# Combination Approaches in Information Retrieval: Words vs. N-grams, and Query Translation vs. Document Translation

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. For monolingual information retrieval, we use a combination strategy that integrates words and n-grams at the ranked list level. In combining words and n-grams, we concentrate on generating several ranked lists showing different retrieval characteristics on word and n-gram indexes by incorporating feedback schemes. For cross-language information retrieval, we attempt a dictionary-based bi-directional combination of query translation and document translation. For both query translation and document translation, their naïve translation is used. Experimental evaluations on CJK monolingual and KC/KJ cross-language retrieval give promising results on our combination approaches: words vs. n-grams, and query translation vs. document translation.*

**Keywords:** *CJK Information Retrieval, fusion strategies, Cross-language information retrieval, Combination of query translation and document translation*

## 1 Introduction

Unlike English, Chinese and Japanese do not use word delimiters in a normal text. In addition, in Korean, no word boundaries exist within an eojeol, a Korean spacing unit that corresponds to a phrasal unit in English. Thus, word segmentation is crucial for the three Asian languages we designate CJK (Chinese, Japanese, and Korean). However, unknown words and segmentation ambiguity obstruct correct word segmentation, resulting in incomplete document representation in information retrieval (IR). To overcome the problem, CJK IR systems employ character-based n-grams as well as words, as indexing units.

In CJK indexing, n-grams are more preferable than words, in terms of complete document representation based on surface terms, because n-grams have the potential to create a superset of a complete lexical term space for a document collection. However, tri-grams or more do not guarantee the completeness, considering that the average word length of the CJK languages is approximately 2. Compared to n-grams, words as an index unit are prone to omit keywords corresponding to necessary concepts, owing to the word segmentation difficulty in the CJK languages. On concept specificity, however, words are superior to n-grams, since n-grams provide only a distributed noisy representation for a keyword that encodes certain concepts, while words enable concentrated representation.

Therefore, in order to mutually compensate deficiencies of words and n-grams, combination approaches of words and n-grams are advocated to obtain better retrieval performance. Thus far, several Chinese monolingual IR (MLIR) literature [8, 11, 12, 13, 15, 16, 27] reports a little success when combining words and n-grams at the indexing unit level or at the ranked list level. However, there are few Japanese or Korean MLIR experiments that evaluate coupling words and n-grams at a large scale.

So, this paper empirically investigates the impact of coupling words and n-grams on CJK monolingual retrieval environment. In combining words and n-grams, we concentrate on generating several ranked lists showing different retrieval characteristics on word and n-gram indexes by incorporating feedback schemes.

In CLIR (Cross-Language IR), a query language can be translated into a document language (query translation), or vice versa (document translation). Query translation (QT) is more widespread, because it is simple, lightweight, and flexible. 'Flexible' means that, in QT, a modification of a bilingual dictionary can be promptly reflected in a CLIR

system. In addition, QT enables its modular utilization in the sense that it can easily convert any existing MLIR systems into its cross-lingual versions, without changing the underlying MLIR systems. However, QT severely suffers from the translation ambiguity problem, resulting from insufficient disambiguation context of queries and translation resources not designed for machine translation purposes. So, QT methods were much explored in order to develop efficient algorithms for resolving translation ambiguity.

On the other hand, document translation (DT) normally requires machine translation (MT) systems or statistical translation methods such as IBM statistical models [2]. Thus, it is computationally expensive in terms of translation time and additional index storage, so it cannot be easily repeated to reflect the changes of MT systems. Moreover, DT generates restricted document representation, which is highly dependent on the performance of a MT system. Therefore, DT is practically barely attempted, although some large-scale document translation approaches [1, 14, 17] were recently reported.

We believe that QT and DT have different disambiguation effects on queries, since they try to resolve source language and target language translation ambiguity, respectively. In other words, some query terms (or queries) could be more effectively disambiguated by QT than DT, or vice versa, because, for the same query, QT and DT may have different degrees of translation ambiguities, according to different translation directions of the same language pair.

So, we adopt a hybrid approach that combines query translation and document translation for CLIR. For both QT and DT, their naïve translation is applied without any separate disambiguation module. That is, QT converts a source language query into a target language query, by simply expanding all (target language) dictionary translations for each source language query term. Similarly, DT creates a source language document for each target language document, by simply replacing all (source language) dictionary translations for each document term.

In summary, this paper empirically explores the following two issues based on our participation systems in the NTCIR-4 CLIR evaluation workshop [10].

1. Combination of words and n-grams in MLIR (for CJK languages)
2. Hybridization of query translation and document translation in CLIR (for Korean-to-Chinese (KC) and Korean-to-Japanese (KJ) language pairs)

To avoid confusion, this paper uses a source language and a target language to refer to a query language and a document language, respectively. For example, in Korean-to-Chinese CLIR, Korean and Chinese are a query language and a document language, respectively. In addition, a source language and a target language always refer to Korean and Chinese, respectively, independent of translation directions in QT (Korean-to-Chinese) or DT (Chinese-to-Korean).

The remainder of this paper is as follows. Section 2 and 3 describe our word/n-gram coupling method for MLIR, and its experimental results, respectively. Section 4 explains a combination approach of QT and DT for CLIR. Section 5 reports CLIR experimental results. Finally, Section 6 gives conclusions.

## 2 Monolingual Information Retrieval

### 2.1 Coupling Words and N-grams

Table 1 shows various stages of coupling words and n-grams in a retrieval system. For example, both words and n-grams are collected from documents to create a single index. At the term weighting stage, term frequencies can be summed over word and n-gram indexes, or document frequencies can be summed or unioned over document postings of words and n-grams, or final term weights obtained from two indexes can be interpolated. At the ranked list stage, we can merge two ranked lists obtained by retrieving documents on each of word and n-gram indexes. We have tested all these coupling methods in Table 1, using NTCIR-3 Korean MLIR test set [4]. However, the results were not remarkable, except for coupling at the ranked list level. So, we selected the ranked list stage for a combination of words and n-grams.

**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 |

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 first selected two representative retrieval models: Okapi probabilistic model approximated by Singhal et al. [26], and Jelinek-Mercer language model [28] with its lamda parameter set to 0.75. Then, on different index units (words and n-grams), the two retrieval models were applied at each retrieval stage of initial retrieval, selection of expansion terms, and second retrieval. Figure 1 shows the flow of generating various ranked lists. In Figure 1, a total of 16 different ranked lists are generated.

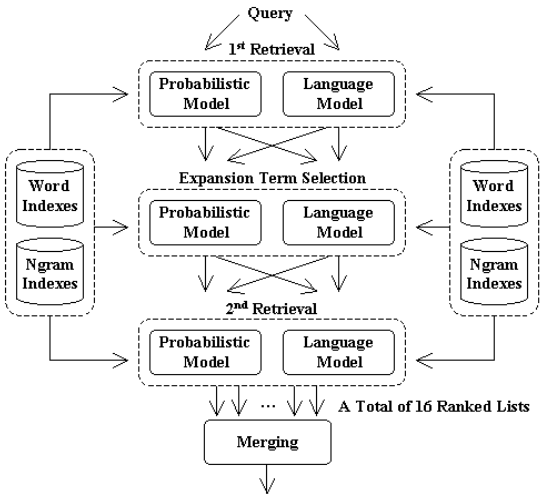


**Figure 1. Combining Words and N-grams from Different Retrieval Models**

After preliminary experiments using NTCIR-3 Korean test set, from a total of 16 combinations, we selected top three ranked lists to be used for merging words and n-grams. A selection measure was the non-interpolated mean average precision (MAP) value, and our selection constraint was to include at least one ranked list from each of word-based and n-gram-based ranked lists. Table 2 shows the selected three ranked lists, where P and L denotes Okapi probabilistic model and Jelinek-Mercer language model, respectively.

**Table 2. NTCIR-4 CJK MLIR 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 a retrieval model of which feedback scheme is used in order to select expansion terms. In the case of the Okapi model (P), Robertson selection value [25] was used for selecting feedback terms, and for the language model (L), we selected Ponte's ratio formula [23]. Robertson selection value S(t) for a term t is defined as Formula (1), where rt is the number of feedback documents containing term t, and nt is the total number of documents containing term t. N is the collection size. R is the number of feedback documents.

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

Ponte's ratio formula [23] sorts promising feedback terms using Formula (2), where *cf* means a collection frequency of term *t*, and $d_k$ is the top k-th document. *N* and *R* have the same meaning as those in Formula (1).

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

## 2.2 Term Extraction

Table 3 summaries terms used at our experiments. We extracted bi-grams and words as CJK index terms from documents, and created separate indexes: word-based and n-gram-based. In order to identify words in CJK languages, we used a CJK tagger developed at our laboratory.

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

| | Terms | Stop-list | CJK Tagger | |
|---|---|---|---|---|
| | | | Lexicon | Accuracy |
| C | Bi-gram, word | None | 160,562 | 95% |
| J | Bi-gram, word | None | 430,251 | 95% |
| K | Bi-gram, word | 374 words | 270,479 | 95% |

In our case, bi-grams are not word-based, but character-based for all CJK languages. More precisely, for Chinese, a sequence of Hanzi characters is sliced into overlapping two-character sub-strings to produce bi-grams. Similarly, for Japanese, from a sequence of the same character class (Hiragana, Katagana, or Kanji), bi-grams are extracted. This character-class-based n-gram indexing was proposed by Ogawa and Matsuda [21]. For Korean, Hanja characters are first converted into the corresponding Hangul characters by a Hanja-to-Hangul mapping table, and then, bi-grams are generated from a sequence of Hangul characters. In addition, for all CJK documents, English words were stemmed using the Porter's algorithm.

To extract query terms from topic files of NTCIR-4 test set, the same method was applied.

# 3 Experimental Results of CJK Monolingual Retrieval

This section reports the experimental evaluations on our word/n-gram coupling strategy using NTCIR-4 CJK MLIR test sets. Each topic has four fields: title

(T), description (D), narrative (N), and concept (C). We evaluated our system 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 non-interpolated mean average precision (MAP) based on relaxed judgments. Details about test collections, topics, and relevance judgments can be found in the NTCIR-4 overview paper [10].

In our experiments, R (the number of feedback documents) was set to 10 and 15 in Formula (1) and (2), respectively. In addition, the number of expansion terms was fixed to 300 and 50 in Formula (1) and (2), respectively. The parameter values were determined from preliminary experiments using NTCIR-3 Korean test set. Finally, in merging ranked lists, a simple score sum was used.

**Table 4. CJK Monolingual Retrieval Results**

| | | 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** |
| | Fusion | **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 |
| | Fusion | **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** |
| | Fusion | **0.5226** | **0.4885** | **0.4846** | **0.5932** | **0.6212*** |

Table 4 shows the retrieval results of Chinese (C), Japanese (J), and Korean (K) MLIR. Note that wP indicates a probabilistic model (P) based on a word index (w), without any feedback step. Similarly, nP or nL is a probabilistic or language model based on an n-gram index (n), respectively, without using a feedback loop. nPPP, nLLL, and wPLP were defined at Table 2. The bold face figures indicate retrieval results of our official runs at NTCIR-4. The underlined figures indicate that retrieval results are the best performance at NTCIR-4. The Chinese and Japanese results correspond to the medium level of performance among all NTCIR-4 participants. For Korean, our system performed close to the best performance (for short queries), or obtained the best (for long queries).

A symbol '*', or '**' is attached to the retrieval result that is statistically significant at a significance level of 0.05 or 0.01, respectively. Statistical tests were performed on each of the following pairs of retrieval methods: nPPP and nP, nLLL and nL, wPLP and wP. A fusion model was considered statistically significant only if the fusion model showed statistical differences for each of nPPP, nLLL, and wPLP. As a significance test, we used the sign test, following a Rijsbergen's argument [24] about the validness of statistical tests.

Comparing words (wP) and n-grams (nP) using the same probabilistic model at the initial retrieval, nP remarkably outperforms wP for Chinese, although statistical difference at the 5% error level was identified only on D and DN query types. For Korean, nP was slightly better than wP. For Japanese, any pair of wP and nP across different query types did not noticeably differ. Thus, this comparison confirms that n-gram-based retrieval performs close to or better than word-based one in CJK languages.

Interestingly, there is clear difference in performances of initial retrieval across CJK languages. Even n-gram-based initial retrievals (nP or nL) showed 0.2351 ~ 0.2539 for Chinese, 0.3642 ~ 0.3897 for Japanese, and 0.4412 ~ 0.4802 for Korean, respectively. For NTCIR-3 CJK MLIR test sets, n-gram-based initial retrieval showed 0.2874, 0.3420, and 0.3562 for Chinese, Japanese, and Korean, respectively. The figures are averages over non-interpolated average precision values of different query types (T, D, C, DN, and TDNC). For the moment, we cannot say the reason, although the common thing is that Japanese and Korean is better than Chinese in n-gram-based retrieval.

Compared to Japanese and Korean, Chinese word-based retrieval (wP) performs clearly less than n-gram-based one (nP). We believe that the reason is related to (1) the unknown real performance of our word segmentation systems on unrestricted corpora, and (2) different degrees of complexity of the segmentation problem for each language. For the first point, considering that our CJK tagger showed a similar performance on controlled test sets, we believe that the low performance of Chinese in Table 4 partly results from the relatively small lexicon size of it (see Table 3). Thus, for Chinese, more unknown words could have prevented word identification much severely. For the second point, Chinese is the most difficult in segmentation. Korean employs delimiters between eojeols, which is similar to a phrasal unit in English. Japanese do not use any word boundaries like Chinese. However, as segmentation clues, Japanese has three different character classes such as Katagana, Hiragana, and Kanji, which are easily identified.

Finally, from the results of Table 4, our fusion method for coupling words and n-grams works only in Korean. That is, only for Korean, the fusion model showed better performance than all three combination models (nPPP, nLLL, and wPLP), while, for Chinese and Japanese, the fusion simply results in

averaging three combination models. We believe that the reason is that the three combination models for fusion were selected from the experimental evaluations on a Korean test set, neither Chinese nor Japanese. Thus, it is believed that probably there exist different sets of top-ranked combination models for Chinese or Japanese, respectively.

# 4 Cross-Language Retrieval

## 4.1 QT vs. DT

Our focus in CLIR is to investigate combining effects of query translation (QT) and document translation (DT) on CLIR retrieval effectiveness, using a dictionary-based naïve translation for both QT and DT. A dictionary-based naïve translation means that each query (or document) term is simply replaced by all its translations in a bilingual dictionary, without performing any disambiguation strategy. A naïve translation can be done on a word-by-word or phrase-by-phrase basis.

Compared to normal document translation by machine translation systems, a naïve DT requires only a bilingual lexicon, and time and space complexity are not severe. Moreover, existing monolingual IR systems can be easily converted to CLIR systems, by creating a source language based index database from a target language based index database through a naïve DT. Thus, the naïve document translation has been attempted by several CLIR researchers [3, 5, 6, 7, 18, 19], as an alternative solution of normal document translation by machine translation (MT) systems. More precisely, they tried to improve the naïve document translation with some disambiguation devices.

Chen and Gey [3, 7] translate a document on a word-by-word basis using a one-to-one bilingual lexicon, which is produced by individually translating all words in a document collection into a query language by an MT system. They call this method *fast document translation*. Similarly, in the approach of Fujii and Ishikawa [6], a document is translated on a phrase-by-phrase basis using an MT system.

Oard and other researchers [5, 18, 19] devised a balanced version for the naïve document translation, where 'balanced' means that each document term is translated into the fixed (or balanced) number of translations in a translated document. First, they order translations for each entry in a bilingual dictionary using their unigram frequencies from a corpus. Then, if the fixed (or balanced) number of translations is set to n, each document term is replaced by its top n translations from the ordered bilingual dictionary. If a document term has only m translations (less than n) in a dictionary, the other n-m translations are obtained from the m translations in a round-robin way to make a total of n translations.

In summary, one group of researchers attempted an unambiguous version of the naïve document translation, and the other group tried a balanced version of it. In terms of disambiguation devices, the former utilized MT systems, and the latter relied on corpus statistics. In the viewpoint of word sense disambiguation (WSD), it is believed that *fast document translation* selects the first sense encoded by a bilingual dictionary of MT systems, and balanced document translation is similar to choosing the most frequent top n senses. Thus, an unambiguous or balanced version of the naïve document translation may undermine document representation, because their schemes of selecting translations do not rely on any context in documents.

In this paper, however, we employ purely the naïve document translation without applying any disambiguation strategies, in order not to omit any probable translations from a translated document. Actually, the naïve document translation is the index time implementation of the Pirkola's method [22], as Oard and Ertunc [20] mentioned. Considering that the Pirkola's method has been very effective on finding relevant target language documents in many CLIR experiments, the naïve document translation itself is expected to help in crossing the language barrier in CLIR.

**Table 5. Disambiguation Effect of QT and DT**

| | Disambiguation Context | | Disambiguation Effect |
|---|---|---|---|
| | Query | Document | |
| Naïve QT | Noisy | Clean | Resolves query language translation ambiguity |
| Naive DT | Clean | Noisy | Resolves document language translation ambiguity |

Table 5 compares disambiguation effects of QT and DT in its naïve translation. Underlying intuition is that correct target language query terms in QT tend to co-occur in target language documents, and correct source language document terms in DT is more likely to co-occur in the original source language query than incorrect ones.

In a naïve QT mode, a target language query has lots of incorrect target language translations. However, documents are their original forms. So, in this case, a retrieval system implicitly disambiguates source language translation ambiguity of queries by matching translated noisy query terms with clean document terms. A naïve DT generates an ambiguous document representation. This is, a translated document has many incorrect source language translations. However, a source language query maintains its original form. So, a naïve-DT-based retrieval system resolves target language translation ambiguity (of matched original document terms) by matching clean source language query terms with translated noisy document terms.

Therefore, a hybrid of QT and DT is advocated even in its naïve translation mode, since different translation directions of the same language pair are

expected to differently influence retrieving relevant documents in CLIR.

## 4.2 Bilingual Dictionaries

Table 6 shows some statistics about our bilingual dictionaries used at NTCIR-4 CLIR, where KC and KJ CLIR were experimented. KJ, JK, KC, and CK dictionaries were extracted from transfer dictionaries of our machine translation (MT) systems . For each language pair (KC or KJ), two versions of bilingual dictionaries are used. For example, our KJ CLIR system uses KJ and JK bilingual dictionaries for a naïve QT and a naïve DT, respectively.

In terms of dictionary ambiguity in Table 6, those of KJ and KC pairs are higher than those of JK and CK pairs, respectively. We believe that the reason is that Korean uses a Hangul writing system, which is not ideographic, but alphabetic and phonetic. Generally, there is a many-to-one mapping relationship between Chinese characters and Hangul characters. For example, both 漢代 (the Han dynasty) and 寒帶 (the frigid zone) are written as the same Hangul word '한대' in Korean.

**Table 6. Bilingual Dictionary Statistics**

| | # of translation pairs | # of source language terms | Dictionary ambiguity |
|---|---|---|---|
| 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 |

## 4.3 Combination of QT and DT

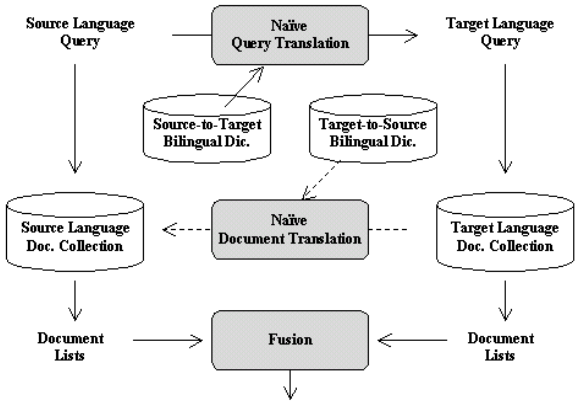


**Figure 2. Combining QT with DT**

As described in Section 4.1, query translation and document translation have different characteristics in resolving translation ambiguity. In a naive QT, its disambiguation effect occurs by co-occurrences of target language query terms within the same document. In a naïve DT, however, co-occurrences of source language query terms within the translated document influence disambiguation of the corresponding target language terms in its original document. Thus, query translation resolves source language translation ambiguity, while document translation 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, QT and DT are expected to have different influence on the same queries. We combine a naïve QT and a naïve DT at the ranked list level, as shown in Figure 2. For fusion of ranked lists, a simple summation was applied.

Thus far, some researchers tried a hybrid approach of QT and DT. In English-to-French bi-directional bilingual retrieval experiments, McCarley [14] translated queries and documents using a bi-directional statistical translation model based on the IBM model [2], which were trained on the same bilingual corpora. He showed that a combination of query and document translation outperforms query translation or document translation alone. In multilingual retrieval experiments, Braschler [1] also obtained better performance through the combination of QT and DT than either ones, where QT and DT were performed by MT systems. Chen and Gey [3] report similar improvements by coupling query translation and document translation. They translate documents using their *fast document translation* method (see Section 4.1), in order to obviate disadvantages of MT systems in CLIR.

In summary, all previous combination approaches reported a positive impact of coupling query translation and document translation, although the quality of translated documents varies according to translation methods such as MT systems [1], statistical models [14], or *fast document translation* [3]. Compared to previous document translation methods, our naïve document translation is likely to produce the worst quality of document translation, since it do not use any separate resolution scheme. So, an evaluation for our hybridization of query and document translation in its naïve translation could provide a baseline performance to any combination approaches of query and document translation.

# 5 Experimental Results of Cross-Language Retrieval

This section describes the experimental evaluations for a combination approach of QT and DT, using NTCIR-4 KC and KJ CLIR test sets. Details about KC and KJ CLIR test sets can be found in the NTCIR-4 overview paper [10].

Table 7 shows the retrieval results for KC and KJ CLIR. QT and DT means the naïve QT and the naïve DT, respectively. The bold face, underline, the symbol '*', or '**' have the same meanings as those of Table 4.

**Table 7. KC and KJ CLIR Retrieval Results**

| | | T | D | C | DN | TDNC |
|---|---|---|---|---|---|---|
| KJ | QT | 0.2861 | 0.3039 | 0.3000 | 0.3763 | 0.3905 |
| | DT | 0.3165 | 0.3207 | 0.3140 | 0.3909 | 0.4039 |
| | QT+DT | 0.3234** | 0.3362** | 0.3241 | 0.4098* | 0.4229* |
| | QT+DT (feedback) | **0.3602** | **0.3601** | **0.3713*** | **0.4471** | **0.4473** |
| KC | QT | 0.1436 | 0.1456 | 0.1584 | 0.1665 | 0.1778 |
| | DT | 0.1551 | 0.1448 | 0.1567 | 0.1937 | 0.2057 |
| | QT+DT | 0.1687** | 0.1731** | 0.1763** | 0.1992** | 0.2089** |
| | QT+DT (feedback) | **0.1892** | **0.1869** | **0.2028*** | **0.2378**** | **0.2469*** |

Comparing QT with DT in both KC and KJ CLIR, DT was on the average better than QT, although the difference between QT and DT was not statistically significant at the 5% error level. As its first reason, we believe that, in KC and KJ language pairs, QT generates more ambiguous terms than DT, since KC and KJ dictionaries for QT are more ambiguous than those of CK and JK for DT, respectively, in terms of dictionary ambiguity (see Table 6).

Another reason is the impact of query structuring. A naïve QT creates an unstructured target language query without any normalization such as sum-to-one normalization [9]. That is, each target language query term equally contributes, irrespective of the number of translations for each source language query term. On the other hand, a naïve DT performs some structuring on queries based on term frequencies (tf) and document frequencies (df) of target language translations. This structuring effect of DT can be understood by interpreting a naïve DT as the Pirkola method. That is, when the Pirkola method creates a pseudo term from a set of translations for each source language query term, a pseudo term is realized with two pseudo sources of evidence: a pseudo tf, and a pseudo df. A pseudo tf is calculated by summing term frequencies of translations that are matched with document terms. A pseudo df is the size of the union of sets of document postings of translations that are matched with document terms.

In Table 7, a hybrid of QT and DT outperformed QT or DT alone in both KC and KJ CLIR. The hybrid system, QT+DT (no feedback), showed a statistical difference against QT at 5% significance level, except for the case of concept (C) queries in KJ-CLIR. We believe that this is because that QT and DT has different disambiguation effects on queries. Figure 3 shows the evidence. In Figure 3, two graphs plot the difference between QT and DT in MAP values for KC and KJ CLIR, respectively. For each topic, 5 symbols (circle, x-mark, triangle, diamond, and square) indicate different query types such as T, D, C, DN, and TDNC. In Figure 3, QT-oriented queries, queries for which QT is better than DT, are clearly distinguished from DT-oriented ones.

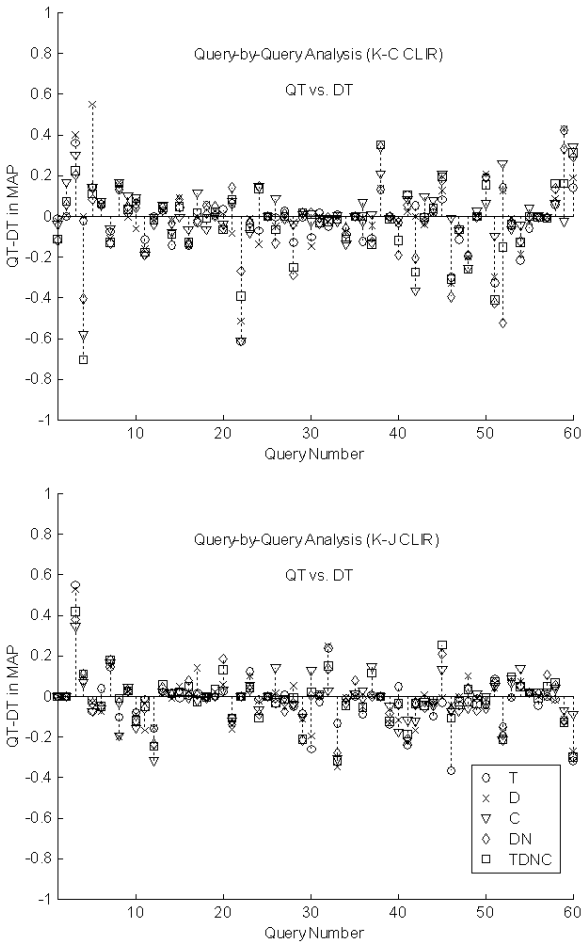


**Figure 3. Comparison of QT and DT in MAP difference**

## 6 Conclusions

For CJK monolingual information retrieval, we employed a word/n-gram coupling strategy that combines several ranked lists generated from words and n-grams indexes by differentiating both retrieval models and expansion term selection schemes. From the experiments on CJK languages, a fusion of top three combination models succeeded only on Korean MLIR, casting a research question that probably there exists a language-dependent set of top combination models for merging of them.

For cross-language information retrieval, a dictionary-based bi-directional query translation and document translation were combined at its naïve translation mode. Experimental evaluations showed that query translation and document translation differ in retrieving relevant documents on a particular query. In addition, query translation and document translation collaboratively helped each other to improve CLIR retrieval effectiveness, even at its default translation.

## Acknowledgements

This work was supported by the KOSEF through

the Advanced Information Technology Research Center (AITrc) and by the BK21 project.

## References


[1] BRASCHLER, M. Combination approaches for multilingual text retrieval. *Information Retrieval* 7, 183-204, 2004.

[2] BROWN, P., PIETRA, S.D., PIETRA, V.D., AND MERCER, R. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19, 2, 263-311, 1993.

[3] CHEN, A., AND GEY, F.C. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval* 7, 149-182, 2004.

[4] CHEN, K.U., CHEN, H.H., KANDO, N., KURIYAMA, K., LEE, S., MYAENG, S.H., KISHIDA, K., EGUCHI, K., AND KIM, H. Overview of CLIR task at the third NTCIR workshop. In *Working Notes of the 3rd NTCIR Workshop Meeting*, Tokyo, Japan, Oct. 2002, 1-38.

[5] DARWISH, K., AND OARD, D.W. CLIR experiments at Maryland for TREC-2002: evidence combination for Arabic-English retrieval. In *Proceedings of the TREC-11 Conference*, Gaithersburg, MD, Nov. 2002.

[6] FUJII, A., AND ISHIKAWA, T. Evaluating multi-lingual information retrieval and clustering at ULIS. In *the 2nd NTCIR Workshop*, Tokyo, Japan, Mar. 2001.

[7] GEY, F.C. Chinese and Korean topic search of Japanese news collections. In *Working Notes of the 4th NTCIR Workshop Meeting*, Tokyo, Japan, Jun. 2004, 214-218, 2004.

[8] HE, H., GAO, J., HE, P., AND HUANG, C. Finding the better indexing units for Chinese information retrieval. In *Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*, Taipei, Taiwan, Sep. 2002, 11-17.

[9] HIEMSTRA, D., AND DEJONG, F. Disambiguation strategies for cross-language information retrieval. *Lecture Notes in Computer Science* 1696, Springer-Verlag, 274-293, 1999.

[10] KISHIDA, K., CHEN, K.H., LEE, S., KURIYAMA, K., KANDO, N., CHEN, H.H., MYAENG, S.H., AND EGUCHI, K. Overview of CLIR task at the fourth NTCIR workshop. In *Working Notes of the 4th NTCIR Workshop Meeting*, Tokyo, Japan, Jun. 2004, 1-59.

[11] KWOK, K.L. Employing multiple representations for Chinese information retrieval. *Journal of the American Society for Information Science* 50, 8, 709-723, 1999.

[12] LEONG, M.K., AND ZHOU, H. Preliminary qualitative analysis of segmented vs bigram indexing in Chinese. In *Proceedings of the TREC-6 Conference*, Gaithersburg, MD, USA, Nov. 1997, 19-21.

[13] LUK, R.W.P., AND KWOK, K.L. A comparison of Chinese document indexing strategies and retrieval models. *ACM Transactions on Asian Language Information Processing* 1, 3, 225-268, 2002.

[14] MCCARLEY, J.S. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA, Jun. 1999, 208-214.

[15] NIE, J.Y., AND REN, F. Chinese information retrieval: using characters or words? *Information Processing and Management* 35, 4, 443-462, 1999.

[16] NIE, J.Y., GAO, J., Zhang, J., AND ZHOU, M. On the use of words and n-grams for Chinese information retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, Sep. 2000, 141-148.

[17] OARD, D.W. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, AMTA*, Penn., USA, Oct. 1998, 472-483.

[18] OARD, D.W., LEVOW, G.A., AND CABEZAS, C.I. CLEF experiments at Maryland: statistical stemming and backoff translation. In *Working Notes of the Cross Language Evaluation Forum Workshop (CLEF-2000)*, Lisbon, Sep. 2000.

[19] OARD, D.W., AND WANG, J. NTCIR-2 ECIR experiments at Maryland: comparing structured queries and balanced translation. In *the 2nd NTCIR Workshop*, Tokyo, Japan, Mar. 2001.

[20] OARD, D.W., AND ERTUNC, F. Translation-based indexing for cross-language retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, Glasgow, Mar. 2002, 324-333.

[21] OGAWA, Y., AND MATSUDA, T. Overlapping statistical segmentation for effective indexing of Japanese text. *Information Processing and Management*, 35, 4, 463-480, 1999.

[22] PIRKOLA, A. The Effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, Aug. 1998, 55-63.

[23] PONTE, J. A language modeling Approach to information retrieval. Ph.D. *Thesis*. University of Massachusetts at Amherst. 1998.

[24] RIJSBERGEN, C.J. van *Information Retrieval*, 2nd edition, Butterworths, 1979.

[25] ROBERTSON, S.E. On term selection for query expansion. *Journal of Documentation* 46, 359-364, 1990.

[26] SINGHAL, A., SALTON, G., MITRA, M., and BUCKLEY, C. Document length normalization. *Information Processing and Management* 32, 619-633, 1996.

[27] TSANG, T.F., LUK, R.W.P., AND WONG, K.F. Hybrid term indexing using words and bigrams. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, Nov. 1999, 112-117.

[28] ZHAI, C., LAFFERTY, J. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, Georgia, USA, Nov. 2001.