# POSTECH at NTCIR-4: CJKE Monolingual and Korean-related Cross-Language Retrieval Experiments

In-Su Kang, Seung-Hoon Na, Jong-Hyeok Lee
Div. of Electrical and Computer Engineering
Pohang University of Science and Technology (POSTECH)
Advanced Information Technology Research Center (AITrc)
San 31, Hyoja-Dong, Pohang, Republic of Korea, 790-784
{dbaisk, nsh1979, jhlee}@postech.ac.kr

## Abstract

*This paper reports our proposal and experimental results at the NTCIR-4 CLIR task, where our team participated in all monolingual IR tracks and Korean-related bilingual IR tracks. For a monolingual IR, we used a combination strategy that integrates words and n-grams at the ranked list level. For a bilingual IR, a pseudo document translation scheme was combined with a default query translation method.*

**Keywords:** *Information Retrieval, Cross-language Information Retrieval, Query Translation, Document Translation, Translation Ambiguity*

## 1 Introduction

Unlike English, Chinese and Japanese do not use word delimiters in a normal text. In Korean, no word boundaries exist within an *eojeol*, a Korean spacing unit that corresponds to a phrasal unit, such as a noun phrase or a prepositional phrase in English. Thus, word segmentation is nontrivial for the three Asian languages we designate CJK (Chinese, Japanese, and Korean). Therefore, most CJK information retrieval (IR) employs character-based n-grams as well as words, as indexing units, in order to complement incomplete word identification.

Thus far, Chinese IR literature reports a little success when combining words and n-grams at the indexing unit level and ranked list level. However, there are few Japanese or Korean IR experiments that attempt to couple words and n-grams at a large scale. We empirically investigate the influence of coupling words and n-grams at the ranked list level in CJK monolingual retrieval. Our intuition is that different retrieval models will show varying performances according to which indexing unit is used. So, in combining words and n-grams, we concentrate on generating several ranked lists with different retrieval characteristics by incorporating various feedback schemes.

Concerning cross-language information retrieval (CLIR), we couple a default query translation (QT) and a pseudo document translation (PDT) method. PDT means a noisy document translation where each (target language) document term is replaced by its (source language) translation equivalents each of which is associated with a translation probability. Without translation probabilities, PDT converts a document into the representation of *a bag of translations*, which was attempted by many researchers as an approximate document translation approach [1,2,3,4]. QT resolves source language translation ambiguity, while PDT disambiguates target language translation ambiguity. That is, QT and PDT are expected to show different disambiguation performance. So, we attempted to combine the two methods.

The remainder of this paper is as follows. Section 2 describes word-ngram coupling methods, and Section 3 explains query translation and PDT techniques. At NTCIR-4, we participated in all SLIR (Single Language IR) tracks (CJK and English) and all Korean-related BLIR (Bi-Language IR) tracks (KC, CK, KJ, JK, KE, EK). Section 4 reports these NTCIR-4 retrieval results and our discussion. Finally, Section 5 gives conclusions.

## 2 Monolingual Information Retrieval

### 2.1 Coupling Words and N-grams

In CJK indexing, n-grams enable complete document representation at the surface term level, although trigrams or more alone do not guarantee completeness considering that the average word length of CJK is approximately 2. Compared to n-grams, words as an index unit are prone to omit necessary concepts, owing to the word segmentation difficulty in CJK languages. However, n-grams provide the distributed representation for each concept, while words themselves represent their concepts. Thus, in terms of concept specificity, words are superior to n-grams. Therefore, combination of words and n-grams are expected to produce better retrieval performances.

Thus far, many Chinese IR papers report performance improvements when words and n-grams are combined at the index level or at the ranked list level. However, there is little Japanese or Korean IR literature that attempts to couple words and n-grams in a large scale. Thus, at NTICR-4, our team seeks to investigate the effects of coupling words and n-grams across CJK languages.

**Table 1. Coupling Stages of Words and N-grams**

| Coupling Stage | Coupling Method | # Indexes |
|---|---|---|
| Index creation | | Single |
| Term weighting | tf merging<br>df merging<br>Weight merging | Multiple |
| Ranked list | Score merging | Multiple |

Table 1 shows various stages of coupling words and n-grams. We have tested word-ngram coupling at term weighting stage, using NTCIR-3 Korean SLIR test set. For example, tf can be summed or averaged over word and ngram indexes. df can be summed or unioned over document postings of words and ngrams. However, the results are not remarkable. So, at NTCIR-4, we are interested in the ranked list stage.

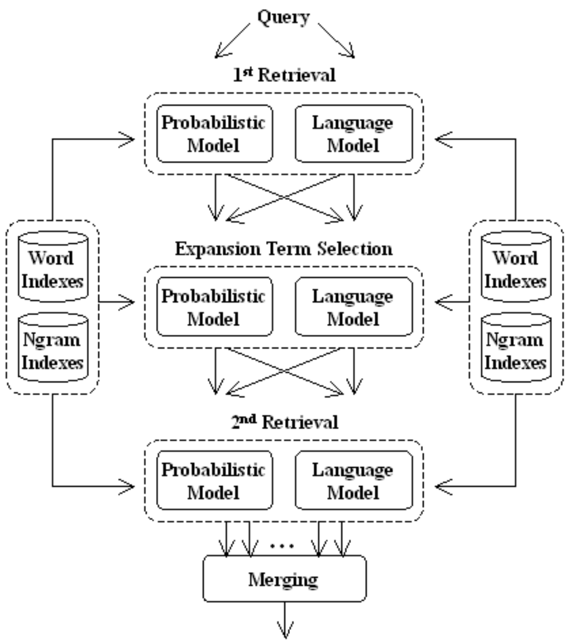


**Figure 1. Combining Words and N-grams from Different Retrieval Models**

The basic idea for coupling words and n-grams at the ranked list stage is to use a variety of ranked lists that are obtained by applying different retrieval models to words and n-grams, respectively. Our intuition is that different retrieval models will show varying performances both on different indexes and different queries.

In order to create various ranked lists, we selected two representative retrieval models, such as the Okapi probabilistic model approximated by Singhal [5], and Jelinek-Mercer language model with its lamda parameter set to 0.75. Then, on different index units (words and n-grams), different retrieval models were applied at each of initial retrieval, selection of expansion terms, and second retrieval. Figure 1 shows the flow of generating various ranked lists.

After preliminary experiments using NTCIR-3 test sets, we selected three coupling strategies for CJK languages from a total of 16 combinations, as shown in Table 2. In Table 2, P and L denotes the Okapi probabilistic model and Jelinek-Mercer language model, respectively. For English SLIR, wPLP and wPPP were used for merging ranked lists, because we did not create n-gram indexes for English.

**Table 2. NTCIR-4 CJK SLIR Coupling Strategy**

| | Index Unit | | |
|---|---|---|---|
| | Word | N-gram | |
| Initial Retrieval | P | P | L |
| Feedback | L | P | L |
| Second Retrieval | P | P | L |
| Abbreviated Notation | wPLP | nPPP | nLLL |

Feedback in Table 2 means the retrieval model to be used in order to select expansion terms. In the case of the Okapi model (P), Robertson selection value [6] was used for selecting feedback terms, and language model (L) used Ponte's ratio formula [7].

Robertson selection value $S(t)$ for a term $t$ is defined as follows.

$$S(t) = r_t \times \log \frac{(r_t + 0.5)(N - n_t - R + r_t + 0.5)}{(n_t - r_t + 0.5)(R - n_t + 0.5)} \qquad (1)$$

In formula (1), $r_t$ is the number of feedback documents containing term $t$, $n_t$ is the total number of documents containing term $t$, $N$ is the collection size, and $R$ is the number of feedback documents.

Ponte's ratio formula [7] is as follows.

$$S(t) = \sum_{k=1}^{R} \log \left( \frac{P(t \mid d_k)}{\frac{cf(t)}{N}} \right) \qquad (2)$$

In formula (2), $cf$ means collection frequency of term $t$, and $N$ is the collection size. $R$ is the number of feedback documents. $d_k$ is top k-th document.

At NTCIR-4, R was set to 15 and 10 in formula (1) and (2), respectively. In addition, the number of expansion terms was 50 and 300 in formula (1) and (2), respectively. Finally, in merging ranked lists, a simple score sum was used.

### 2.2 Term Extraction

Table 3 summaries terms used at NTCIR-4 CLIR task. Basically, we extracted bi-grams and words as CJK index terms, and created separate indexes: word-based and ngram-based. As English terms, *pgram(n,k)* means a string concatenation of prefix **n**-grams of **k** consecutive words. For example, from "… computer virus …", a pgram(3,2) *com_vir* is obtained if we extract a concatenation of prefix **3**-grams of **2** consecutive words. *pgram* was devised to support phrasal terms in English.

**Table 3. Terms used at NTCIR-4**

| Language | Terms | Stoplist |
|---|---|---|
| Chinese | Bi-gram, word | None |
| Japanese | Bi-gram, word | None |
| Korean | Bi-gram, word | 374 words |
| English | Word, pgram(3,2) | 322 words |

In order to identify words in CJK languages, we used CJK morphological analyzers developed at our laboratory. For English, stemming was performed using Porter's algorithm.

### 2.3 Preliminary Experiments

Table 4 shows preliminary Korean SLIR experiments using the NTCIR-3 test set. For example, *wP*-- means a probabilistic retrieval model (*P*) based on a word index (*w*), without any feedback (--). *wPLP* corresponds to the second column in Table 2. The last row in Table 4 refers to merging ranked lists of three different retrieval models wPLP, nPPP, and nLLL. Performances are incrementally improved through the processes of feedback and coupling.

**Table 4. Preliminary Korean SLIR Experiments using NTCIR-3 Test Set**

| | T | D | C | TDNC |
|---|---|---|---|---|
| wP-- | 0.2887 | 0.2561 | 0.3214 | 0.4029 |
| wL-- | 0.2880 | 0.2153 | 0.3198 | 0.3734 |
| nP-- | 0.3247 | 0.2805 | 0.3328 | 0.4267 |
| nL-- | 0.3090 | 0.2496 | 0.3277 | 0.4203 |
| wPLP | 0.3878 | 0.3412 | 0.3939 | 0.4804 |
| nPPP | 0.3435 | 0.3139 | 0.3560 | 0.4701 |
| nLLL | 0.3477 | 0.3467 | 0.3928 | 0.4809 |
| wPLP+nPPP+nLLL | 0.3818 | 0.3738 | 0.4365 | 0.5011 |

## 3 Cross-Language Information Retrieval

At NTCIR-4 BLIR, we were interested in the following language pairs: KC, CK, KJ, JK, KE, and EK. In BLIR, a query language can be translated into a document language (query translation), or vice versa (document translation). We adopted a hybrid approach that combines query translation and document translation at the ranked list level.

### 3.1 Bilingual Dictionaries

Table 5 shows some statistics about our bilingual dictionaries used at NTCIR-4 CLIR. KE and EK dictionaries are general-purpose dictionaries gathered from Web dictionaries designed for human users. KJ, JK, KC, and CK dictionaries are extracted from transfer dictionaries created for machine translation (MT) systems[1].

**Table 5. Bilingual Dictionary Statistics**

| | # of translation pairs | # of source language terms | Dictionary ambiguity |
|---|---|---|---|
| KE | 421,459 | 170,025 | 2.48 |
| EK | 513,015 | 202,099 | 2.54 |
| KJ | 420,650 | 303,199 | 1.39 |
| JK | 434,672 | 399,220 | 1.09 |
| KC | 113,312 | 81,750 | 1.39 |
| CK | 127,560 | 109,614 | 1.16 |

### 3.2 Query Translation (QT) Method

While preparing NTCIR-4 BLIR track, we developed a statistical word sense disambiguation (WSD) method [8] for our query translation approach. Unfortunately, however, for some reasons, we could not complete gathering all required probabilities for the method from NTCIR-4 document collections. So, we submitted our official BLIR runs using a default QT method, where a target language query is created from a source language query by replacing each source language query term with all its translations in a bilingual dictionary.

### 3.3 Pseudo Document Translation (PDT)

Pseudo document translation (PDT) translates a target language document into a source language pseudo document at the surface term level. That is, each (target language) document term is replaced by all its (source language) translation equivalents each

[1] Our laboratory has developed MT systems for CJK: COBALT-JK/KJ (Collocation-based Language Translator between Korean and Japanese), and TOTAL (Translator Of Three Asian Languages)

of which is attached with a translation probability. More formally, given a target language document $D_T = t_1,\ldots,t_n$ where $t_i$ is a i-th term in $D_T$, its source language pseudo document is $D_S = \langle s_{11}, p_{11}\rangle,\ldots, \langle s_{ij}, p_{ij}\rangle,\ldots, \langle s_{nm}, p_{mn}\rangle$, where $s_{ij}$ is a j-th translation of $t_i$, and $p_{ij}$ is a translation probability that $t_i$ is translated into $s_{ij}$. $p_{ij}$ can be estimated from a source language corpora. At NTCIR-4, however, we simply used all $p_{ij}$ set to 1.0.

Predictably, PDT generates an ambiguous document representation. We expect that some of this ambiguity will be resolved by document co-occurrence constraints of source language query terms.

Compared to normal document translation by machine translation systems, PDT requires only a bilingual lexicon, and time and space complexity are not severe. Moreover, an existing monolingual IR system can be easily adapted to a CLIR system, by creating a source language based index database from a target language based index database through PDT.

Compared to query translation, PDT requires more storage and additional dictionary lookup time, when indexing documents. However, query translation needs more retrieval time.

Table 6 shows preliminary BLIR experiments based on PDT using NTCIR-3 test set for KC and KJ language pairs. 'Base' means a default QT method of Section 3.2.

**Table 6. Preliminary BLIR Experiments based on PDT using NTCIR-3 Test Set**

| | | T | D | C | TDNC |
|---|---|---|---|---|---|
| KC | Base | 0.1061 | 0.0913 | 0.1232 | 0.1710 |
| | PDT | 0.1568 | 0.1422 | 0.1766 | 0.2045 |
| KJ | Base | 0.2120 | 0.2155 | 0.2465 | 0.2967 |
| | PDT | 0.2546 | 0.2418 | 0.2622 | 0.3347 |

## 3.4 Language-Dependent PDT

### Transliteration of Chinese characters

The CJK languages share ideographic Chinese characters. For example, Japanese and Korean uses *Kanji* and *Hanja* (or Sino-Korean) as Chinese characters, respectively. Throughout China, Japan, and Korea, the meanings of Chinese characters are almost the same, although the forms may differ slightly according to particular localizations of Chinese characters.

In Korea, however, Sino-Korean words are typically written in *Hangul*, the Korean alphabet. *Hangul* is not ideographic, but alphabetic and phonetic. Generally, there is a many-to-one mapping between *Hanja* and *Hangul*. For example, both 漢代 (Han dynasty) and 寒帶 (the frigid zone) are written as the same word 한대 in Korean. So, most Sino-Korean words written as *Hangul* may cause homographs.

In CLIR retrieving documents in Chinese or Japanese using Korean as a query language, however, transliterating Chinese characters in those documents into *Hangul* helps to increase the vocabulary coverage of a bilingual lexicon, which alleviates the vocabulary mismatch problem. For example, in a Korean-to-Japanese (KJ) bilingual lexicon for KJ CLIR, suppose that a Korean word 고궁 contains its translation equivalents 古宮 (an old palace), and 孤窮 (lonesome and poor). In that case, however, 故宮 (an old palace), and 固窮 (endure loneliness and poverty) are another possible translations of 고궁. Then, by transliterating *Kanji* characters in Japanese documents into *Hangul*, a Korean query term 고궁 can be matched with any one of 古宮, 孤窮, 故宮, and 固窮, independent of KJ dictionary coverage. Note that, in the above example, translations of 고궁 may not be legal words in Japanese, because they were selected from a Korean dictionary only for this example.

In addition, transliteration of Chinese characters into *Hangul* may be useful for mitigating an unknown word problem in KC or KJ CLIR. For example, an unknown word 金大中 (a former Korean president, Kim Dae Jung) in Chinese or Japanese documents can be matched with a Korean query term 김대중 by transliterating 金大中 into 김대중.

In this respect, in translating Chinese or Japanese documents into Korean pseudo documents for KC or KJ CLIR at NTCIR-4, their Chinese characters (*Hanzi* or *Kanji*) were transliterated into *Hangul*, and inserted into Korean pseudo documents together with dictionary translations.

## 3.5 Combination of Query Translation and Pseudo Document Translation

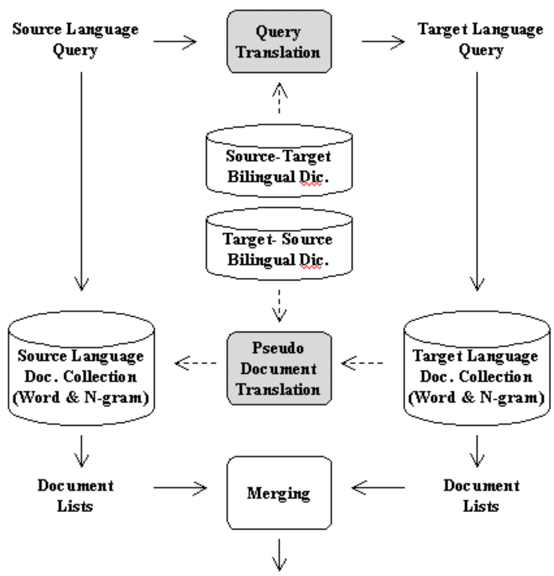


**Figure 2. Combining Query Translation and PDT**

Query translation and PDT has different characteristics in resolving translation ambiguity. In a default query translation, its disambiguation effect occurs by co-occurrences of target language query terms within the same document. In PDT, however, co-occurrences of source language query terms within the same document influence disambiguation of the corresponding target language terms in the document.

Thus, query translation resolves source language translation ambiguity, while PDT disambiguates target language translation ambiguity. Given a particular language pair for CLIR, one of the two translation directions would be easier than the other in terms of translation ambiguity resolution. In other words, query translation and PDT are expected to have different influence on the same queries.

Therefore, we combine query translation (QT) and PDT, as shown in Figure 2. Under our BLIR architecture of Figure 2, we selected word-ngram coupling strategies for BLIR tracks at NTCIR-4, as shown in Table 7.

**Table 7. NTCIR-4 BLIR Coupling Strategy**

| Language Pair | Methods | |
|---|---|---|
| | QT | PDT |
| KC | nPLP | nPLP |
| KE | wPLP | None |
| KJ | wPLP | nPLP |
| EK, CK, JK | wPLP, nPLP | None |

The reason why we use n-grams in BLIR is to alleviate word mismatch problem by increasing dictionary coverage on a target document collection. The procedure of ngram-based BLIR is as follows. First, a document collection is indexed with n-grams. Next, a word-based target language query is obtained from a source language query by looking up a word-based bilingual dictionary. Then, each word-based query term is sliced into several ngram-based query terms to produce an ngram-based target language query. Finally, the ngram-based query is used to search an ngram-based index.

# 4 Retrieval Results and Discussion

This section reports the retrieval results of our official runs submitted to NTCIR-4 CLIR task, using non-interpolated average precision (AvgPre). Each topic has four fields: title (T), description (D), narrative (N), and concepts (C). We submitted our runs using T, D, C, DN, and TDNC. Relevance judgments are divided into two categories: rigid, and relaxed. In this paper, we report all retrieval results using AvgPre based on relaxed judgments. Details about test collections and relevance judgments can be found in the NTCIR-4 overview paper [9].

## 4.1 CJKE SLIR Track

Table 8 shows the retrieval results of Chinese (C), Japanese (J), Korean (K), and English (E) SLIR tracks. Largely, retrieval using longer topics obtained better results than using shorter topics across different languages. The bold face figures indicate retrieval results of our official runs. N/A means that the retrieval result is not available at this time.

In all languages, feedback techniques consistently improve performance, independent of different query types. Interestingly, the language model is not better than the probabilistic model at initial retrieval. However, at feedback stage, the language model recovers its performance enough to outperform probabilistic model. As we expected, different coupling strategies behave differently in Table 8. In addition, Table 8 shows that top n coupling strategies may be different according to languages. For example, our selection of nPPP, nLLL, and wPLP was determined from the experience on NTCIR-3 Korean SLIR experiments. So, in the case of NTCIR-4 Korean SLIR, performance differences are negligible among nPPP, nLLL, and wPLP. However, in Chinese or Japanese SLIR, performance differences are remarkable. Therefore, the fusion of three ranked lists having different retrieval characteristics was effective only on Korean.

**Table 8. SLIR Performance**

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| C | nP-- | 0.2297 | 0.2069 | 0.2562 | 0.2855 | 0.2911 |
| | nL-- | 0.2050 | 0.1823 | 0.2365 | 0.2708 | 0.2809 |
| | wP-- | 0.1603 | 0.1533 | 0.1789 | 0.2281 | 0.2358 |
| | nPPP | 0.2532 | 0.2398 | 0.2681 | 0.2983 | 0.3060 |
| | nLLL | 0.2699 | 0.2686 | 0.2856 | 0.3019 | 0.3046 |
| | wPLP | 0.1853 | 0.2016 | 0.2049 | 0.2503 | 0.2693 |
| | nPPP+nLLL+wPLP | **0.2584** | **0.2535** | **0.2703** | **0.2968** | **0.3103** |
| J | nP-- | 0.3650 | 0.3424 | 0.3496 | 0.4346 | 0.4570 |
| | nL-- | 0.3260 | 0.3101 | 0.3141 | 0.4274 | 0.4435 |
| | wP-- | 0.3647 | 0.3715 | 0.3426 | 0.4439 | 0.4561 |
| | nPPP | 0.3844 | 0.3842 | 0.3926 | 0.4539 | 0.4856 |
| | nLLL | 0.4056 | 0.4282 | 0.4207 | 0.4924 | 0.5024 |
| | wPLP | 0.4226 | 0.4103 | 0.3806 | 0.4715 | 0.4875 |
| | nPPP+nLLL+wPLP | **0.4211** | **0.4119** | **0.4105** | **0.4741** | **0.4963** |
| K | nP-- | 0.4515 | 0.4198 | 0.4450 | 0.5249 | 0.5598 |
| | nL-- | 0.4091 | 0.3674 | 0.4081 | 0.4896 | 0.5318 |
| | wP-- | 0.4285 | 0.4184 | 0.4370 | 0.5111 | 0.5383 |
| | nPPP | 0.4660 | 0.4347 | 0.4499 | 0.5610 | 0.6040 |
| | nLLL | 0.4967 | 0.4623 | 0.4496 | 0.5592 | 0.5873 |
| | wPLP | 0.4900 | 0.4771 | 0.4611 | 0.5806 | 0.5859 |
| | nPPP+nLLL+wPLP | **0.5226** | **0.4885** | **0.4846** | **0.5932** | **0.6212** |
| E | wP-- | 0.3704 | 0.3618 | 0.3681 | 0.4270 | 0.4503 |
| | wPLP | 0.3924 | 0.3697 | 0.3753 | 0.4755 | 0.4986 |
| | wPPP | N/A | N/A | N/A | N/A | N/A |
| | wPLP + wPPP | **0.3898** | **0.3670** | **0.3729** | **0.4731** | **0.4962** |

### 4.2 Korean-related BLIR Track

Table 9 shows retrieval results of Korean-involved BLIR tracks, such as KJ, JK, KC, CK, KE, and EK language pairs. The notations *qt* and *pdt* in Table 9 means the default QT method and the pseudo document translation method, respectively. The bold face figures indicate retrieval results of our official runs at NTCIR-4. The others correspond to our post-experiments. N/A means that the retrieval result is not available at this time.

JK and CK performed better than KJ and KC, respectively. We believe that one of the reasons is due to differences between degrees of translation ambiguities of the corresponding bilingual lexicons, as shown in Table 5. In addition, KJ and JK are better than KC and CK. At this time, we expect that the reason results from the difference between dictionary coverage. The KJ/JK dictionaries are three times more than KC/CK dictionaries, as shown in Table 5. Among other language pairs, EK performances are quite poor, compared to those of CK and JK. Currently, we are analyzing the cause of the result.

**Table 9. BLIR Performance**

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| KJ | wP-- (qt) | 0.2861 | 0.3039 | 0.3000 | 0.3763 | 0.3905 |
| | nP-- (pdt) | 0.3165 | 0.3207 | 0.3140 | 0.3909 | 0.4039 |
| | wP-- (qt) + nP-- (pdt) | 0.3234 | 0.3362 | 0.3241 | 0.4098 | 0.4229 |
| | wPLP(qt)+nPLP(pdt) | **0.3602** | **0.3601** | **0.3713** | **0.4471** | **0.4473** |
| JK | wP-- (qt) | 0.3559 | 0.3431 | 0.3451 | 0.4243 | 0.4450 |
| | nP-- (qt) | 0.3490 | 0.3501 | 0.3587 | 0.4536 | 0.4607 |
| | wP-- (qt) + nP-- (qt) | 0.3634 | 0.3666 | 0.3833 | 0.4632 | 0.4773 |
| | wPLP(qt)+nPLP(qt) | **0.4559** | **0.4306** | **0.4593** | **0.5383** | **0.5446** |
| KC | nP-- (qt) | 0.1436 | 0.1456 | 0.1584 | 0.1665 | 0.1778 |
| | nP-- (pdt) | 0.1551 | 0.1448 | 0.1567 | 0.1937 | 0.2057 |
| | nP-- (qt) + nP-- (pdt) | 0.1687 | 0.1731 | 0.1763 | 0.1992 | 0.2089 |
| | nPLP(qt)+nPLP(pdt) | **0.1892** | **0.1869** | **0.2028** | **0.2378** | **0.2469** |
| CK | wP-- (qt) | 0.3466 | 0.3193 | 0.3364 | 0.4004 | 0.4299 |
| | nP-- (qt) | 0.3572 | 0.3342 | 0.3466 | 0.4099 | 0.4355 |
| | wP-- (qt) + nP-- (qt) | 0.3663 | 0.3463 | 0.3557 | 0.4259 | 0.4538 |
| | wPLP(qt)+nPLP(qt) | **0.4343** | **0.4314** | **0.4083** | **0.5060** | **0.5138** |
| KE | wP-- (qt) | 0.1958 | 0.1876 | 0.2186 | 0.2031 | 0.2468 |
| | wPLP(qt) | **0.2647** | **0.2622** | **0.2860** | **0.2435** | **0.2805** |
| EK | wP-- (qt) | 0.1123 | 0.0955 | 0.0626 | 0.0729 | 0.0846 |
| | nP-- (qt) | N/A | N/A | N/A | N/A | N/A |
| | wP-- (qt) + nP-- (qt) | N/A | N/A | N/A | N/A | N/A |
| | wPLP(qt)+nPLP(qt) | **0.1260** | **0.1060** | **0.0627** | **0.0730** | **0.0849** |

In KJ and KC BLIR, PDT-based BLIR was slightly better than QT-based BLIR, except for only KC concept queries. The reason is as follows. QT-based BLIR retrieves noiseless target language documents, using a noisy target language query with incorrect translations, while PDT-based one retrieves noisy source language documents with incorrect translations, using a noise-free source language query. So, in QT-based BLIR, incorrect query terms with high idf values may cause negative effect on retrieval effectiveness. In PDT-based BLIR, correct document terms may not exhibit their full weights, since each document term has lower idf value than its normal value. The reason is that after PDT, a document posting of each document term is the union of document postings of its original terms. We believe that incorrect query terms with high idf values causes worse effect on BLIR performance than correct document terms with lower idf values.

In addition, in KJ and KC BLIR, we obtained a consistent improvement, through the combination of QT and PDT. This result indicates that disambiguation effects on a document are different, according to languages. In other words, documents on which QT-based disambiguation works are not the same as the documents on which PDT-based one is effective, because some relevant documents may have lower translation ambiguity in terms of the query language than that in terms of the document language, and vice versa.

Moreover, the coupling of words and n-grams as well showed a consistent improvement in JK and CK BLIR. Interestingly, ngram-based BLIR was better than word-based one, except for only JK title queries. This means that simple conversion of word translations into ngrams can mitigate word mismatch problem of word-based indexing to a certain degree, although there are noisy ngrams.

Predictably, feedback techniques improve performance significantly in JK and CK language pairs, noticeably in other language pairs except EK pairs. So, this result reconfirms that the use of feedback is a cheap performance-enhancing device in BLIR, independent of language pairs.

Table 10 shows the distribution of averages of AvgPre across different combinations of query fields by language pairs. Except EK, all bilingual retrieval showed about 50% ~ 90% performances of full-fledged single language information retrieval.

**Table 10. Distribution of Averages of AvgPre of Table 9**

| | Average of AvgPre | % SLIR |
|---|---|---|
| KJ | 0.3972 | 0.90 |
| JK | 0.4857 | 0.90 |
| KC | 0.2127 | 0.76 |
| CK | 0.4588 | 0.85 |
| KE | 0.2674 | 0.65 |
| EK | 0.0905 | 0.17 |

## 5 Conclusions

For NTCIR-4 SLIR, we employed a word-ngram coupling strategy that combines several ranked lists generated from words and n-grams indexes by differentiating both retrieval models and expansion term selection schemes. For NTCIR-4 BLIR, we experimented with a hybrid strategy that combines a default query translation and a document translation based on a pseudo document translation scheme.

## Acknowledgements

This work was supported by the KOSEF through the Advanced Information Technology Research Center (AITrc) and by the BK21 project.

## References

[1] Oard, D.W., Levow, G.A., and Cabezas, G.I. CLEF Experiments at Maryland: Statistical Stemming and Backoff Translation. *CLEF-2000*: 176-187. 2000.

[2] Oard, D.W. and Wang, J. NTCIR-2 ECIR Experiments at Maryland: Comparing Pirkola's Structured Queries and Balanced Translation. *NTCIR-2001*. 2001.

[3] Oard, W. and Ertunc, F. Translation-based Indexing for Cross-Language Retrieval. *ECIR-2002*: 324-333. 2002.

[4] Darwish, K. and Oard, D.W. CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval. *TREC-2002*. 2002.

[5] Singhal, A., Salton, G., Mitra, M., and Buckley, C. Document Length Normalization. Information Processing and Management, 32(5):619-633. 1996.

[6] Robertson, S.E. On Term Selection for Query Expansion. *Journal of Documentation*, 46:359-364. 1990.

[7] Ponte, J. A Language Modeling Approach to Information Retrieval. *Ph.D. Thesis*. University of Massachusetts at Amherst. 1998.

[8] Kang, I.S., Na, S.H., and Lee, J.H. Influence of WSD on Cross-Language Information Retrieval. *Proceedings of the 1st International Joint Conference on Natural Language Processing*. 2004.

[9] Kishida, K., Chen, K.H, Lee, S.H., Kuriyama, K., Kando, N, Chen, H.H., Myaeng, S.H., and Eguchi, K. Overview of CLIR Task at the Fourth NTCIR Workshop. *NTCIR-2004*. 2004.