IASL System for NTCIR-6 Korean-Chinese Cross-Language Information Retrieval

Yu-Chun Wang [12], Cheng-Wei Lee [13], Richard Tzong-Han Tsai [1], Wen-Lian Hsu[1]

[1]Institute of Information Science, Academia Sinica, Taiwan, R.O.C
[2] Department of Electrical Engineering, National Taiwan University, R.O.C
[3] Department of Computer Science, National Tsing-Hua University, R.O.C

{albyu,aska,thtsai,hsu}@iis.sinica.edu.tw,

Abstract

This paper describes our Korean-Chinese cross-language information retrieval system for NTCIR-6. Our system uses a bilingual 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.

Keywords: *Korean-Chinese cross-language information retrieval, query translation, transliteration*

1. Introduction

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 developped, like Google, Yahoo, AltaVista, and son 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-languaged web search engines.

Therefore, the research of cross language information retrieval 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.

This paper reports on our participation in the Korean-Chinese cross-language information retrieval (CLIR) task at the NTCIR-6 workshop. 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. 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 official results of our participation in the NTCIR-6 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

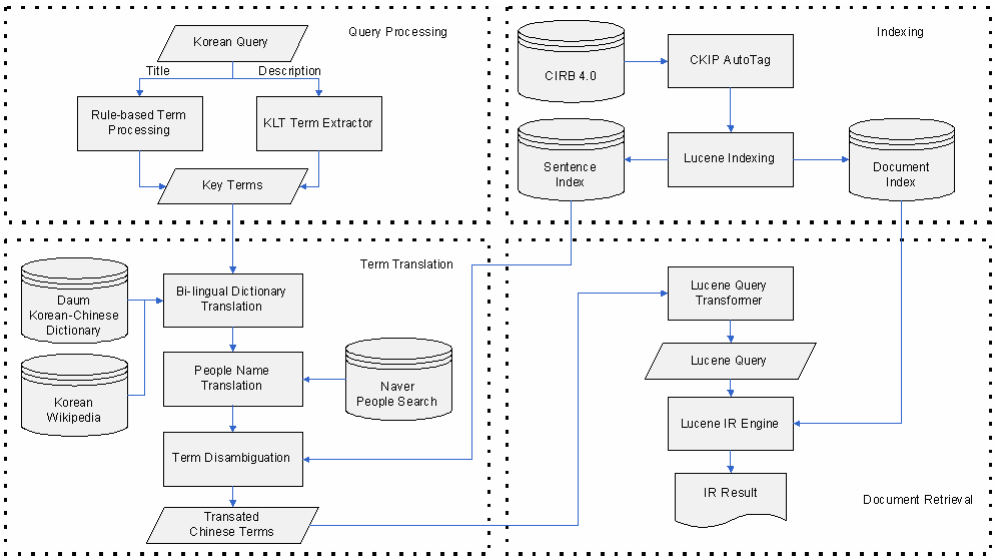

Figure 1 – System Architecture

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. We use two different segmentation methods, one for the title of the query and the other for the descriptive part.

Due to the characteristics of the Korean language, the titles of queries written in Korean are comprised mainly of nouns. We use spaces to split the title into several eojeols, and then remove the postpositions at the end of the eojeols according to our predefined rules.

For the descriptive part of a Korean query, we use the KLT Term Extractor[1], developed by Kookmin University in Korea, to extract vital key words 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 [2]. 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, 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 rules listed in Table 1, which are based on the properties of Korean morphemes, 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.

2.2.2. Person Name Translation

The NTCIR topics contain some personal names not listed in Korean Wikipedia. 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[3], 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

Table 1 The Rules for Splitting Korean Terms Based on Morphemes

Number of Character	Separation
3	ABC→A, BC
	ABC→AB, C
4	ABCD→AB, CD
	ABCD→A, BCD
	ABCD→ABC, D
5	ABCDE→AB, CDE
	ABCDE→ABC,DE
6	ABCDEF→AB, CD, EF
	ABCDEF→ABC, DEF
7	ABCDEFG→AB, CD, EFG
	ABCDEFG→AB, CDE, FG
	ABCDEFG→ABC, DE, FG
8	ABCDEFGH→AB, CD, EF, GH
9	ABCDEFGHI→AB, CD, EF, GHI
10	ABCDEFGHIJ→AB, CD, EF, GH, IJ

* Each capital letter represents one Hangul character.

Central News Agency (CNA) to translate it into Chinese and the proceed to the next stage.

2.2.3. 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) [4] 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:

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

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. The higher the translation candidate's MI score is, the higher weight is assigned to it in the retrieval module.

2.2.4. Chinese Document Indexing

CIRB 4.0 documents are pre-processed to remove noise and then segmented by CKIP AutoTag[5, 6] to obtain words and part-of-speech (POS). We use Lucene[7], an open source information retrieval engine, to index Chinese documents. Our index is based on Chinese characters.

2.2.5. Lucene Queries

After chunking 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. Official Results and Analysis

The main metric to evaluate the performance of information retrieval is Mean Average Precision (MAP). 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

Table 2. The Official Results of Cross Language IR

Run	Rigid		Relax	
	MAP	R-prec	MAP	R-prec
IASL-K-C-T-01	0.1118	0.1420	0.1392	0.1781
IASL-K-C-D-01	0.1022	0.1331	0.1274	0.1760

precisions computed for each query. Besides, R-precision is also a good metric which is provided by the evaluation result of NTCIR-6. R-precision is the precision among the front of R 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 to compute among all 140 topics that we submitted.

We submitted two CLIR runs:

- **IASL-K-C-T-01**: a run using a Korean title field to retrieve Chinese documents.
- **IASL-K-C-D-01**: a run using a Korean description field to retrieve Chinese documents.

Table 2 shows the performance of our Korean-Chinese CLIR system. The performance 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 all 140 topics provided by NTCIR. 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) in topic 3 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. In topic 24, 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 topic 46 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. In topic 41, 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) in topic 46 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. In topic 53, the word "인터넷" (internet) 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 Korean Processing Rules

If a term is not defined in our dictionaries, we split it into several shorter terms by the predefined rules discussed in Section 2.2.1. In some cases, however, the rules do not segment a term correctly. For example, in topic 58, for the term 비접촉형(contactless), the correct segmentation is 비(not)-접촉(contact)-형(type). However, by our rules, it is segmented as 비접-촉형 so that the wrong word, 비접(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 topic 18, the word "10 대" refers to people aged between 10 and 19. Similarly, "20 대" means people aged from 20 to 29. Therefore, the corresponding translation of the word "10 대" 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 topic 39, "외국인 노동자" (foreign worker) is translated into "外國人勞工" (foreign worker) by our system. However, in Tai-

wanese 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 The same problem also occurs in topic 96. Our translation 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, which was the only entry in the NTCIR-6 CLIR K-C task. 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 result of the official runs. 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.

5. Acknowledgments

This research was supported in part by the National Science Council under Grant No. NSC94-2752-E-001-001-PAE.

We would like to thank the Chinese Knowledge and Information Processing group (CKIP) in Academia Sinica for providing us with AutoTag for Chinese word segmentation.

References

[1] "KLT Term Extractor for Information Retrieval," Natural Language Processing Laboratory, Kooknin University.

[2] Korean-Chinese Dictionary, Daum. http://cndic.daum.net

[3] "Naver People Search." http://people.naver.com

[4] Hee-Cheol Seo, Sang-Bum Kim, Ho-Gun Lim, Hae-Chang Rim, "KUNLP System for NTCIR-4 Korean-English Cross-Language Information Retrieval," *NTCIR*, 2004.

[5] "CKIP AutoTag," Academia Sinica. http://ckipsvr.iis.sinica.edu.tw

[6] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Cheng-Lung Sung, Chia-Wei Wu, Yu-Ren Chen, Shih-Hung Wu, Wen-Lian Hsu, "ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA," *NTCIR*, 2005.

[7] "Apache Lucene," The Apache Software Foundation. http://lucene.apache.org