# Simple Query Translation Methods for Korean-English and Korean-Chinese CLIR in NTCIR Experiments

Myung Gil Jang*, Pyung Kim, Yun Jin, Suk-Hyun Cho, Sung Hyon Myaeng
Dept. of Computer Science, Chungnam National University, Korea
* ETRI, Korea

mgjang@etri.re.kr, {pyung, wkim, shcho, shmyaeng}@cs.cnu.ac.kr

## Abstract

*The main goal of our participation in the NTCIR Workshop is to evaluate relatively simple yet practical methods for CLIR using Korean queries for English and Chinese documents. We employed dictionary-based query translation methods for both cases but with different translation ambiguity resolution techniques. The Korean-English CLIR was quite successful, but the Korean-Chinese CLIR resulted in unexpectedly low performance. While our analysis is still in progress, we found several problems related to the bilingual dictionary used for this experiments and identified issues to be considered in the future.*

## Keywords:

Korean-English CLIR, Korean-Chinese CLIR, translation ambiguity resolution, dictionary-based query translation

## 1. Introduction

The main goal of our participation in the NTCIR Workshop is to evaluate relatively simple yet practical methods for CLIR using Korean queries for English and Chinese documents. The results would help us understand the base lines in the tasks, i.e., Korean-English (K-E) & Korean-Chinese (K-C) CLIR, which could be used for further research into more sophisticated methods.

Among several approaches to CLIR, we opted for the method of translating queries based on a bilingual dictionary. While a simple dictionary based query translation method with a thesaurus-based disambiguation technique was employed for the K-C CLIR task, a carefully designed method of using mutual information from the target corpus was used for the Korean-English CLIR task.

In order to measure and compare effectiveness of the aforementioned methods, we submitted runs for English-English and Chinese-Chinese mono-lingual retrieval and those for manually translated queries for K-E and K-C, in addition to the K-E & K-C CLIR runs.

## 2. Korean-English CLIR

### 2.1 Query translation

In essence, our query translation is based on a Korean-English bilingual dictionary and target corpus statistics. We used a bilingual dictionary with about 190,000 entries developed in ETRI for machine translation to translate a Korean query into a list of terms in English. Since the Korean terms in the original query often have many senses, some of the translated terms are not related to the meaning of the original query, causing the translation ambiguity problem.

As in Fig. 1, a Korean query is analyzed to select content-bearing terms in our system. English terms generated from the direct translation of the Korean terms then go through the disambiguation module. Unrelated terms are eliminated and the surviving terms are weighted, based on some calculated likelihood of each term representing the meaning of the original query terms. The term selection and weighting process is guided by the mutual information values

calculated for all word pairs from the target corpus [1].

## 2.2 Indexing and retrieval

English documents are indexed with standard methods: stop word elimination and stemming. Translated English query terms are processed in the same way.

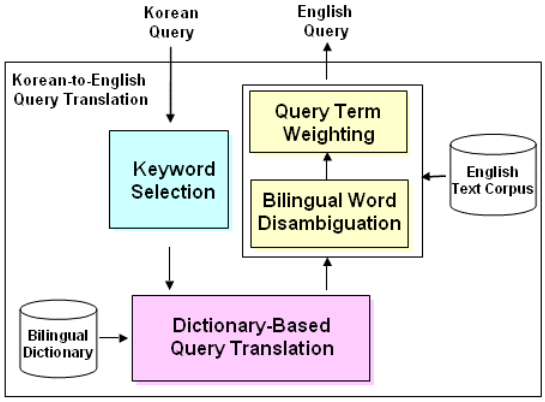


Fig. 1. The query translation process

For retrieval, our system adopted the 2-Poisson model [2]:

$$W = \sum_{i=1}^{n} (w_{di} \times w_{qi})$$

$$w_{di} = \frac{tf_{di}}{k_1\left((1-b)+b\cdot\frac{\text{document length}}{\text{average document length}}\right)+tf_{di}} \times \log\frac{N-n+0.5}{n+0.5}$$

$$w_{qi} = tf_{qi}$$

In our runs, $k_1$ was set to 2.0 and $b$ to 0.75 as suggested by [3].

## 2.3 Result Analysis

Among the K-E CLIR runs we submitted with the group ID CECIR, the run #7 gave the best result (11-point average precision value being 0.3169). In order to put the value in perspectives, we obtained a value for the case of E-E mono-lingual retrieval result (0.4383) as in Table 1. This value is not much different from the E-E retrieval average using TDNC across the runs submitted for this workshop (0.4557).

Table 1. K-E CLIR Comparison

| Run Type | E-E TDNC Avg. | E-E CECIR | K-E CECIR TDNC* |
|---|---|---|---|
| 11-pt. Avg. Pr. | 0.4557 | 0.4383 | 0.3169 |

* Run #7

The effectiveness of the CLIR is about 72.3% of that of the mono-lingual retrieval case in this experiment. Considering the fact that a significant number of content-bearing words could not be translated at all, due to the limitation of the bilingual dictionary, the ratio is quite remarkable. While this ratio is larger than most of other figures reported in the literature for simple dictionary-based query translation methods, it is smaller than the ratio 85% we obtained in the previous experiments [Jang et al., 1999].

In the following sub-sections, we report on our detailed analyses of individual queries.

### 2.3.1 Successes

In the case of Topic 2, the K-E result is slightly better than the E-E result. The content-bearing part of the topic #2 is:

*<TITLE>Joining WTO </TITLE>*
*<DESC>*
*Find possible problems that industries will meet after Taiwan's joining WTO.*
*</DESC>*
*<NARR>*
*It has taken Taiwan 10 years to get in to WTO. The Council For Economic Planning and Development, Chung-Hua Institution for Economic Research and Taiwan Institution for Economic Research evaluated the beneficial result of joining WTO. Related contents are supposed to include the evaluation contents, the advantages and disadvantages and the effects on agriculture, industry and business. If the documents only describe the opinions, comments, and attitudes of the America and other countries, or the political and diplomatic issues, they will be regarded irrelevant.*
*</NARR>*
*<CONC>*
*Taiwan, WTO, agriculture, industry, benefits, economy, World Trade Organization*
*</CONC>*
</TOPIC>

In the translated query from the Korean version, most of the important content-bearing words were correctly translated. Since only a few words were not translated at all (e.g. *Chung-Hua, joining, political, diplomatic, beneficial*) or incorrectly translated as in Table 2, they didn't seem to affect the retrieval effectiveness.

Table 2. Examples for incorrectly translated words

| Words in the English Topic | Incorrect translations |
|---|---|
| council | committee |
| issue | dispute |
| benefits | advantage |

In the case of the run #7, we also observed that most of the query words, except *Taiwan*, had similar weights, resulting in little improvement in effectiveness. An interesting case is that although the word *WTO* is highly important in this query, the weight was distributed among the *WTO* and the three words in *world trade organization*, giving an unexpectedly low weight on *WTO* that actually appears in documents as an important term. If the weighting scheme had worked, the result could have been better.

### 2.3.2 Failures

We analyzed the Topic 13 because the K-E CLIR result is considerably lower than the mono-lingual result. The topic is:

*<TITLE>Province-refining</TITLE>*
*<DESC>*
*Find the content of Province-refining enactment and Mr. James Soong's attitudes after the Province-refining*
*</DESC>*
*<NARR>*
*Taiwan Province and Taiwan province assembly have become history since December 20th, 1998. The temporary province function and organization regulation start applying on December 21st. Related contents include province-refining regulation and on what it is based, which regulations stop applying, which ones start applying, the purpose of province-refining, James Soong's reflections, attitudes and other related comments. Effects on individual because of province-refining will be regarded as irrelevant.*
*</NARR>*
*<CONC>*
*Province-refining, James Soong, Taiwan Province, chairman of Taiwan Province, Province assembly, budget*
*</CONC>*

There were considerable number of words that could not be translated (e.g. province, enactment, Mr. James Soong, chairman, Province assembly) and incorrectly translated words as in Table 3.

Table 3. Incorrectly translated words for Topic 13.

| Original words | Translated words |
|---|---|
| refining | change |
| reflection | opinion |
| Taiwan | Taiwan<br>Formosa |

While the dictionary has a limited coverage of general words, its lack of proper nouns (e.g. Soong) had a big impact on the poor result since relevant documents tend to have many occurrences of the proper nouns.

In the run #7, heavy weights were only given to *change, Formosa, regulation, Taiwan*, but *change* and *Formosa* are not correct translations as shown in Table 3. As a result, run #6 where no weighting was given to the query terms outperformed run #7 that incorporated the weighting scheme. In other words, absence of some important query terms plus assignment of high weights to incorrectly translated terms made the retrieval precision very low.

### 2.3.3 Discussions

Based on the failure analyses, we came to a conclusion that in the dictionary-based query translation, the coverage of the dictionary is most critical in improving the retrieval effectiveness, especially when there is a reasonable way of resolving translation ambiguities. In particular, proper nouns such as person names and geographical location names seem to play an important role. It may be feasible to combine a corpus-based method when a word does not appear in the dictionary.

In the case of run #7, in addition, one-to-many translations have a positive impact of increasing recall especially when the sense disambiguation is not very clear. However, the method also allows for incorrect words to be used for retrieval, degrading retrieval effectiveness. As in the case of the Topic 2, in order to avoid the problem of

distributing a weight to several translations, only one of which actually appears in relevant documents, there should be a way of gathering weights.

## 3. Korean-Chinese Cross Language IR

### 3.1 Query translation

Translation of Korean queries into Chinese queries was conducted using a Korean-Chinese bilingual dictionary. It contains about 150,000 entries and is still under development in ETRI for Korean-Chinese machine translation tasks.

For ambiguity resolution in query translation, we used a hierarchical structure of semantic codes embedded for nouns. In this structure, nouns are divided into concrete nouns, abstract nouns, and activity nouns, each of which has its own hierarchy of depth four.

Our ambiguity resolution method basically relies on the semantic distance between words in the hierarchy. Given all the translations for each of the Korean query terms, their mutual semantic distances are calculated. By adding the distance values for each translation, we obtain its semantic distance to other query terms.

We observed several problems in the query translation process. Like the K-E CLIR, the most serious problem arises from named entities such as geographical location names and person names that are often misrecognized. 6.3% of Chinese query terms belongs to these categories, and the ratio increases to 9.3% when foreign words are counted together.

For instance, the word '漢代' (Han Dynasty period) in the original topic statement in Topic 1 was expressed as a phrase '한 왕조' in Korean. Since this phrase can be interpreted as ‘one dynasty’ rather a particular dynasty with a proper noun, it was reduced to ‘왕조’ by the morphological analyzer and subsequently translated into '王朝' (dynasty). As a result, the unique meaning of the original query word, i.e. the proper noun, was lost in the translation process.

Person names were extremely difficult to translate correctly. For instance, '崔琦', a person name, in Topic 6 is expressed as '추 이' in Korean whereas '宋楚瑜' is expressed as '쏭추위'. Although the last two syllables of the latter is the same as the two syllables of the former, their origins are completely different. Furthermore, the former is written with a space between the two syllables, a writing convention in Korea.

More complicated problems arise when three languages are involved. In Topic 17, for example, a person name '다케시 기타노' (Takeshi Gitano) was correctly translated from '北野武' the original query created in Japanese (i.e. SLANG is Japanese). However, it is difficult for Korean language processors to translate the six syllable person name into the three Kanji characters since the Korean word is an expression of a word based on the sound rather than the meaning. Furthermore, it is not clear how the Japanese person name would be expressed in Chinese documents. This kind of problems occurs with terminologies, too.

Based on our analysis, some Korean topic statements contain improperly translated terms, especially person and place names, and these terms influenced the retrieval results seriously. Furthermore, our query translation method based on the incomplete dictionary itself caused many problems since it has very limited coverage of proper nouns, terminology, and foreign words.

### 3.2 Indexing and retrieval

Chinese documents were processed in a very simple manner; once sentences were recognized, they were segmented by space, numbers, English words, and punctuation marks. Without any other segmentation, bi-grams were used as index terms. The same method was applied to the processing of topic statements. Like the K-E CLIR, we used our retrieval system that ranks documents based on the 2-Posson model.

### 3.3 Result Analysis

The aforementioned scheme was tested using 42 topic statements. As in the case of K-E CLIR, we submitted C-C monolingual retrieval runs for comparisons. The 11-point average precision (0.3329) for our mono-lingual retrieval run, i.e. C-C run, is slightly higher than that of the average C-C TDNC runs submitted for this workshop (0.2967).

However, our K-C CLIR results are considerably lower than the average C-C TDNC result. This is primarily due to the poor coverage of the Korean-Chinese bilingual dictionary we used. In order to understand the relationship between the translation coverage (i.e. to what

extent the source query terms were found in the bilingual dictionary) and retrieval effectiveness, we calculated the ratio between the 11-point average precision of the K-C runs and that of the C-C runs for various translation coverage values, as in Fig. 2. The average of translation coverage from all the topics is 0.15. In order to eliminate the bias from the inherent difficulty of some queries, which resulted in poor effectiveness values even in C-C retrieval, we used the ratio between K-C and C-C.

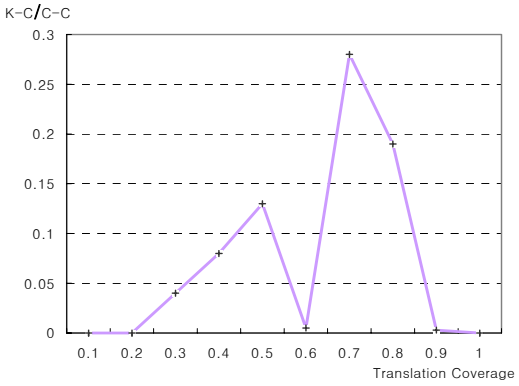


Fig. 2. The effects of translation coverage

Overall, the graph shows a general trend that as the translation coverage becomes higher, the K-C/C-C ratio increases. The drop at the translation coverage value 0.6 was due to the unusual characteristics of the topic 9 and topic 17 whose translation ratio happened to be about 0.6. Their lengths are very short with many proper nouns whose translations are totally incorrect. The drops beyond 0.7 are explained in the same way.

We give qualitative analysis results for individual topic statements in the following.

### 3.3.1 Successes

Topic 18 is one of the rare successful queries. This topic is unique in that it contains no person/place names and foreign words either in the Korean or Chinese queries. In addition, it generates the largest number of Chinese translations from the Korean query. That is, almost all the words in the Korean query have Chinese translations found in the bilingual dictionary. Finally, the resulting Chinese translations are relatively unambiguous words with clear meaning like *religion* and *crime*. This confirms the hypothesis that proper nouns and the coverage of the bilingual dictionaries are the major contributing factors for low effectiveness in K-C CLIR.

### 3.3.2 Failures

Aside from the low performance of the C-C retrieval, which may be attributed to the indexing method, the low performance of the K-C CLIR seems to come from the inadequacy of the Korean-Chinese bilingual dictionary.

The number of words extracted from the Korean topic statements (39 words on the average) is smaller than the number of Chinese words (57 words on the average) from the original topic statements. The difference may indicate that the Korean topic statements are not as expressive as the original Chinese topic statements, and it partially explains the low performance. However, we believe that the difference in numbers is primarily due to the nature of Chinese and Korean languages and the characteristics of the Korean morphological analyzer.

A more important reason is the fact that only 75.4% of Korean query words, i.e., 29.7 words on average, have at least one translation in the bilingual dictionary. Among the translations, 22.8% (6.79 per topic on average) belong to foreign words, inappropriate expressions, or the simplified form that is not used in Taiwan.

We compared manual translation of Korean query words and the dictionary-based translations. Only 7.5 words per query were common between the manually generated queries and automatically generated queries. This indicates how incomplete the bilingual dictionary was.

In the following, we summarize the problems and issues that need to be addressed when Korean-Chinese bilingual dictionaries are used for query translations.

- Lack of terminology
  '전자상거래' (electronic commerce) in Topic 4, and '인공위성' (satellite) in Topic 9 are examples for non-existent words in the dictionary.

- Lack of proper nouns
  '세계무역기구' (WTO) in Topic 8 and '타이타닉' (Titanic) are examples that do not appear in the dictionary.

- Incorrect or unnatural expressions
  Some translations are inadequate for native

Chinese speakers, due to differences in colloquialism, word order differences, and other conventions unfamiliar to foreigners. Despite the differences, bilingual dictionaries tend to contain expressions used in Korea since many Korean words were originated from Chinese. Table 4 shows some examples.

Table 4. Differences in expressions

| Incorrect | Correct |
|---|---|
| '生長率' | '成長率' |
| '亞細亞' | '亞洲' |
| '亞非利加' | '非洲' |
| '學生父母' | '家長' |

- Differences between the Simplified Form and the Original Form
  The Simplified From is used in China whereas the Original Form is used in Taiwan. Even for the Original Form, some words used in the mainland and Taiwan are different. For instance, '臺灣' (Taiwan) is used in the mainland whereas '台灣' is used in Taiwan.

- Mapping between a word and a phrase
  When a two-word phrase can be expressed in a single word in Chinese, for example, the automatic translation process may end up producing two Chinese words whose combination is totally different from the Korean phrase. For instance, '머리 스타일' (hair style) must be translated into a single word 髮型, but it may be translated into two separate words corresponding to the two Korean words. Another example is 'TV 방송국' (TV station) correctly translated into a single word '電視台'.

- Differences in expressing numbers
  Numbers such as years are often expressed in digits in Korea (e.g. 1998), but are expressed with Chinese characters in Chinese documents (e.g. 一九九八)

- Word order differences
  Although some Korean words or at least each character consisting of a word have been originated from Chinese, the order of component word or characters for a compound word may differ from each other. For instance, '核反對運動' (Anti-nuclear movement) is expressed as '反核運動' in Chinese with the order of the first two characters reversed. This kind of differences is not always reflected in bilingual dictionaries.

- Lack of richness
  In comparison with the variety of expressions in Chinese documents, the bilingual dictionary contains limited translations for the entries. For example, '수상' has only two translations '首相' and '水上', there are many other possibilities: '得?/愁傷/好奇/受傷/授?/首相/數上/水上/受賞/大臣 (http://www.openchina.co.kr).

### 3.3.3 Discussions

While we are still in the process of analyzing the low performance compared to the NTCIR average, we conjecture that the bi-gram method we employed had some problems. In this paper, however, we limit our discussion to the issues related to query translations, rather than the indexing and retrieval issues.

Although the query translation method is simpler than other techniques for CLIR, the performance depends heavily on the quality of the dictionary used for translations.

## 4. Conclusion

Our analysis of individual translations of query words and the overall performances show that it is very important that bilingual dictionaries be made with caution and sufficient knowledge on both languages. In particular knowledge on contemporary language usages, terminology, and foreign words seem most critical.

Ambiguity resolutions and term weight assignments help improving retrieval effectiveness. However, the amount of gain from ambiguity resolution seems to be surpassed by the loss from insufficient coverage of bilingual dictionaries.

Since languages evolve, dictionaries to be used for CLIR must be upgraded periodically, or supplemented with such things as corpus statistics reflecting contemporary writings.

## References

[1] Myung-Gil Jang, Sung Hyon Myaeng and Se

Young Park, 1999, "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, MD.

[2] Robertson, S.E. and Walker S., “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval,” In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232-241, 1994.

[3] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M., “Okapi at TREC-3,” In Proceedings of the 3rd Text Retrieval Conference, pp. 109-126, 1995.