Korean-Chinese Cross-Language Information Retrieval Based on Extension of Dictionaries and Transliteration

Yu-Chun Wang[†‡], Richard Tzong-Han Tsai[†§], Hsu-Chun Yen[‡], Wen-Lian Hsu[†]

[†]Institute of Information Science, Academia Sinica, Taiwan
[‡]Department of Electrical Engineering, National Taiwan University, Taiwan
[§]Department of Computer Science and Engineering, Yuan Ze University, Taiwan

{albyu,thtsai}@iis.sinica.edu.tw
yen@cc.ee.ntu.edu.tw
hsu@iis.sinica.edu.tw

Abstract

This paper describes our Korean-Chinese cross-language information retrieval system. Our system uses a bi-lingual dictionary to perform query translation. We expand our bilingual dictionary by extracting words and their translations from the Wikipedia site, an online encyclopedia. To resolve the problem of translating Western people's names into Chinese, we propose a transliteration mapping method. We translate queries form Korean query to Chinese by using a co-occurrence method. When evaluating on the NTCIR-6 test set, the performance of our system achieves a mean average precision (MAP) of 0.1392 (relax score) for title query type and 0.1274 (relax score) for description query type.

摘要

本文描述我們所提出之韓中雙語跨語言檢索系統。我們採用韓中雙語辭典進行問題之翻譯，並利用線上維基百科以及韓國 Naver 網站來擴增我們雙語辭典的覆蓋率。此外，針對韓文中西方人名的翻譯，我們提出一音譯對應的搜尋方法。對於韓中翻譯時的歧義性問題，我們採用 Mutual Information 方法來解決。我們使用 NTCIR-6 之 test set 測試我們韓中跨語言檢索系統之效率，其使用標題部分進行查詢時之 Mean average precision (MAP) 之結果爲 0.1392；使用敍述部分進行查詢時之 MAP 爲 0.1274。

Keywords: Korean-Chinese cross-language information retrieval, query translation
關鍵詞：韓中跨語言資訊檢索，問題翻譯

1 Instroduction

The contents of whole Internet are growing explosively due to the improvement of the computer and web technology. Besides English, the web pages written in other languages also increase tremendously. In order to get the useful information from the Internet, many advanced modern search engines are developed, like Google[1], Yahoo[2], AltaVista[3], and so on. However, for the users that do not have any knowledge about other languages, it is impossible to get the information in other languages by current single-language web search engines.

Therefore, the research of cross language information retrieval (CLIR) is rising quickly. Cross language information retrieval systems allow the users to input the key words in their own languages and then the systems will retrieve the relevant documents written in the other language that the users want to search based on the queries the users inputted.

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://www.altavista.com

There are many different approaches of CLIR. The first is the translation methods. There are two kinds of approaches usually adopted: translation approach and statistical approach. The translation approach uses the bilingual dictionaries, ontology, or thesaurus to translate either queries or documents. The statistical approach uses pre-constructed bilingual corpora to extract the cross-lingual associations without any language translation methods [1–3]. The translation approach is restricted with the coverage and the precision of the dictionaries. The statistical approach can extract bilingual lexicons automatically; however, it bases on a well-constructed and large-scaled bilingual corpus which requires a lot of human effort.

In translation approach, there are two different targets to do the translation. One is document translation; the other is query translation. The document translation approach is to translate all the documents in the collection from the target language to the source language the users use. Then, while the users give the query, the system will do a monolingual information retrieval. The query translation translates the query in the source language that user inputted into the target language and then retrieval the documents which is written in the target language. The document translation approach is possible if there exists a high quality machine translation system. [4, 5] However, the document translation approach is not very practical when the documents are not stable or can be updated frequently, like the web text retrieval.

In this paper, we propose a Korean-Chinese cross-language information retrieval system. We adopt the query-translation approach because it is effective. Moreover, the translation method, which is dictionary-based, does not involve a great deal of work. In CLIR, the most serious problem is that unknown words cannot be translated correctly. To resolve the problem, we utilize Wikipedia, an online encyclopedia, to expand our dictionary to make higher coverage of vocabulary. Another difficult issue involves translating Western people's names written in Korean into Chinese. As a solution, we propose a transliteration mapping method to deal with the problem.

The remainder of the paper is organized as follows. In Section 2, we give an overview of our system and describe its implementation, including the translation and indexing methods adopted. In Section 3, we detail the evaluation results of our CLIR system based on the topics and the document collections provided by NTCIR CLIR task, and discuss the effectiveness of our method, as well as some problems that have to be solved. Finally, in Section 4, we present our conclusions and indicate the direction of our future work.

2 System Description

Figure 1 shows the architecture of our CLIR system. It is comprised of four stages. First, a Korean query is chunked into several key terms, which are then translated into Chinese by three dictionaries. In the third stage, we disambiguate the translated terms and transform them into a Lucene query. Finally, the query is sent to the Lucene IR engine and the answer is retrieved.

2.1 Query Processing

Unlike English, Korean written texts do not have word delimiters. Spaces in Korean sentences separate eojeols, which are composed of a noun and a postposition, or a verb stem and a verb ending. Therefore, Korean text has to be segmented. There are two types of queries that the users might make. One is composed of several key words; the other is an natural language sentence. Therefore, we use two different segmentation methods to deal with these two query types separately.

Due to the characteristics of the Korean language, the keyword-typed queries written in Korean are comprised mainly of nouns. We use spaces to split the title into several eojeols, and

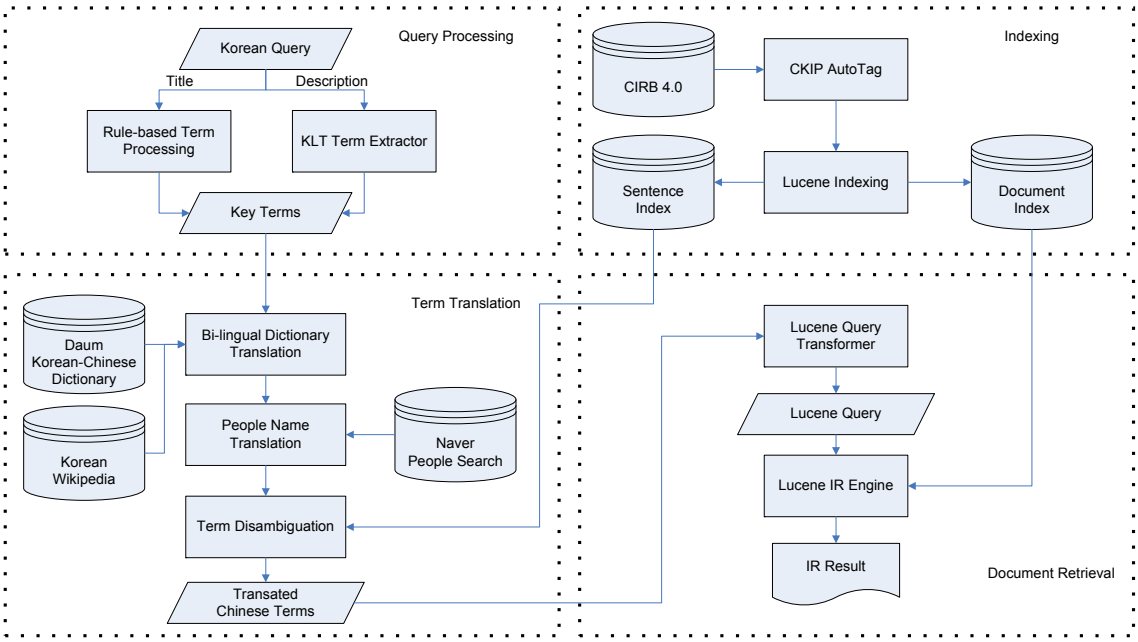

Figure 1: System Architecture of Our CLIR System

then remove the postpositions at the end of the eojeols according to our predefined rules.

For the natural language sentence of a Korean query, we use the KLT Term Extractor[4], developed by Kookmin University in Korea. KLT term extractor will do the word segmentation to extract vital key words which are useful for information retrieval and remove stop words.

2.2 Query Translation

2.2.1 Bilingual Dictionary Translation

Due to copyright restrictions, we use the free online Korean-Chinese dictionary provided by the Daum Korean web site[5]. We send the key terms obtained in the query processing stage to the online dictionary. However, as a general bilingual dictionary is not suitable for proper nouns, we use Wikipedia[6], an online encyclopedia, to expand our dictionary. In Wikipedia, an item might contain inter-language links to the same item in Wikipedia written in other languages. Therefore, we send a Korean term to Korean Wikipedia. If it contains an inter-language link to Chinese Wikipedia, we can find the corresponding Chinese word. This method is very efficient because it yields accurate Chinese translations of Korean words.

The Daum Korean-Chinese dictionary is written in simplified Chinese, as are many pages in Chinese Wikipedia. We use a simple mapping table to convert simplified Chinese characters to traditional Chinese characters.

If some terms cannot be found in the Daum dictionary or Wikipedia, we apply the maximal matching algorithm to split a long term into several shorter terms. Then, the shorter terms are sent to the Daum dictionary and Wikipedia to search for Chinese translations.

[4] http://nlp.kookmin.ac.kr/HAM/kor/index.html
[5] http://cndic.daum.net
[6] http://www.wikipedia.org

2.2.2 Person Name Translation

The person names may often appear in the query. Although Wikipedia contains many famous people's names around the world, some people's names are still excluded. Therefore, it is necessary to deal with this person name translation problem. Unlike Korean-English or Korean-Japanese CLIR, transliteration methods are not appropriate for Korean-Chinese CLIR because so many Chinese characters have the same pronunciation in Korean. Besides, to translate Japanese personal names, Korean uses the Hangul alphabet to pronounce the names of Japanese people; however, Chinese uses original Chinese characters with Mandarin pronunciation, instead of Japanese pronunciation of Chinese characters. Thus, transliteration methods are not useful in this context. To solve the problem, we use Naver People Search[7], a database containing the basic profiles of famous people, including their original names. We can submit person names in Korean to Naver people search and get their original names. If the original name is composed of Chinese characters, it is clearly Chinese, Japanese, or Korean; therefore, we can send it to next stage directly, i.e., the disambiguation stage. If, however, the original name is in English, we use the English name translation table provided by Taiwan's Central News Agency (CNA)[8] to translate it into Chinese and the proceed to the next stage.

2.3 Term Disambiguation

In the past, the Korean language adopted many Chinese words. More than half of its vocabulary comprises Chinese words. Now, however, Koreans use Hangul, an alphabet writing system, instead of Chinese characters, which is an ideograph writing system. As a result, many different Chinese loanwords have the same pronunciation when written in the Hangul alphabet. For example, the four different Chinese loanwords with different meanings: "理想" (ideal), "以上" (above), "異常" (unusual), and "異狀" (indisposition) are written in the same way as the Hangul word "이상" because their pronunciation is the same in Korean. This creates a very serious ambiguity problem when Korean is translated into Chinese. Therefore, choosing the correct translation term among translation candidates is important.

For each term in a given query Q, there may be several possible translation candidates. To select the best translation term among all the candidates, we must not only consider the original query term qt but also consider all the other terms in Q and their translation candidates. We denote the j-th translation candidate for the i-th term qt_i in Q as tc_{ij}. We adopt the mutual information score (MI score) [6] to evaluate the co-relation between the tc_{ij} and all translation candidates of all the other terms in Q. The MI score of tc_{ij} given Q is calculated as follows:

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

where $Z(qt_x)$ is the number of translation candidates of the x-th query term qt_x; $Pr(tc_{ij}, tc_{xy})$ is the probability that tc_{ij} and tc_{xy} co-occur in the same sentence; and $Pr(tc_{ij})$ is the probability of tc_{ij}. The values of the probabilities are obtained from Chinese Information Retrieval Benchmark (CIRB) Chinese corpus which is provided by NTCIR CLIR task [7]. The higher the translation candidate's MI score is, the higher weight is assigned to it in the retrieval module.

[7] http://people.naver.com

[8] http://client.cna.com.tw/name/

2.4 Chinese Document Indexing

The Chinese documents we use is CIRB 4.0 documents which is provided by NTCIR. The CIRB 4.0 documents are pre-processed to remove noise and then segmented by CKIP AutoTag [8] to obtain words and part-of-speech (POS). We use Lucene[9], an open source information retrieval engine, to index Chinese documents. Our index is based on Chinese characters.

2.5 Lucene Queries

After processing a Korean query into several terms and translating it into Chinese, we transform the Chinese terms into a Lucene Query. Different Chinese terms are separated by a space, which means an "OR" operator in the Lucene format. If a term has different translation candidates, the weight of the candidate with highest mutual information score will be increased by 1 by the boost operator. The other candidates are boosted by a weight that is the reciprocal of the total number of candidates. The boost operation affects the ranking of the documents the Lucene returns. The default boost value of each terms is 1, and we decrease the weight of the candidates with lower mutual information score to make them not affect the ranking so much.

3 Evaluation and Analysis

In order to evaluate our CLIR system, we use the topics and the document collections which is provided by NTCIR-6 CLIR task [7]. The topics contains 50 Korean queries composed of four parts: title, description, narration, and keywords. We use these topics as the queries that users inputted in our system.

The main metric to evaluate the performance of information retrieval is Mean Average Precision (MAP) [9]. Average precision is based on the whole list of documents returned by the system and emphasizes returning more relevant documents earlier. The Mean Average Precision is the mean value of the average precisions computed for each query. Besides, R-precision [10] is also a good metric which is the precision among the front of R relevant documents.

There are two kinds of relevance judgments: Rigid and Relax. A document is rigid relevant if it is highly relevant; a document is relax relevant if it is highly relevant or partial relevant. Our evaluation is based on the 50 topics which is selected by NTCIR-6 CLIR task to compute among all 140 topics they provided.

In order to evaluate the effectiveness of our CLIR system, we build a monolingual Chinese IR system for comparison. NTCIR-6 CLIR test set also contains the Chinese topics which meanings are the same as Korean ones. We use these Chinese topics as queries and apply CKIP AutoTag to do Chinese word segmentation and remove Chinese stop words. Then, we use Lucene search engine we use in our CLIR system to retrieve related Chinese documents.

We do the four different runs:

- **KC-title-run**: a run using a Korean title field to retrieve Chinese documents.
- **KC-description-run**: a run using a Korean description field to retrieve Chinese documents.
- **CC-title-run**: a monolingual Chinese run using Chinese title field to retrieve Chinese documents.

[9] `http://lucene.apache.org/`

Table 1: Evaluation Results

Run	Rigid		Relax	
	MAP	R-precision	MAP	R-precision
KC-title-run	0.1118	0.1420	0.1392	0.1781
KC-description-run	0.1022	0.1311	0.1274	0.1760
CC-title-run	0.1501	0.1961	0.2141	0.2747
CC-description-run	0.1567	0.2111	0.2157	0.2788

- **CC-description-run**: a monolingual Chinese run using Chinese description field to retrieve Chinese documents.

Table 1 shows the performance of our Korean-Chinese CLIR system and the monolingual Chinese IR system. The performance of Korean-Chinese CLIR is not as good as that of Chinese monolingual IR. We have investigated why it is difficult to retrieve high precision answers to some queries.

3.1 Problems of Bilingual Dictionaries

We use a general bilingual dictionary and Wikipedia to translate most of the words in 50 topics provided by NTCIR-6 CLIR task. Although we have used Wikipedia to expand our dictionary, there are some problems that cause translations to fail. The first problem is that there are still some unknown words. For example, the word "배아" (embryo) is not listed in the dictionaries. The other problem is that the dictionaries do not always have the proper translation candidates of the words and terms in queries. For instance, the word "감청" (monitor) is not translated correctly because the dictionary lacks the correct translation and provides another translation instead, i.e., "紺青" (deep blue). Also, the word "암" (cancer) in one topic is translated as "岩" (rock), "庵" (nunnery), and "雌" (female), but no correct translation, i.e., "癌" (cancer).

3.2 Different Phraseology Used in Taiwan and China

The Daum Korean-Chinese dictionary that we use was written people studying Mainland Chinese, i.e., Pinyin. However, the CIRB 4.0 document collection contains Taiwanese newspapers. Taiwanese people use traditional Chinese characters, whereas Mainland Chinese people use simplified characters. Besides the difference in characters, the vocabulary and grammar used in Taiwan and China are slightly different. The differences between Taiwanese Chinese and Mainland Chinese can make IR difficult.

The following are some examples of the difficulties we face. The term "휴대폰" (mobile phone) is translated into Mainland Chinese word as "移動電話" (the phone that can move); however, the correct word used in Taiwan is "手機" (the machine held in the hand). The word "유전자" (gene) is translated to "遺傳子" (the factor of heredity) , not to correct word "基因" (the Mandarin transliteration of the English word "gene") used in Taiwan. The word "인터넷" (internet) in some topics is translated to "互聯網" (the net connecting to each other) , but the correct word used in Taiwan is "網際網路" (cyber network).

3.3 The Limitations of Maximal Matching Algorithm

If a term is not defined in our dictionaries, we split it into several shorter terms by the maximal matching algorithm discussed in Section 2.2.1. In some cases, however, the algorithm do not

segment a term correctly. For example, for the term "비 접 촉 형"(contactless), the correct segmentation is 비(not)-접 촉(contact)-형(type). However, it is segmented as 비 접-촉-형 so that the wrong word, "비 접"(convalescing), is retrieved.

3.4 Different Expressions Used in Korean and Chinese

In some topics, different expressions used in Korean and Chinese may cause translation problems. In one of the topic, the word "10대" refers to people aged between 10 and 19. Similarly, "20대" means people aged from 20 to 29. Therefore, the corresponding translation of the word "10대" in this topic is "青少年" (teenager). However, our system translates the numbers and the Hangul characters separately so that the final translation is "10代" (ten generations). This is a semantic problem that our system has difficulty coping with.

Another problem relates to abbreviations used in Chinese. For instance, in another topic, "외국인 노동자" (foreign worker) is translated into "外國人勞工" (foreign worker) by our system. However, in Taiwanese newspapers, the abbreviation "外勞" , which is composed of the first characters of the two words : "外國人" (foreigner) and "勞工" (worker), is used more frequently. Our translation in one of the topic for the phrases "원자능 반대" is "反對 核能" (anti-nuclear), but the abbreviation "反核" is frequently used.

4 Conclusions and Future Works

We have described our Korean-Chinese CLIR system. It is based on a query-translation approach and uses a general Korean-Chinese dictionary and Wikipedia to translate words and terms. To obtain person names, we use the Naver people search website and the CNA transliteration table to translate the names.

We have evaluated the performance of our Korean-Chinese CLIR system with the Korean topics and the Chinese document collection which is provided by NTCIR-6 CLIR task. Our translation method is effective, but there are still some cases where the precision is low. We believe the problems are due to the limitations of the dictionaries, the different phraseology used in Taiwan and China, and the expressions used in Chinese and Korean.

In our future work, we will apply a Chinese thesaurus to overcome the problem of different Chinese phraseology and use more bilingual dictionaries to reduce the number of unknown words. We will also incorporate a query expansion method into our CLIR system to improve its precision.

References

[1] S. Dumais, T. Letsche, M. Littman, and T. Landauer, "Automatic cross-language retrieval using latent semantic indexing", in *AAAI Symposium on CrossLanguage Text and Speech Retrieval*. 1997, American Association for Artificial Intelligence.

[2] Bob Rehder, Michael L. Littman, Susan Dumais, and Thomas K. Landauer, "Automatic 3-language cross-language information retrieval with latent semantic indexing", in *Sixth Text REtrieval Conference (TREC-6)*, 1997.

[3] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking, "Translingual information retrieval: Learning from bilingual corpora", *Artificial Intelligence*, vol. 103, pp. 323–345, 1998.

[4] Oh-Wook Kwon, I.S. Kang, J-H Lee, and G.B. Lee, "Cross-language text retrieval based on document translation using japanese-to-korean mt system", in *NLPRS*, 1997, pp. 101–106.

[5] Douglas W. Oard and Paul Hackett, "Document translation for the cross-language text retrieval at the university of maryland", in *the Sixth Text REtrieval Conference (TREC-6).*

[6] Hee-Cheol Seo, Sang-Bum Kim, Ho-Gun Lim, and Hae-Chang Rim, "Kunlp system for ntcir-4 korean-english cross-language information retrieval", in *NTCIR-4*, Tokyo, 2004.

[7] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, and Hsin-Hsi Chen, "Overview of clir task at the sixth ntcir workshop", *Proceedings of NTCIR-6 Workshop Meeting*, 2007.

[8] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, and Wen-Lian Hsu, "Asqa: Academia sinica question answering system for ntcir-5 qa", in *NTCIR-5*, Tokyo, 2005.

[9] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison, "A study of information seeking and retrieving", *Journal of the American Society for Information Science*, vol. 39, no. 3, pp. 161–176, 1988.

[10] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.