seems to have some positive effect.

3.2 English-Chinese (E-C) Bilingual Retrieval

For EC-CLIR, Systran MT was employed to provide a basis translation of a query. This is augmented with our web-based E-C translation procedure. Last year, words left un-translated by Systran were sent to our web-based translation algorithm. This year, a different procedure was followed. For title queries, all words were used for web translation. For description queries only extracted named entity terms were sent to our web-based translation and output added to the query. This way, entity names may get better chance of hitting a correct translation. These generate our first pair of submissions pircs-E-C-T-01 and –D-03.

Table 2 shows that these submissions achieve between 59%-64% (title) and 69%-73% (description) of our best C-C monolingual retrieval using rigid precision measures. Description queries have better results than titles (MAP .2682 vs. .2459). This reflects our NTCIR-2 findings that longer queries generally perform better in E-C CLIR. Comparing title runs pircs-E-C-T-01 with –C-C-T-01, there are only 10 q^{EC} queries performing better and 2 equal to the corresponding monolingual q^{C} queries. Comparing description runs –E-C-D-03 with –C-C-D-01, there are 14 q^{EC} queries performing better and 1 equal to the corresponding monolingual q^{C} According to the sign test, bilingual results are significantly worse at the 5% level compared to monolingual results.

This year we also experimented with web-based pre-translation expansion with related entity names. An original English query was first used for web probing and the returned snippets or documents help to define an expansion term list. We truncate the list to the top 20. BBN's IdentiFinder was employed to detect entities and these were sent to our web-based translation and merged with the Systran-translated query. We hope that the web is sufficiently large that it can cover practically all domains of the queries, and bring in related named entity. Results are reported as pircs-E-C-T-02 and –D-04, and they can be compared with pircs-E-C-T-01 and –D-02.

As seen in Table 2, this pre-translation expansion was not successful and depresses result compared to no pre-translation expansion (e.g. title rigid MAP .2309 vs. .2459, description .2528 vs. .2682).

A fifth run pircs-E-C-T-05 with title queries was submitted using a larger segmentation dictionary for short-word retrieval. This run can be compared with the –T-02 and is a few percent better in MAP but worse in P10. However, when this dictionary was used for description runs, it does not improve results.

The un-submitted rows !E-C-x-01sys in Table 2a show translation using Systran only without web. Worth noting is that query #14 "nanotechnology" has

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-C-T-01	75	**.2459**	.2880	.2530	.2478
% mono	*86*	*62*	*59*	*63*	*64*
E-C-T-02	76	.2309	.2960	.2590	.2365
% mono	*87*	*58*	*61*	*65*	*61*
E-C-T-05	75	.2456	.2840	.2550	.2428
% mono	*80*	*63*	*59*	*62*	*65*
!E-C-T-01sys	67	.2021	.2388	.2235	.2021
% mono	*77*	*51*	*49*	*56*	*52*
Description Queries					
E-C-D-03	83	**.2682**	.3500	.2960	.2661
% mono	*88*	*69*	*73*	*72*	*71*
E-C-D-04	86	.2528	.3120	.2750	.2557
% mono	*91*	*65*	*65*	*67*	*69*
!E-C-D-03sys	78	.2276	.2940	.2650	.2388
% mono	*83*	*59*	*61*	*64*	*64*

a) Rigid Assessment (number relevant = 1885)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-C-T-01	72	.2975	.4000	.3640	.2870
% mono	*84*	*64*	*66*	*67*	*64*
E-C-T-02	73	.2826	.4000	.3640	.2851
% mono	*85*	*61*	*66*	*67*	*64*
E-C-T-05	72	**.3004**	.3980	.3660	.2938
% mono	*80*	*65*	*64*	*66*	*67*
Description Queries					
E-C-D-03	80	**.3235**	.4380	.4050	.3275
% mono	*89*	*70*	*71*	*74*	*74*
E-C-D-04	83	.3196	.4180	.3970	.3121
% mono	*92*	*69*	*67*	*72*	*71*

b) Relax Assessment (number relevant = 3052)

Table 2a,b: E-C Bilingual Results for 50 Queries of Types T, D.

no output by Systran and the precision is zero. The result is much worse than E-C-T-01 (MAP .2021 vs..2459). This demonstrates the usefulness of web translation for terminology and names.

4 Retrieval with the Japanese Collection

Monolingual Japanese (J-J) Retrieval

Our aim with J-J retrieval is to see how far simple n-gram methods can work. Japanese texts are encoded with three character sets: Kanji, Hiragana and Katakana. They were isolated independent of each other. For Kanji, we used our usual overlapping bigram processing as in Chinese, but without stopword removal. For Hiragana, the same is done except that stopword/character removal was performed first using a small list. They are retained in case queries get very short. For Katakana, we use non-overlapping 4-grams to segment any long strings from left to right. These become our indexing terms, and when tested on NTCIR-4 gave reasonable results. Since Katakana is employed mainly for transliteration of foreign words, they may be important indexing terms. We assume that the longer such terms match, the higher their matching weight should be. Thus, if there is a string of 8 Katakana characters, we would form two indexing terms. It also allows some partial matching if for example the 8 Katakana matches only the first or last four.

Table 3 shows results of our experiments. In general longer queries perform better, with title, description and description+narrative queries returning rigid MAP values of .2980, .3018 and .3775. Recall R% achieves over 86% to 94%. For comparison, the overall maximum and median rigid MAP are title (.4193, .3246), description (.3823, .3018), and any (.448, .335). Our approach appears to perform progressively better with longer queries.

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
J-J-T-01	86	**.2980**	.3617	.3138	.3098
Description Queries					
J-J-D-02	87	**.3018**	.3638	.3213	.3067
Description + Narrative Queries					
J-J-DN-03	94	**.3775**	.5085	.4170	.3854

a) Rigid Assessment (number relevant = 2112)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
J-J-T-01	85	**.3993**	.5362	.4670	.4039
Description Queries					
J-J-D-02	85	**.4043**	.5277	.4809	.4013
Description + Narrative Queries					
J-J-DN-03	92	**.5018**	.7319	.6266	.4890

b) Relax Assessment (number relevant = 4190)

Table 3a,b: J-J Monolingual Results for 50 Queries of Types T, D, DN.

5 Retrieval with the Korean Collections

5.1 Monolingual Korean (K-K) Retrieval

As in Chinese, retrieval of Korean text was done for both bigram and word indexing, and later combined. Word indexing employed the nouns analyzed by KLT software.

Because titles are short, mostly nouns and do not vary much in different morphological forms, bigram indexing was done without doing suffix stripping first. This was not true with the longer description queries. Two runs were submitted: pircs-K-K-T-01 and -D-02. Results are shown in Table 4. Description runs perform better than titles, and both have very high recall percentages of over 95% (rigid R%).

The overall best and median title submissions have rigid MAP values of .5586 and .4468 respectively. Our title rigid MAP value of .4490 is above median. For description queries, the best and median rigid MAP are .5079 and.4541. Our value of .4816 is about 5% below the best, and above median.

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
K-K-T-01	97	**.4490**	.5220	.4640	.4170
Description Queries					
K-K-D-02	95	**.4816**	.5480	.4910	.4585

a) Rigid Assessment (number relevant = 1829)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
K-K-T-01	94	**.4903**	.6000	.5530	.4791
Description Queries					
K-K-D-02	93	**.5335**	.6500	.5940	.5241

b) Relax Assessment (number relevant = 2683)

Table 4a,b: K-K Monolingual Results for 50 Queries of Types T, D.

5.2 English-Korean (E-K) Bilingual Retrieval

English queries were translated by Systran, Lni MT and web translation, or dictionary with web. Their outputs were concatenated to form E-K translated queries. Web translation is done for all title query words. For description, it is analyzed by MINIPAR, stopwords removed, and adjectives, verbs, nouns, phrases and noun-noun phrases are retained for web translation. Un-translated English words (except for capital acronyms) are deleted. These are our basis pircs-E-K-T-01 and -D-03 submissions. They returned the best overall E-K bilingual results of .3975 and .4092 rigid MAP for title and description queries respectively. Precision values range from 85 to 92% of monolingual values for title, and 81 to 89% for description queries. Comparing E-K-T-01 with K-K-T-01, 19 q^{EK} queries

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-K-T-01	91	**.3975**	.4460	.4130	.3848
% mono	*94*	*89*	*85*	*89*	*92*
E-K-T-02	87	.3774	.4320	.3790	.3650
% mono	*90*	*84*	*83*	*82*	*88*
!E-K-T-01-sys	85	.3281	.3740	.3220	.3160
!E-K-T-01-web	80	.3154	.3540	.3210	.3104
!E-K-T-01-dic	66	.2841	.3100	.2840	.2723
Description Queries					
E-K-D-03	90	**.4092**	.4660	.4000	.4077
% mono	*95*	*85*	*85*	*81*	*89*
E-K-D-04	86	.3938	.4540	.4020	.3796
% mono	*91*	*82*	*83*	*82*	*83*
!E-K-D-03-sys	83	.3830	.4160	.3630	.3647
!E-K-D-03-web	79	.3527	.4460	.4150	.3483
!E-K-D-03-dic	67	.2777	.3140	.2740	.2776

a) Rigid Assessment (number relevant = 1829)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-K-T-01	89	**.4335**	.5220	.4900	.4253
% mono	*95*	*88*	*87*	*89*	*89*
E-K-T-02	84	.4132	.5060	.4460	.3925
% mono	*89*	*84*	*84*	*81*	*82*
Description Queries					
E-K-D-03	88	**.4510**	.5580	.4900	.4514
% mono	*95*	*85*	*86*	*82*	*86*
E-K-D-04	84	.4406	.5440	.4940	.4224
% mono	*90*	*83*	*84*	*83*	*81*

b) Relax Assessment (number relevant = 2683)

Table 5a,b: E-K Bilingual Results for 50 Queries of Types T, D.

performed better, one equal and 30 performed worse than monolingual q^K queries. Comparing E-K-D-03 with K-K-D-02, 21 q^{EK} queries performed better, one equal and 28 performed worse than monolingual q^K queries. According to the sign test, there is no significant difference between these monolingual and bilingual results at the 5% significance level. A scatter plot of the K-K-T-01 vs. E-K-T-01 rigid APs is shown in Fig.2. Queries above the diagonal have monolingual AP better and vice versa. The correlation coefficient of .77 is fairly high.

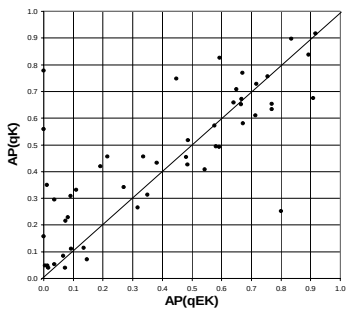

Fig.2: Title Query Scatter Plot – AP(q^K) vs. AP(q^{EK}) (rigid)

The other two submissions pircs-E-K-T-02 and pircs-E-K-D-04 made use of the English-Korean bilingual dictionary (instead of Systran and Lni MT). This output is concatenated with web translation. The purpose is to see if dictionary translation is comparable to the use of MT. For each English word, all dictionary translation mappings are captured. Quite often, long definitions are found in the mappings and lead to noisy output. We deal with this by capturing only those mappings with a single noun. Each query may now have several sets of Korean nouns for each English word. We filter the output further by choosing one set with the least ambiguity (i.e. with the smallest number of mapping) as anchor. Candidates in other sets are ranked with respect to the anchor based on their log likelihood ratio evaluated using target collection term co-occurrence frequencies with the anchor. Only the top three candidates of each set and the anchor set were kept as the final translation query output. An illustration of this process for Query 2 is shown below:

a) dictionary output after keeping only single nouns:

PRESIDENT 대통령 회장 사장 총재 학장 총장 의장 장관 교장
PERU 페루
ALBERTO 알베르토
SCANDAL 세상 스캔들 치욕 의옥 추문 중상
BRIBE 뇌물 증회

b) disambiguated output after anchor processing:

PERU 페루
ALBERTO 알베르토
BRIBE 뇌물 증회
SCANDAL 스캔들 추문 세상
PRESIDENT 대통령 의장 총장

From Table 5, it is seen that this dictionary with web translation approach is about 4 to 5% worse than MT (title rigid MAP of .3774 vs. .3975, and description MAP of .3938 vs. .4092). Table 5a also shows three un-submitted rows for title (and three for description) that use individual un-concatenated translation as query. It is seen that dictionary translation by itself performs much worse than MT or web translation. However, concatenation with web translation enhances its effectiveness much closer to that of MT plus web. It is possible that with more experimentation, one could close this gap.

5.3 Chinese-Korean (C-K) Pivot Bilingual Retrieval via English

We also submitted Chinese-Korean retrieval via English as pivot. For pircs-C-K-T-01 and pircs-C-K-D-03, Chinese queries were translated to English

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-K-T-01	80	**.2889**	.3200	.3120	.2845
%mono	*83*	*65*	*61*	*67*	*68*
C-K-T-02	78	.2715	.3080	.2930	.2876
%mono	*81*	*61*	*59*	*63*	*69*
Description Queries					
C-K-D-03	83	.3086	.3480	.3210	.2997
%mono	*88*	*64*	*63*	*65*	*65*
C-K-D-04	74	.3112	.3500	.3170	.3169
%mono	*78*	*65*	*64*	*64*	*69*
C-K-D-05	72	**.3263**	.3680	.3420	.3220
%mono	*76*	*68*	*67*	*69*	*70*

a) Rigid Assessment (number relevant =1829)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-K-T-01	76	**.3238**	.3960	.3750	.3155
%mono	*81*	*66*	*66*	*68*	*66*
C-K-T-02	74	.3206	.3900	.3710	.3203
%mono	*79*	*65*	*65*	*67*	*67*
Description Queries					
C-K-D-03	*80*	*.3463*	*.4220*	*.4010*	*.3549*
%mono	*86*	*65*	*65*	*68*	*68*
C-K-D-04	72	.3439	.4280	.4000	.3544
%mono	*77*	*64*	*66*	*67*	*68*
C-K-D-05	67	**.3583**	.4460	.4200	.3581
%mono	*72*	*67*	*69*	*71*	*68*

b) Relax Assessment (number relevant = 2683)

Table 6a,b: C-K Pivot Bilingual Results for 50 Queries with Types T, D.

using Systran MT and web translation. Web translation was done only on entities extracted by IdentiFinder from the Chinese queries. The resultant English were rendered to Korean by Systran and web translation as in E-K. Table 6a shows that rigid MAP values are .2889 (65% of K-K) for title and .3086 (64% of K-K) for description. Other precision measures range from 61% to 68% of monolingual K-K, which is quite respectable, considering the transitive nature of translation. 16 q^{CK} queries are better than q^{K} while 34 are worse. Sign test shows that this difference is significant at the .05 level. Fig.3 shows a scatter plot of KK title AP vs. CK similar to Fig.2. There are many more points falling

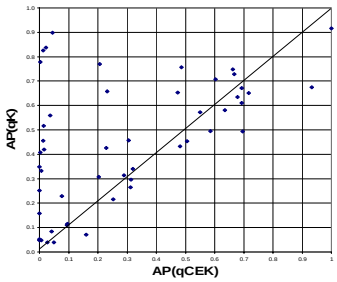

Fig.3: Title Query Scatter Plot – AP(q^{K}) vs. AP(q^{CEK}) (rigid)

near the y-axis (compared to the KK vs. EK plot of Fig.2) which means that many translated q^{CK} queries returned very low AP for the corresponding q^{K} queries with good AP.

The next two submissions pircs-C-K-T-02 and –D-04 make use of bilingual dictionary with web translation in place of Systran with web. Description queries surprisingly out-perform MT slightly, while for title it is reverse. Apparently, our dictionary translation can favor longer queries. It hints that our E-K dictionary with web translation may be a viable alternative to MT.

A fifth submission pircs-C-K-D-05 makes use of Chinese web-assisted expanded query (pircs-C-C-D-05). The expansion is reduced by entity extraction first, which are then translated by Systran and web. The English output is then used in E-K dictionary and web translation. Unlike monolingual runs, this entity expansion provides better result for CK description queries: with rigid MAP value .3263 (68% of K-K). Other precision values for this run vary between 67 to 70% of monolingual. The entities brought in by pre-translation web expansion at the q^{C} stage can help to improve the final q^{CEK} queries for retrieval.

In this pivot BLIR, the q^{CE} queries are crucial for the downstream q^{CEK} retrieval based on E-K translation. In NTCIR-4 [5] and [6], we pointed out that, for this purpose, the quality of q^{CE} queries cannot be judged by its English retrieval. Wordings of q^{CE} may not be appropriate for English retrieval and lead to bad C-E vs. E-E retrieval comparison; yet such a query might be adequate for translation to Korean so that its q^{CEK} retrieval may give good C-K vs. E-K retrieval comparison. We show in Fig.4 the scatter plot of AP(q^{CEK})-AP(q^{EK}) vs. AP(q^{CE})-AP(q^{E}). (C-E and E-E retrievals are discussed in Section 6).

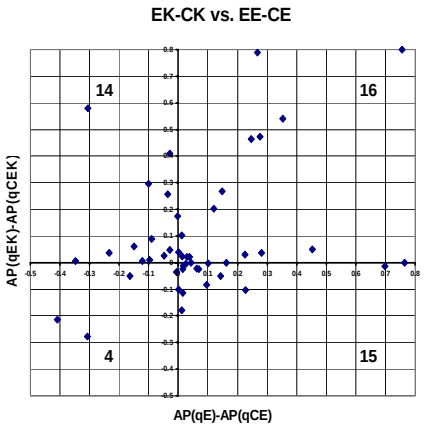

Fig.4: Title Query Scatter Plot – AP(q^{EK})-AP(q^{CEK}) vs. AP(q^{E})-AP(q^{CE}) (rigid)

The lower half of Fig.4 denotes q^{CEK} queries that outperform q^{EK}. Out of these, it is seen that a substantial number (15, lower right quadrant) gives worse CE retrieval (compared to EE), yet recovers after translation to Korean for CEK retrieval

(compared to EK). The other 4 q^{CE} queries (lower left quadrant) outperform q^{E}, and continue to outperform q^{EK} after further translation to q^{CEK}.

6 Retrieval with English Collections:

6.1 Monolingual (E-E) English Retrieval

English monolingual retrieval was performed to provide a basis for evaluating our Chinese-English cross-lingual results. These are tabulated in Table 7. Porter's stemming, stopword removal, 2-word phrases were employed for indexing, as well as PRF procedure in our retrievals.

pircs-E-E-T-02 and –D-01 are runs using only the respective topic sections. pircs-E-E-T-03 and –D-04 are runs that were enhanced with 20 expansion terms

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-E-T-02	88	.3970	.4653	.4235	.4066
E-E-T-03	91	**.4026**	.4594	.4276	.4063
Description Queries					
E-E-D-01	85	.3895	.4755	.4357	.4032
E-E-D-04	89	**.4241**	.4796	.4449	.4106

a) Rigid Assessment (number relevant =3073)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
E-E-T-02	89	.4701	.6041	.5378	.4625
E-E-T-03	92	**.4751**	.5857	.5408	.4672
Description Queries					
E-E-D-01	86	.4369	.5796	.5265	.4421
E-E-D-04	89	**.4722**	.5857	.5449	.4674

b) Relax Assessment (number relevant = 4064)

Table 7a,b: E-E Monolingual Results for 49 Queries with Types T, D.

from web probing first. These run pairs are directly comparable to see effects of using web enhancement. This web expansion of query before retrieval helps description MAP (rigid) by >8%, but only affects titles slightly but positively.

6.2 Chinese-English (C-E) Bilingual Retrieval

For C-E CLIR, Systran MT was employed to translate the title or description queries as a basis. BBN IdentiFinder was used to extract named entities from the Chinese queries. These entity names were sent to our web software for online translation. This output is then concatenated with the previous Systran output. These form the queries for our submissions: pircs-C-E-T-02 and pircs-C-E-D-01. They provided between 80% to 91% of various monolingual E-E retrieval precision effectiveness. Description queries have better performance than title queries.

Another pair of submissions, pircs-C-E-T-03 and –D-04, consists of the above queries but concatenated with expansion from the web. Each Chinese query was used to probe the web, and returned expansion

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-E-T-02	79	**.3339**	.4184	.3551	.3262
%mono	*89*	*84*	*90*	*84*	*80*
C-E-T-03	78	.3227	.3449	.3286	.3214
%mono	*86*	*80*	*75*	*77*	*79*
Description Queries					
C-E-D-01	86	**.3556**	.4122	.3908	.3540
%mono	*101*	*91*	*87*	*90*	*88*
C-E-D-04	83	.3490	.4224	.3949	.3463
%mono	*93*	*82*	*88*	*89*	*84*
C-E-D-05	75	.2692	.3367	.3092	.2808
%mono	*84*	*63*	*70*	*69*	*68*

a) Rigid Assessment (number relevant = 3073)

pircs-	R%	MAP	P10	P20	R.Pre
Title Queries					
C-E-T-02	78	**.3664**	.4898	.4214	.3599
%mono	*88*	*78*	*81*	*78*	*78*
C-E-T-03	77	.3590	.4082	.3878	.3592
%mono	*84*	*76*	*70*	*72*	*77*
Description Queries					
C-E-D-01	87	**.3902**	.4898	.4653	.3887
%mono	*101*	*89*	*85*	*88*	*88*
C-E-D-04	84	.3856	.4959	.4714	.3941
%mono	*94*	*82*	*85*	*87*	*84*
C-E-D-05	75	.2997	.4000	.3673	.3145
%mono	*87*	*69*	*69*	*70*	*71*

b) Relax Assessment (number relevant = 4064)

Table 8a,b: C-E Bilingual Results for 49 Queries with Types T, D.

terms were passed through IdentiFinder to pick up entity names. Only these names were translated by our web software and added to the above queries. These return results that are slightly inferior to the first pair without expansion. Their monolingual comparison was worse (between 75% - 89% for various precision comparisons) because web expansion helps the corresponding E-E retrieval results.

The fifth submission pircs-C-E-D-05 uses Systran translation without web-assisted translation. It is seen that its rigid MAP value of .2692 is much inferior to that for –D-01 of .3556. Because many topics have entity names, this again points out the importance to have an entity translation facility for CLIR.

7 Conclusion

We introduced an English-Korean dictionary for E-K bilingual retrieval. Although its E-K results are about 5% deficient compared to using MT, its application to C-K bilingual retrieval via English as

pivot outperforms Systran MT. The dictionary has to be augmented with entity-oriented web translation in order to be competitive. Entity translation is shown to be important for our good bilingual results. Pre-retrieval expansion of Chinese queries via the web does not help C-C or C-E retrieval, but has positive effects for C-K pivot bilingual retrieval. This process also helps to improve E-E monolingual retrieval.

Acknowledgment

We like to thank BBN for the use of their IdentiFinder software.

References

[1] K. Kishida. K. H. Chen, S. Lee, K. Kuriyama, N. Kando, H. H. Chen, S. H .Myaeng. Overview of CLIR task at the fifth NTCIR Workshop. In: Proceedings of the Fifth NTCIR Workshop Meeting, 2005.

[2] Kwok, K.L, Deng, P, Sun, H.L, Xu, W, Dinstl, N, Peng, P. & Doyon, J. CHINET – a Chinese name finder for document triage. Proc. of 2005 International Conference on Intelligence Analysis. 2005. (http://analysis.mitre.org/ proceedings_agenda.htm#papers)

[3] Bikel, D.M, Miller, S, Schwartz, R & Weischedel, R. A high-performance learning name-finder. In: Proc. on Conference of Applied Natural Language Processing, 1997.

[4] Kwok, K.L, Grunfeld, L, Sun, H.L & Deng, P. TREC2004 robust track experiments using PIRCS. In: Information Technology: The Fourteenth Text REtrieval Conference TREC-2004. Available at: http://trec.nist.gov/pubs/ trec13/papers/queens-college. robust.pdf

[5] Kwok, K.L, Dinstl, N & Choi, S. "NTICR-4 Chinese, English Korean Cross Language Retrieval Experiments using PIRCS". In: Working Notes of the Fourth NTCIR Workshop Meeting. NII: Tokyo, 2004. Available at http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/

[6] Kwok, Kui Lam, Choi, Sora and Dinstl, Norbert. Rich results from poor resources: NTCIR-4 monolingual and cross-lingual retrieval of Korean texts using Chinese and English. To be published in ACM TALIP.