

Korean-Chinese Person Name Translation for Cross Language Information Retrieval

Yu-Chun Wang^{ab}, Yi-Hsun Lee^a, Chu-Cheng Lin^{ac},
Richard Tzong-Han Tsai^{d*}, Wen-Lian Hsu^a

^aInstitute of Information Science, Academia Sinica, Taiwan

^bDepartment of Electrical Engineering, National Taiwan University, Taiwan

^cDepartment of Computer Science and Information Engineering, National Taiwan University, Taiwan

^dDepartment of Computer Science and Engineering, Yuan Ze University, Taiwan

{albyu, rog, as1986}@iis.sinica.edu.tw

thtsai@saturn.yzu.edu.tw

hsu@iis.sinica.edu.tw

*corresponding author

Abstract. Named entity translation plays an important role in many applications, such as information retrieval and machine translation. In this paper, we focus on translating person names, the most common type of name entity in Korean-Chinese cross language information retrieval (KCIR). Unlike other languages, Chinese uses characters (ideographs), which makes person name translation difficult because one syllable may map to several Chinese characters. We propose an effective hybrid person name translation method to improve the performance of KCIR. First, we use Wikipedia as a translation tool based on the inter-language links between the Korean edition and the Chinese or English editions. Second, we adopt the Naver people search engine to find the query name's Chinese or English translation. Third, we extract Korean-English transliteration pairs from Google snippets, and then search for the English-Chinese transliteration in the database of Taiwan's Central News Agency or in Google. The performance of KCIR using our method is over five times better than that of a dictionary-based system. The mean average precision is 0.3490 and the average recall is 0.7534. The method can deal with Chinese, Japanese, Korean, as well as non-CJK person name translation from Korean to Chinese. Hence, it substantially improves the performance of KCIR.

Keywords: Person Name Translation, Korean-Chinese Cross Language Information Retrieval

1. Introduction

Named entity (NE) translation plays an important role in machine translation, information retrieval, and question answering. It is a particularly challenging task because, although there are many online bilingual dictionaries, they usually lack domain specific words or NEs. Furthermore, new NEs are generated everyday, but the content of bilingual dictionaries cannot be updated frequently. Therefore, it is necessary to construct a named entity translation (NET) system.

Economic ties between China and Korea have become closer as China has opened its markets further, and demand for the latest news and information from China continues to grow rapidly in Korea. One key way to meet this demand is to retrieve information written in Chinese by using Korean queries, referred to as Korean-Chinese cross-language information retrieval (KCIR). The main challenge involves translating NEs because they are usually the main concepts of queries. In Chen (1998), the authors romanized Chinese NEs and selected their

English transliterations from English NEs extracted from the Web by comparing their phonetic similarities with Chinese NEs. Al-Onaizan and Knight (2002) transliterated an NE in Arabic into several candidates in English and ranked the candidates by comparing their occurrences in several English corpora. In the above works, the target languages are alphabetic; however, in K-C translation, the target language is Chinese, which uses an ideographic writing system. Korean-Chinese NET is much more difficult than NET considered in previous works because, in Chinese, one syllable may map to tens or hundreds of characters. For example, if an NE written in Korean comprises three syllables, there may be thousands of translation candidates in Chinese.

In this paper, we focus on translating person names, and propose an effective hybrid method to improve the performance of our Korean-Chinese cross-language information retrieval system

2. Difficulties in Korean-Chinese Person Name Translation for IR

In this section, we discuss the phenomena observed in the transliteration of person names in Korean and Chinese. We begin with a brief review of the relationship between the Korean and Chinese languages.

Korean is an Altaic language, while Chinese is a Sino-Tibetan language; hence, their phonology and grammar are quite different. Due to a long history of contact with Chinese, Koreans adopted Chinese characters and incorporated a lot of Chinese vocabulary into their language. Chinese characters used in Korean are called “Hanja”, and Chinese loanwords used in Korean are called Sino-Korean words.

The pronunciation of Hanja in Korean is very different from modern Chinese, Mandarin, because it follows the pronunciation of Middle Chinese; thus, it has not undergone many of the sound changes evident in modern Chinese. Interestingly, Song (2005) mentioned that over 52 percent of the words in the modern Korean vocabulary are Sino-Korean

In 1443, Koreans invented their own alphabetic writing system called “Hangul”. Each Hanja character has a corresponding Hangul character based on its Korean pronunciation. However, Hanja is only used in some limited domains now.

2.1. Korean Name Translation

Korean and Chinese name systems are very similar. Because of historical links, almost all Koreans have names that are exclusively Hanja (Han- "Chinese, "-ja "characters"). Therefore, the most straightforward way to translate a Korean name into Chinese is to adopt its Hanja equivalent. Take the Korean president's name “노무현” (No Mu-Hyeon) as an example. In this case, we can adopt the Hanja equivalent “盧武鉉” (Lu Wu-Xuan) directly. However, if a Korean's Hanja name is unknown, the name is translated character by character. Each Hangul character is basically translated into its corresponding Hanja character. For example, the name of the Korean actor “조인성” (Cho In-Seong) is usually translated as “趙仁成” because ‘조’ is mapped to ‘趙’, ‘인’ is mapped to ‘仁’, and ‘성’ is mapped to ‘成’. However, that translation may not be the same as the actor's Hanja name. In addition, some Hangul characters do not have corresponding Chinese characters, so Chinese characters with similar pronunciations are used to translate the Hangul characters. Take the Korean actress “김하늘” (Kim Ha-Neul) for example. Her given name “하늘” (“ha-neul” meaning “sky”) is a native Korean word that has no corresponding Hanja characters. We use “荷娜” (He-Na) or “哈嫩” (Ha-Nen), which have similar pronunciations to translate “하늘” (ha-neul). These examples show that there may be many Chinese translations for a Korean name. This phenomenon makes Korean-Chinese information retrieval more difficult because reporters usually use one or two common translations to write articles. However, we cannot guarantee that our translations are the most common ones.

2.2.Chinese Name Translation

To translate a Chinese person name written in Korean, we consider two ways that are used to translate a Chinese person name into spoken Korean. The first method uses Sino-Korean pronunciation. For example, consider the name “馬英九” (Ma Ying-Jiu, the ex-chairman of the Kuomintang (KMT), a Taiwanese political party); its Sino-Korean pronunciation is “마영구” (Ma Yeong-Gu). However, in recent years, Koreans have started to transliterate a Chinese person name based on its Mandarin pronunciation. Therefore, the name “馬英九” is transliterated to “마잉주” (Ma Ing-Ju). Translating Chinese person names by either method is a major challenge because one Hangul character corresponds to several Chinese characters that have the same pronunciation in Korean. This results in thousands of possible combinations of Chinese characters, making it very difficult to choose the right one. Therefore, we must develop different techniques to find the correct Chinese translation that is used in articles.

2.3.Japanese Name Translation

Chinese and Korean use different strategies to translate Japanese person names. Korean transliterates a Japanese person name into Hangul characters based on the name’s pronunciation in Japanese, whereas, Chinese speakers use the name in Kanji directly. Take the Japanese ex-premier “小泉純一郎” (Koizumi Junichiro) for example. In Korean, his name is transliterated into “고이즈미 준이치로” (Ko-i-jeu-mi Jun-i-chi-ro). In contrast, the Kanji name “小泉純一郎” (Xiao quan chun yi lang) is used directly in Chinese. Therefore, it is very difficult to translate Japanese names written in Korean into Chinese based on phonetic information.

2.4.Non-CJK Name Translation

In both Korean and Chinese, transliteration methods are used to translate non-CJK person names. Korean uses the Hangul alphabet for transliteration. Because of the phonology of Korean, some phonemes are changed during translation because the language lacks such phonemes as described in Oh (2003) In contrast, Chinese transliterates each syllable in a name into Chinese characters with similar pronunciation. Although there are some conventions for selecting transliteration characters, there are still many possible alternatives. For instance, Greenspan has several Chinese transliterations, such as “葛林斯班” (Ge-lin-si-ban) and “葛林斯潘” (Ge-lin-si-pan). In summary, it is difficult to match a non-CJK person name transliterated from Korean with its Chinese transliteration due to the latter’s variations. However, this task is the key to retrieving Chinese articles by using Korean queries.

3. Our Method

We now describe our Korean-Chinese person name/NE translation method for dealing with the problems described in Section 2. We either translate NE candidates from Korean into Chinese directly, or translate them into English first and then into Chinese.

3.1.Named Entity Selection

The first step is to identify which words in a query are NEs. In general, Korean queries are composed of several eojeols, each of which is composed of a noun followed by the noun’s postposition, or a verb stem followed by the verb’s ending. We remove the postposition or the ending to extract the key terms, and then select person name candidates from the key terms. Next, the maximum matching algorithm is applied to further segment each term into words in

the Korean-Chinese bilingual dictionary¹. If a segment’s length is equal to one, the term is regarded as an NE candidate to be translated.

3.2.Using Wikipedia for Translation

Wikipedia is a multilingual online encyclopedia comprised of content written by volunteers all over the world. Unlike traditional encyclopedias, the number of articles in Wikipedia increases rapidly, and each article usually lists hyperlinks to other relevant content. Currently, Wikipedia is available in 252 languages. It is a highly consistent, human-made corpus.

Each article in Wikipedia has an inter-language link to other language editions, which we exploit to translate NEs. An NE candidate is first input to the Korean Wikipedia, and the title of the matched article’s Chinese version is treated as the NE’s translation in Chinese. However, if the article lacks a Chinese version, we use the English edition’s version to acquire the NE’s translation in English. The English translation is then transliterated into Chinese by the method described in Section 3.5.

3.3.Using the Naver People Search Engine for Translation

The Naver people search engine is a translation tool that maintains a database of famous people’s basic profiles. If the person is from China, Japan, or Korea, the search engine returns his/her name in Chinese. For example, if we input the Japanese actor’s name “사나다 히로유키” (Sanada Hiroyuki) to the Naver people search engine, it will return his Japanese name “真田広之” with Chinese characters. In such cases, we can adopt the retrieved name directly. However, for other nationalities, the Naver search engine returns person names in English, and we have to translate them into Chinese. The translation method is also described in Section 3.5.

3.4.Web-Based Korean-English Transliterations

Obviously, the above methods cannot cover all possible translations used in newspaper articles. Therefore, we propose a web-based transliteration method. First, each NE candidate, NEC, is input to Google to retrieve snippets of relevant documents in the first ten pages. Second, we use the following template to extract the NEC’s English translation from the snippets.

$$NEC_K(e_1e_2e_3\dots e_n),$$

where NEC_K represents the NEC in Hangul characters and $e_i \in$ English alphabet. The string $e_1e_2e_3\dots e_n$ is regarded as the NEC’s English translation. In Section 3.5, we describe the method used to further transliterate an NEC’s English translation into Chinese.

3.5.Searching English-Chinese Transliteration in the CNA Database and Google

In this section, we discuss two methods that we use to transliterate English names generated by the above Korean-English translation methods into Chinese. The first obtains the Chinese translations of English names from Taiwan’s Central News Agency (CNA) database², which stores all the transliterations used by CNA since 1954. The second method exploits the Web to extract other possible Chinese transliterations not available in the CNA database. The latter have a significant influence on IR’s performance. The English name NECE is also input to Google and snippets are extracted from the first 10 returned pages. Then, we use the following template to extract the Chinese translation:

$$W_{boundary}C_1C_2C_3\dots C_m(NEC_E),$$

¹ <http://cndic.daum.net>

² <http://client.cna.com.tw/name/>

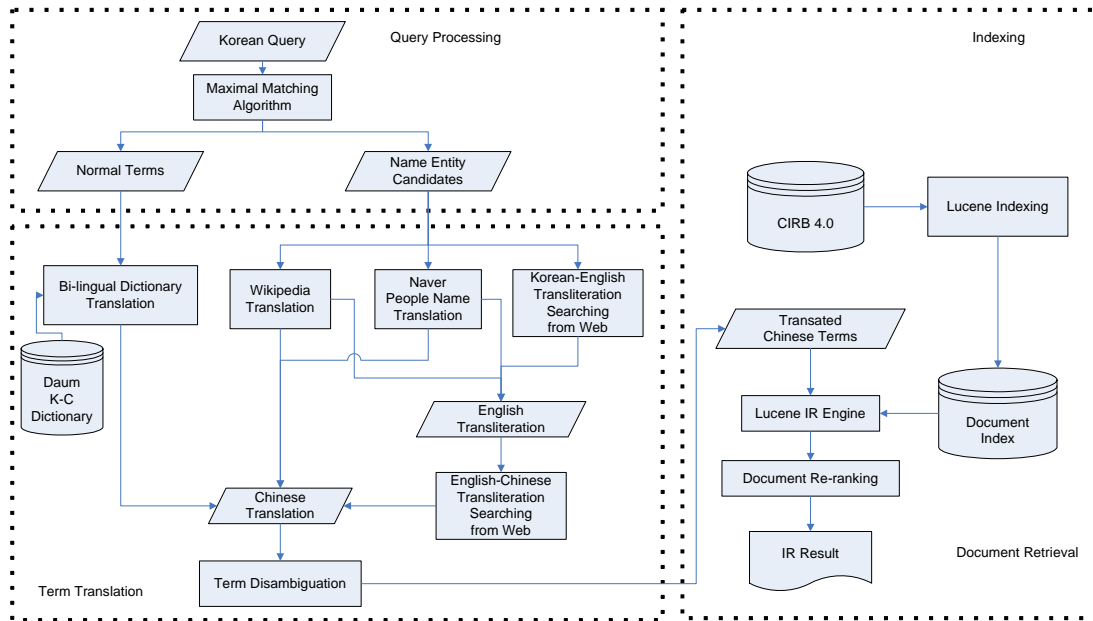


Figure 1: System Architecture of our CLIR system

where $w_{boundary}$ represents boundary words such as punctuation, titles, occupations, or nationalities; and $c_i \in$ Chinese characters. The string $c_1c_2c_3\dots c_m$ is regarded as the NEC_E 's Chinese translation.

4. System Description

We construct a Korean-Chinese cross language information retrieval (KCIR) system to determine how our person name translation methods affect KCIR's performance. A Korean query is translated into Chinese and then used to retrieve Chinese documents, as shown in Figure 1. The following are the four stages of our KCIR system.

4.1. Query Processing

First, the postposition or verb ending in each *eojeol* is removed. Then, NE candidates are selected using the method described in Section 3.1.

4.2. Query Translation

Key terms not selected as NE candidates are sent to the online Daum Korean-Chinese dictionary to get their Chinese translations, while NE candidates are translated into Chinese by the methods described in Sections 3.2 - 3.5. The Daum Korean-Chinese dictionary is written in simplified Chinese, as are many pages in Chinese Wikipedia. We use the conversion tool provided by Microsoft .Net Framework to convert simplified Chinese characters into traditional Chinese characters.

4.3. Term Disambiguation

A Hangul word may have many meanings. For instance, the word “이상” has four meanings: “理想” (ideal), “以上” (above), “異常” (unusual), and “異狀” (indisposition) because these four Sino-Korean words are written as the same Hangul word. This phenomenon causes ambiguity during information retrieval. To solve the problem, we adopt the mutual information score (MI score) to evaluate the co-relation between a translation candidate tc_{ij} for a term qt_i and all

translation candidates for all the other terms in Q ; tc_{ij} 's MI score given Q is calculated as follows:

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

where $Z(qt_x)$ is the number of translation candidates of the x -th query term qt_x ; tc_{xy} is y -th translation candidate for qt_x ; $\text{Pr}(tc_{ij}, tc_{xy})$ is the probability that tc_{ij} and tc_{xy} co-occur in the same sentence; and $\text{Pr}(tc_{ij})$ is the probability of tc_{ij} . Only the translation candidate with the highest score is used for retrieval.

4.4. Document Indexing and Retrieval Model

We use the Lucene information retrieval engine to index all documents and the bigram index based on Chinese characters. The Okapi BM25 function described in Robertson (1996) is used to score a retrieved document's relevance. The function is

$$\sum_{T \in Q} w \frac{(k_1 + 1) tf (k_3 + 1) qtf}{(K + tf)(k_3 + qtf)} + k_2 \cdot |Q| \cdot \frac{avdl - dl}{avdl + dl},$$

where Q is a query containing term T ; w is the Robertson-Sparck Jones weight described in Robertson (1988); K is $k_1((1-b)+b \cdot dl/avdl)$; k_1 , b , k_2 , and k_3 are parameters whose values are set to 3, 1, 3, and 0.3 respectively; tf is the term frequency within a specific document, and qtf is the term frequency within the topic from which Q was derived; dl is the document length; and $avdl$ is the average document length.

In addition, we employ the following document re-ranking function described in (Yang et al., 2007):

$$\sqrt{\frac{(\sum_{i=1}^K df(t, d_i) \times f(i)) / K}{DF(t, C) / R}} \times \sqrt{|t|} \quad df(t, d_i) = \begin{cases} 1 & t \notin d_i \\ 0 & t \in d_i \end{cases},$$

where d_i is the i th document; R is the total number of documents in the collection C ; $DF(t, C)$ is the number of documents containing a term t in C ; and $|t|$ is t 's length, $f(i)=1/\text{sqrt}(i)$.

5. Evaluation and Analysis

To evaluate our KCIR system, we use the topic collection and document collection of the NTCIR-5 and NTCIR-6 CLIR tasks. The document collection is the Chinese Information Retrieval Benchmark (CIRB) 4.0, which contains news articles published in four Taiwanese newspapers from 2000 to 2001. The topics have four fields: title, description, narration, and concentrate words. We select 18 topics containing person names and use the title field as the input query because it is similar to the queries input to search engines. The nationalities of the person names in the 18 topics are shown in Table 1.

Table 1: Nationalities of Person Names

Nationality	Count
Chinese	2
Japanese	4
Korean	4
non-CJK	9

Table 2: Evaluation Results

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
baseline	0.0491	0.0671	0.2765	0.2834
baseline + Wikipedia	0.1112	0.1443	0.4570	0.4578
baseline + person name translation	0.2835	0.3490	0.7382	0.7534
Google translation	0.1048	0.1312	0.4837	0.4893
Chinese monolingual	0.2698	0.3396	0.7708	0.7882

We construct five runs as follows:

- **Baseline:** using a Korean-Chinese dictionary-based translation.
- **Baseline + Wikipedia only:** the baseline system plus the Wikipedia translation.
- **Baseline + Person Name Translation Methods:** the baseline system plus our translation methods, namely, Wikipedia, the Naver people search engine, and web-based transliteration.
- **Google Translation:** using the Google translation tool.
- **Chinese monolingual:** using the Chinese versions of the 18 topics given by NTCIR directly.

We use the Mean Average Precision (MAP) and Recall in Saracevic (1988) to evaluate the performance of IR. NTCIR provides two kinds of relevance judgments: Rigid and Relax. A document is rigid-relevant if it is highly relevant to the topic; and relax-relevant if it is highly relevant or partially relevant to the topic.

The evaluation results demonstrate that our method improves KCIR substantially, as its performance is more than five times better than that of the baseline system. Interestingly, it is even better than Chinese monolingual IR. Wikipedia translation improves the performance, but not markedly because Wikipedia cannot cover some names. Google translation is not very satisfactory either, since many person names cannot be translated correctly. In the following, we analyze why our method can improve the overall performance and handle difficult cases. We also explain why the IR system with our person name translation method performs better than Chinese monolingual IR.

5.1. Effectiveness of Wikipedia

Wikipedia is a useful tool for translating famous person names. In our topics, names like “김대중” (Kim Dae-jung, South Korea’s ex-president), “김정일” (Kim Jong-il, North Korea’s leader), “주룽지” (Zhu Rong-ji, China’s ex-premier), and “빈라덴” (Osama bin Laden) are all translated correctly by Wikipedia and improve the performance of IR. In addition to person names, Wikipedia is also very useful for translating other kinds of NEs.

5.2. Effectiveness of the Naver People Search Engine

We observe that names, especially Japanese and some non-CJK person names, can be successfully translated by the Naver people search engine; for example, “코엔” (William Cohen, the ex-Secretary of Defense of U.S.) and “이치로” (Ichiro Suzuki, a Japanese baseball player). Therefore, the Naver search engine is effective for KCIR.

5.3. Effectiveness of Web-based Korean-English Transliteration

Our web-based method can successfully translate most non-CJK names that cannot be found in Wikipedia or the Naver people search engine. For example, our template can extract the

following non-CJK names from Google snippets successfully: “제니퍼 카프리아티” (Jennifer Capriati, the American tennis player), “데니스 티토” (Dennis Tito, the first space tourist), and “웬호 리” (Wen-ho Lee, the American scientist who stole nuclear secrets for China).

5.4. Effectiveness of Searching English-Chinese transliteration

All English names generated by the Naver people search or derived by Korean-English web-based transliteration can be successfully transliterated into Chinese by our English-Chinese transliteration method. Notably, our method can extract a larger number of possible Chinese transliterations. Take “Tito” for example: its six common Chinese transliterations: “迪托” (di-tuo), “蒂托” (di-tuo), “帝托” (di-tuo), “提托” (ti-tuo), “提多” (ti-duo), and “狄托” (di-tuo) can be extracted by our approach. This result is similar to that derived by query expansion. Under our method, the rigid MAP of this topic achieves 0.8361, which is much better than that of the same topic in the Chinese monolingual run (0.4459) because the Chinese topic has only one transliteration “帝托” (di-tuo).

5.5. Error Analysis

Person names that cannot be translated correctly can be divided into two categories. The first contains names not selected as NE candidates. The two Japanese person names “후지모리” (Alberto Fujimori, Peru’s ex-president) and “모리” (Yoshiro Mori, the ex-premier of Japan) are in this category. In the name “후지모리” (Fujimori), the first two characters “후지” (hind legs) and the last two characters “모리” (profiting) are Sino-Korean words, so the name is regarded as a compound word, not an NE. The Japanese surname “모리” (Mori) is the same because it is also a Sino-Korean word.

The other category contains names with few relevant web pages, like the two non-CJK names “홀링스위스” (Holingswiss) and “안토니오 토디” (Antonio Toddy). We can only obtain a few relevant web documents from web sites related to NTCIR. This means that, except for NTCIR, these names do not appear in any of the web documents maintained by Google. They might be a error transliteration or very obscure.

6. Conclusion

In this paper, we describe the difficulties that arise in translating person names from Korean to Chinese for IR. We propose a hybrid method for Korean-Chinese person name translation that exploits Wikipedia, the Naver people search engine, and the Google search engine. To evaluate our method, we use the topic and document collection of the NTCIR CLIR task. Our method’s performance on KCIR is over five times better than that of a dictionary-based translation system. Moreover, its average MAP score is 0.3490, which is even better than that of the Chinese monolingual IR system. The proposed method can deal with Chinese, Japanese, Korean, as well as non-CJK person name translation. Hence, it substantially improves the performance of KCIR.

References

- Al-Onaizan, Y. and K. Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pp. 400-408.
- Chen, H.-H., S.-J. Huang, Y.-W. Ding and S.-C. Tsai. 1998. Proper name translation in cross-language information retrieval. *Proceedings of 17th COLING and 36th ACL*, pp. 232-236.
- Oh, M. 2003. English Fricatives in Loanword Adaption. *Explorations in Korean Language and Linguistics*, pp. 471-87.
- Robertson, S.E. and K.S. Jones. 1988. Relevance weighting of search terms. *Taylor Graham Series In Foundations Of Information Science*, pp. 143-160.
- Robertson, S.E., S. Walker, M. Beaulieu, M. Gatford and A. Payne. 1996. Okapi at TREC-4. *Proceedings of the Fourth Text Retrieval Conference*, pp. 73-97.
- Saracevic, T., P. Kantor, A. Y. Chamis and D. Trivison. 1988. A Study of Information Seeking and Retrieving. *Journal of the American Society for Information Science*, 39(3),161-76.
- Song, J. J. 2005. *The Korean Language Structure, use and context*. Oxon: Routledge.
- Yang, L., D. Ji and M. Leong. 2007. Document reranking by term distribution and maximal marginal relevance for Chinese information retrieval. *Information Processing and Management*, 43(2), 315-26.