Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

KUNLP System for NTCIR-4 Korean-English Cross-Language Information Retrieval

Hee-Cheol Seo, Sang-Bum Kim, Ho-Gun Lim and Hae-Chang Rim
Dept. of Computer Science and Engineering, Korea University
1, 5-ka, Anam-dong Seongbuk-Gu, Seoul, 136-701, Korea
{hcseo, sbkim, hglim, rim}@nlp.korea.ac.kr

Abstract

This paper describes our Korean-English cross-language information retrieval system for NTCIR-4. Our system is based on a query translation approach with a bilingual dictionary and co-occurrence information between English terms in English corpus. In this year, we have focused on translation of unknown words. We have expanded the existing bilingual dictionary by gathering some of the Korean-English translation pairs for Korean words from Web manually. For other unknown not contained in the expanded bilingual dictionary, we automatically transliterated into English using pre-constructed mapping table. Some issues for processing Korean queries and documents are also described, such as identification of Korean phrases. On evaluation collections for NTCIR-4, performance of our system is 30.25% for description query type, 33.33% for title query type, and 32.47% for combination query type of description and narrative in relax scoring. Post-submission experiments show that our expanded dictionary and transliteration mechanism improve the performance of our system.
Keywords: *Korean-English cross-language information retrieval, query translation, transliteration, Korean-English Dictionary*

1 Introduction

We participated in Korean-English cross-language information retrieval(CLIR) task of NTCIR-4 workshop. We adopted a query-translation approach using a bilingual dictionary, because the approach is known to be very simple and effective. Among several issues in the query translation approach, we focused on translation of unknown word, which is one of the major reasons to drop down performance of CLIR systems. In order to reduce the problem of unknown words, we expanded our bilingual dictionary by collecting translation information of unknown words from Web, and automatically transliterated the unknown words that are not still registered in our dictionary into English. In Korean-English CLIR, to handle Korean queries become another issue. We describe a mechanism to extract Korean query terms from Korean queries by considering characteristics of Korean.

This paper is organized as follows: Next section describes our query translation method including bilingual dictionary, transliteration method for unknown words, query term extraction and translation equivalent selection. Section 3 presents a document retrieval method and a query expansion method. In Section 4, we show our official results in NTCIR-4 and analyze failures of our system. In Section 5, post-submission experiments are presented for the effectiveness of our dictionary and our transliteration method. Finally, we conclude and describe our future works in Section 6.

2 Query Translation

2.1 Bilingual Dictionary Expansion

CLIR system only based on a bilingual dictionary requires the perfect dictionary, which contains all query terms in a query and their translation equivalents. However, it is impossible to obtain such dictionary since a huge number of words newly spring up day by day in the real world.

For this reason, we cannot avoid unknown word problem in the CLIR system based on the bilingual dictionary. The unknown words in a query cannot be translated properly by the system. It makes the performance of CLIR system get lower than that of single language information retrieval (SLIR) systems.

In order to reduce the problem of unknown words, we collected translation information of unknown words from Web manually, and added them to our general-purpose dictionary. Thus our dictionary contains 34,122 new entries, whose types are Korean place name, company name, country name, person name, pet name, street name, person name in Bible, place name in Bible, medical terminology, vegetable name and so on. Table 1 presents some examples.

Table 1. Examples of Additional Entries

word type	examples
Korean place	대전-Taejon, 종로-Jongro
company	삼성-Samsung, 야후-Yahoo
country	감비아-Gambia
person	라이언-Ryan, 모건-Morgan
pet	라빈-raven, 로빈-robin
place in Bible	예루살렘-Jerusalem
person in Bible	야곱-Jacob, 유다-Judas
medical science	골수염-Osteomyelitis
vegetable	양상치-lettuce, 파슬리-Parsley

Table 2. Example : Transliteration of Korean Unknown Word

Kor. word	아키타
mapping table	아 - A, Ah 키 - Key, Ci, Ki, Ky, Cy 타 - Ta
Transliteration	AKeyTa, ACiTa, AKiTa, AKyTa, ACyTa AhKeyTa, AhCiTa, AhKiTa, AhKyTa, AhCyTa

Entries in our dictionary have pairs of Korean terms and their English translations. Korean terms are ones of Korean morphemes, eojeols, and phrases. In Korean, a morpheme is a basic unit with meaning, an eojeol is composed of several morphemes, and a phrase is composed of several eojeols or morphemes.

2.2 Unknown Word Transliteration

There are still unknown words for our expanded dictionary. Hence, some techniques are required to translate the unknown words. A well-known technique is to transliterate the unknown words into document language, in Korean-English CLIR, into English. We also adopted the transliteration approach, and developed one mechanism.

Before describing our transliteration mechanism, we must explain characteristics of Korean on the transliteration. Korean language is a phonetic language, and one syllable represents one sound. For example, Korean word "서희철" consists of three syllables, "서", "희", and "철", and each syllable is pronounced "seo", "hee", and "cheol", respectively. Since we can map each syllable to its corresponding English string, all the unknown words also can be transliterated into English if we have such a mapping table.

For the automatic transliteration for Korean unknown words, we have used a large mapping table between Korean syllables and their English transliterations, which contains 2,350 Korean syllables[1] and their English transliterations. Each syllable has 1.83 English transliterations on average, since some syllables can be transliterated in several English strings.

With the mapping table, unknown words are transliterated by combining with every English transliterations of Korean syllables in the words. Table 2 shows one example for Korean unknown word "아키타", which consists of three syllables, "아", "키", and "타". Each syllable has 2, 5, and 1 English transliterations respectively, and 10 English transliterations are made for the words "아키타".

Unlike other transliteration techniques resulting in just one transliteration for one word, our technique generates all the candidate transliterations. The candidate transliterations for unknown words are regarded as the translation equivalents of the unknown words in the following query term translation phase, and the most probable transliteration is selected in the phase. For instance, 10 candidate transliterations in Table 2 are regarded as translation equivalents of "아키타" and are transferred to the query term translation phase.

2.3 Query Term Translation

In order to translate query terms in queries, our system firstly identifies the query terms from queries, secondly collects translation equivalents of query terms by looking up a bilingual dictionary or by transliterating the unknown word query terms, and finally translates the query terms into English by selecting the most probable translation equivalents.

As described in the section 2.1, entries of our bilingual dictionary contain Korean morphemes, eojeols, and phrases. Hence, the query terms must be one of three types in order to get translation equivalents of them.

Before extracting query terms, we must consider structures of Korean sentences(or queries). A Korean sentence is composed of several eojeols, which are divided by blanks, and a Korean query is the same because a query is a kind of a sentence. Eojeols are composed of one or more morphemes, but there are not blanks between morphemes in an eojeol. Korean phrases consist of several eojeols or morphemes. Table 3 shows a Korean example sentence, where POS tag N is noun, V is verb, P is particle, E is ending and S is symbol. The table shows that an eojeol "실험에" consists of two morphemes, "실험" and "에", and a phrase "핵 실험" consists of one eojeol "핵" and one morpheme "실험" of an eojeol "실험에".

Among a morpheme, an eojeol, and a phrase, a

[1] Among several coding systems for Korean, KSC5601-1987 Wansung code consists of these 2,350 syllables, which can represent most Korean words.

Table 3. Components of a Korean sentence

Sentence	지하 핵 실험에 대한 항의에 관한 문서를 찾는다.
Eojeol(Morpheme/POS)	지하(지하/N), 핵(핵/N), 실험에(실험/N, 에/P), 대한(대하/V, ㄴ/E), 항의에(항의/N, 에/P), 관한(관하/V, ㄴ/E), 문서를(문서/N, 를/P), 찾는다.(찾/V, 는다/E ./S)
Phrase	핵 실험

phrase consists of the largest number of morphemes and so has the least number of translation equivalents, while morpheme has the most number of translation equivalents. Hence, it is desirable to identify Korean phrases firstly, eojeols secondly and morphemes finally as query terms.

In order to identify Korean phrases at queries which exist in the dictionary, eojeols in queries must be segmented into morphemes since phrases can consist of several morphemes as well as eojeols. For the sake of eojeol segmentation, we adopted a Korean part-of-speech(POS) tagger[2], which segments eojeols into morphemes and assigns the morphemes to POSs. After eojeol segmentation, phrases are identified by phrase patterns in Table 4. Due to complexity, we assumed that phrases consist of maximum two eojeols. Priority in the table means the order of each pattern that is applied to phrase extraction. In other words, "eojoel + eojeol" pattern whose priority is 1 is firstly applied, and if the pattern succeeds to identify a phrase, other patterns are not applied. Otherwise, next pattern "morpheme+eojeol" is applied. The number of morpheme in each pattern decides the priority, and the first pattern "eojeol+eojeol" has the largest number of morphemes in usual. In the table, there is a Korean phrase "정보 검색", which is in the bilingual dictionary and means "information retrieval". The phrase is extracted by the pattern eojeol("정보") + eojeol("검색") from "지능형 정보 검색", and by the pattern eojeol("정보")+morpheme("검색") from "지능형 정보 검색을" and so on.

Korean eojeols are extracted easily because they are divided by blanks. Among eojeols in queries, the eojeols that exist in the bilingual dictionary but are not components of phrase query terms are selected as query terms. For example, from "지능형 정보 검색", an eojeol "지능형" is identified as a query terms, but "정보" and "검색" are not identified because they are components of a phrase query term "정보 검색".

Finally, morpheme query terms are identified. Among morphemes in queries, the morphemes whose POSs are nouns are selected. Unlike phrase query terms and eojeol query terms, morpheme query terms do not have to be headwords in the dictionary since unknown morphemes are transliterated.

When query terms are identified, our system collects translation equivalents of the query terms and selects just one translation equivalent per each query term. To do that, we adopted mutual information between translation equivalents, and chose one translation equivalent with the highest score per each query term. The score of a translation equivalent is calculated as:

$$score(te_{ij}|Q) = \sum_{x=1,x\neq i}^{n} \sum_{y=1}^{Z(qt_x)} MI(te_{ij};te_{xy}) = \sum_{x=1,x\neq i}^{n} \sum_{y=1}^{Z(qt_x)} \frac{Pr(te_{ij},te_{xy})}{Pr(te_{ij})Pr(te_{xy})}$$

where $score(te_{ij}|Q)$ is the score of the j-th translation equivalent of the i-th query term given a query Q, n is the number of query terms in the query Q, $Z(qt_x)$ is the number of translation equivalents of the x-th query term qt_x, $MI(te_{ij};te_{xy})$ is the mutual information value between te_{ij} and te_{xy}[2], $Pr(te_{ij},te_{xy})$ is the probability that te_{ij} and te_{xy} cooccur in the same sentences, and $Pr(te_{ij})$ is the probability of te_{ij}. The values of probabilities are obtained from the English target documents. When more than two translation equivalents have the same highest score, our system randomly selected just one translation equivalent among them.

There are English words or numbers in some queries, which exist in the relevant documents usually. Hence, we add English words and numbers in source queries to translated queries.

3 Document Retrieval and Query Expansion

3.1 Document Retrieval

In order to retrieve English documents using English translated queries, we used Okapi BM25[4], which is one of the most successful techniques in an information retrieval. The values of parameters for BM25 are set to k1 = 20, b = 0.75, and k3 = 1000. All terms except stopwords in documents and in queries are used in BM25 after stemmed by a Porter stemmer.

[2]Mutual information value is calculated with two translation equivalents stemmed by a Porter stemmer.

Table 4. Patterns for Korean Phrase : an example phrase is "정보 검색"

pattern	priority	example for pattern
eojeol+eojeol	1	지능형 정보 검색
eojeol+morpheme	2	지능형 정보 검색을
morpheme+eojeol	3	지능형정보 검색
morpheme+morpheme	4	지능형정보검색을

3.2 Query Expansion

Query expansion is well known to improve the performance of CLIR system. There are two types of query expansion in CLIR: one is pre-translation query expansion and the other is post-translation query expansion. Pre-translation query expansion adds additional terms to a source query before a query translation phase, while post-translation query expansion adds additional terms to a translated query after a query translation phase. The additional terms by pre-translation query expansion are helpful to translate the source query into a document language correctly, and the additional terms by post-translation query expansion are helpful to retrieve the relevant documents. Some previous papers[1][5] showed that both types of query expansion are useful for CLIR.

For the query expansion, pseudo-relevance feedback measure of Okapi BM25 was adopted. The terms in top 10 documents retrieved by BM25 are sorted according to their relevance weight, the terms whose weights are larger than 0 and in top 10. The relevance weight of terms is calculated as[4]:

$$RW(t) = r_t log\frac{N}{n_t} - log\binom{R}{r_t} - logV \qquad (1)$$

where $RW(t)$ is the relevance weight of the term t, R is the total number of the relevant documents(In this paper, R is set to 10), r_t is the number of the relevant documents in which the term t occurs, N is the size of the collection, n_t is the number of the documents which contain the term t, and V is the size of the vocabulary. The argument of the second logarithm is the number of ways we can choose r_t from R, and is calculated as $\frac{R!}{r_t!(R-r_t)!}$.

For the pseudo-relevance feedback, document collections are required. For pre-translation query expansion, Chosunilbo and Hankookilbo collection are used as Korean collections, which are evaluation document collections for NTCIR-4. In Korean document retrieval, only nominal words are used as query terms or document terms, which Korean POS tagger or a noun-extractor[3] can identify. A noun extractor extracts only nominal words in a sentence, while a POS tagger assigns all words to a POS and identifies the nominal words with the POS information. In NTCIR-3, we had compared these two tools in English-Korean CLIR task and had observed that the contributions of two tools are similar. In NTCIR-4, we used a noun-extractor instead of POS tagger, since the noun-extractor is faster than the POS tagger.

For post-translation query expansion, English target document collections for Korean-English CLIR in NTCIR-4 are used. Document retrieval model is the same as the model in Section 3.1.

4 Official Results and Failure Analysis

We have submitted three CLIR runs:

- KUNLP-K-E-T-01(hereafter KET01) : a run using a Korean title field to retrieve English documents.
- KUNLP-K-E-D-02(hereafter KED02) : a run using a Korean description field to retrieve English documents
- KUNLP-K-E-DN-03(hereafter KEDN03) : a run using both Korean description and Korean narrative fields to retrieve English documents.

Table 5 presents our official results, and also the results of the single language information retrieval(SLIR) system for comparison purposes:

- EET : a run using a English title field to retrieve English documents.
- EED : a run using a English description field to retrieve English documents.
- EEDN : a run using both English description and English narrative fields to retrieve English documents.

The SLIR system is the same model as English document retrieval of CLIR system(Section 3.1), including the query expansion mechanism. We evaluated the SLIR system after submitting the official runs.

Table 5 shows that performance of our CLIR system is not as high as that of SLIR system. We have analyzed some queries with lower precision than SLIR, and found some reasons: problems of bilingual dictionary, problem of transliteration, and problem of POS tagging.

Table 6. Unknown words in queries

query ID	query topic	unknown Korean word
003	Embryonic Stem Cells	배아 (Embryonic)
011	Tiananmen Event, Global Signing Movement	천안문 (Tiananmen)
028	Self-Defense Force, Amendment, Law	자위대 (Self-Defense Force)
041	Cellular Phone, Internet, Service	핸드폰 (cellular phone)
057	Daily Life, Environmentally Friendly	친 환경적(Environmentally Friendly)

Table 5. Official Results and results of single language information retrieval

	relax		rigid	
run	avg. P.	R-prec.	avg. P.	R-prec.
KET01	0.3333	0.3507	0.2382	0.2574
EET	0.4151	0.4267	0.3306	0.3442
KED02	0.3025	0.3252	0.2250	0.2385
EED	0.3974	0.4185	0.3144	0.3252
KEDN03	0.3247	0.3492	0.2637	0.2869
EEDN	0.4151	0.4687	0.3306	0.3907

4.1 Problems of Bilingual Dictionary

Though our dictionary was expanded with additional information in order to reduce unknown words, there are still unknown words in queries. Since some unknown words represents key point of queries, and cannot be translated correctly by any transliteration mechanism, the unknown words bring about very serious problem in CLIR system. Table 6 shows queries with low precision due to the unknown words. In the table, for example, query 011 is related to "Tiananmen event" but our dictionary does not provide any translation information of "천안문", which is Korean translation equivalent of "Tiananmen".

In addition to the unknown words, there are some words that our bilingual dictionary does not contain proper translation equivalents. "해상" in query 019[3] means "incidents at sea", but the translation equivalents related to the meaning do not exist in our dictionary. Also "복제" in query 030[4] must be translated as "cloning", but the dictionary does not contain "cloning" as a translation equivalent of "복제".

4.2 Problem of Transliteration Method

We transliterated unknown words into English. However, our transliteration method suggested incorrect transliterations occasionally. Table 7 shows the queries with low precision due to the incorrect transliterations. Transliteration errors result from the lack of mapping information between Korean syllables and their English transliterations. For example, "던" in "조던" of query 006 does not have English transliteration "dan" and "스" in "스텔스" of query 051 does not contain English transliteration "th". If this mapping information is added to our mapping table, the transliteration errors will be reduced.

4.3 Problem of Korean Part-of-Speech Tagger

Korean part-of-speech(POS) tagger identifies morphemes in queries in our system. Errors of POS tagger have bad effects on the query term extraction, and that problem is propagated next phases. There are some queries with low precision due to errors of POS tagger.

"왕단" in query 011 is a proper noun in Korean, but POS tagger splits it into two morphemes "왕" and "단". The problem is that "왕단" represents different thing from "왕" and "단". That is to say, "왕단" indicates a person name, while "왕" means a king, and "단" means a platform. "구로사와" in query 012 indicates a name of a japanese person, "Kurosawa", but it is splitted into "구로", "사" and "와" by POS tagger. Hence out system extracts incorrect query terms, and thus translate the query incorrectly. "성추행" in query 045 is also splitted into "성추" and "행" by POS tagger. though it consists of two morphemes "성" and "추행". "성추" is a unknown word and "행" means "row", while "성" means "sex" and "추행" means "disgraceful conduct".

5 Post-submission Experiments

After submitting our runs, we have tried to find how much our expanded bilingual dictionary and our transliteration method improve the performance of CLIR system.

[3]Korean query is "국제적인, 사건, 해상" and it means "International incidents at Sea".

[4]Korean query is "동물, 복제기술" and it means "Animal Cloning Technique".

Table 7. Examples of incorrect transliteration

query ID	transliterated word	correct transliteration	incorrect transliteration
002	죠니워커	Jonnie Walker	JoNiWoeCeo, JoNiWoeKeo, ...
006	조던	Jordan	Jordon, Jodeon, ...
013	게이조	Keizo	Kericho, Keleecho, ...
051	스텔스	stealth	Stels, Seutels, ...

Table 8. Effectiveness of expanded bilingual dictionary on relax relevance

query type	Dict		Exp. Dict	
	avg. P.	R-prec.	avg. P.	R-prec.
title	0.2690	0.2944	0.3330	0.3507
desc.	0.2595	0.2879	0.3025	0.3252
d + n	0.3062	0.3316	0.3247	0.3492

Table 9. Effectiveness of Transliteration on Relax relevance

query type	w/o transliteration		with transliteration	
	avg. P.	R-prec.	avg. P.	R-prec.
title	0.3332	0.3518	0.3330	0.3507
desc.	0.2912	0.3116	0.3025	0.3252
d + n	0.3112	0.3370	0.3247	0.3492

5.1 Effectiveness of the expanded bilingual dictionary

In this experiment, we try to observe how much the additional information about the proper nouns and terminologies in a specific domain improve the performance of our CLIR system. For the comparison, we performed the experiments with and without the additional information. Table 8 shows the results, where ”Dict” means the general-purpose dictionary without the additional information, and ”Exp. Dict” means our expanded dictionary with the additional information. ”d+n” in query type indicates that both description field and narrative field are used as a query. In the table, it is observed that the additional information is useful, particularly for title query type. It seems that unknown words have fatal effects on document retrieval of short queries.

5.2 Effectiveness of Transliteration

In this experiment, we observed the contribution of the transliteration technique to CLIR. Table 9 shows the results, where ”w/o transliteration” represents the results without transliteration phase, and ”with transliteration” represents the results with transliteration phase. ”d+n” in query type indicates that both description field and narrative field are used as a query. The performance is improved slightly when using transliteration.

6 Conclusions and Future works

We described our system for Korean-English CLIR in NTCIR-4, which is the best system in the task, though only two teams participated in the task. Our system is based on a query translation approach, and uses an expanded bilingual dictionary and adopts transliteration mechanism in order to translate unknown words.

Our system extracts query terms considering the characteristics of Korean language, which is an agglutinative language, and where eojeols consist of several morphemes and are divided by blanks in sentences, and phrases consist of several morphemes or eojeols. In order to extract query terms, our system identifies morphemes by Korean part-of-speech tagger, eojeols by blanks, and phrase with some patterns.

We have evaluated our expanded bilingual dictionary and transliteration mechanism after submitting our official runs, and observed that the additional information in our dictionary is very useful, particularly in short queries and the transliteration mechanism is also useful though the rate of improvement is not high.

In the based of failure analyses, we will improve our system by using several bilingual dictionaries to reduce the problem of unknown words, and by expanding the mapping table for a transliteration to generate the correct transliteration.

References

[1] J. Gao, J.-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou, and C. Huang. Trec-9 clir experiments at msrcn. In *Proceedings of the Nineth Text REtrieval Conference(TREC-9), NIST special publication, 500-249*, 2000.

[2] J. D. Kim, H. S. Rim, and H. C. Rim. Twoply HMM: A part-of-speech tagging model based on morpheme-unit considering the characteristics of korean. *Journal of Korean Information Science Society*, 24(12(B)):1502–1512, 1987.

[3] D.-G. Lee, S.-Z. Lee, and H.-C. Rim. An efficient method for korean noun extraction using noun occurrence characteristics. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, 2001.

[4] S. E. Robertson and S. Walker. Okapi/keenbow at TREC-8. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference(TREC-8)*, pages 151–161, Gaithersburg, 1999. NIST special publication 500-246.

[5] J. Xu and R. Weischedel. Trec-9 cross-lingual retrieval at bbn. In *Proceedings of the Nineth Text REtrieval Conference(TREC-9), NIST special publication, 500-249*, 2000.