Learning Patterns from the Web to Translate Named Entities for Cross Language Information Retrieval

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

[†]Institute of Information Science, Academia Sinica, Taiwan
[‡]Department of Eletrical Engineering, National Taiwan University, Taiwan
[§]Department of Computer Science and Engineering, Yuan Ze University, Taiwan
albyu@iis.sinica.edu.tw
thtsai@saturn.yzu.edu.tw
hsu@iis.sinica.edu.tw
[*]corresponding author

Abstract

Named entity (NE) translation plays an important role in many applications. In this paper, we focus on translating NEs from Korean to Chinese to improve Korean-Chinese cross-language information retrieval (KCIR). The ideographic nature of Chinese makes NE translation difficult because one syllable may map to several Chinese characters. We propose a hybrid NE translation system. First, we integrate two online databases to extend the coverage of our bilingual dictionaries. We use Wikipedia as a translation tool based on the inter-language links between the Korean edition and the Chinese or English editions. We also use Naver.com's people search engine to find a query name's Chinese or English translation. The second component is able to learn Korean-Chinese (K-C), Korean-English (K-E), and English-Chinese (E-C) translation patterns from the web. These patterns can be used to extract K-C, K-E and E-C pairs from Google snippets. We found KCIR performance using this hybrid configuration over five times better than that a dictionary-based configuration using only Naver people search. Mean average precision was as high as 0.3385 and recall reached 0.7578. Our method can handle Chinese, Japanese, Korean, and non-CJK NE translation and improve performance of KCIR substantially.

1 Introduction

Named entity (NE) translation plays an important role in machine translation, information retrieval, and question answering. It is a 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 bilingual dictionaries cannot update their contents 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 et al., 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. Yaser Al-Onaizan (Al-Onaizan and Knight, 2002)

transliterated an NE in Arabic into several candidates in English and ranked the candidates by comparing their counts in several English corpora. Unlike the above works, whose target languages are alphabetic, 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 possible translation candidates in Chinese.

In this paper, we propose an effective hybrid NET method which can help improve performance of cross-language information retrieval systems. We also describe the construction of a Korean-Chinese CLIR system able to evaluate the effectiveness of our NE translation method.

2 Difficulties in Korean-Chinese Named Entity Translation for IR

2.1 Korean NET

Most Korean NEs originate from Hanja. Therefore, the most straightforward way to translate a Korean name into Chinese is to use its Hanja equivalent. Take the name of Korea's president, "노무현" (No Mu-hyeon), as an example. We can directly convert it to its Hanja equivalent: "盧武鉉" (Lu Wu-Xuan). Or in the case of the city name "부산" (Pusan/釜山/Fu-shan) and the company name "삼성" (Samsung/三星/San-xing), Chinese also presents Hanja equivalents.

If the Hanja name is unknown, the name is translated character by character. Each Hangul character is basically translated into a corresponding Hanja character. For example, the name of the Korean actor "조인성" (Cho In-seong) is usually translated as "趙仁成" (Zhao Ren-cheng) because '조' is mapped to '趙', '인' mapped to '仁', and '성' mapped to '成'. However, that translation may differ from the person's given Hanja name.

For native Korean NEs which have no corresponding Hanja characters, we must turn to transliteration or convention. Take the name of South Korea's capital "서울" (Seoul) as an example. Before 2005, Chinese media and government used the old Hanja name of the city "漢城" (Han-cheng), which was used during Joseon dynasty (A.D. 1392–1910). However, after 2005, Chinese switched to using the transliteration "首爾" (Shou-er) instead of "漢城" at the request of the Seoul Metropolitan Government. This example illustrate how more than one Chinese translation for a Korean name is possible, a phenomenon which, at times, makes Korean-Chinese information retrieval more difficult.

2.2 Chinese NET

To translate a Chinese NE written in Hangul, we begin by considering the two C-K NET approaches. The older is based on the Sino-Korean pronunciation and the newer on the Mandarin.

For example, "臺灣" (Taiwan) used to be transliterated solely as "대만" (Dae-man). However, during the 1990s, transliteration based on Mandarin pronunciation became more popular. Presently, the most common transliteration for "臺灣" is "타이완" (Ta-i-wan), though the Sino-Korean-based "대만" is still widely used. For Chinese personal names, both ways are used. For example, the name of Chinese actor Jackie Chan ("成龍" Cheng-long) is variously transliterated as "성룡" Seong-ryong (Sino-Korean) and "청룽" Cheong-rung (Mandarin).

Translating Chinese NEs by either method is a major challenge because each Hangul character may correspond to several different Chinese characters that have similar pronunciations in Korean. This results in thousands of possible combinations of Chinese characters, making it very difficult to choose the most widely used one one.

2.3 Japanese NET

Japanese NEs may contain Hiraganas, Katakanas, or Kanjis. For each character type, J-C translation rules may be similar to or very different from K-C translation rules. Some of these rules are based on Japanese pronunciation, while some are not. For NEs composed of all Kanjis, their Chinese translations are generally exactly the same as their Kanji written forms. In contrast, Japanese NEs

are transliterated into Hangul characters. Take "名古屋" (Nagoya) for example. Its Chinese translation "名古屋" is exactly the same as its Kanji written form, while its pronunciation (Ming Gu Wu) is very different from its Japanese pronunciation. This is different from its Korean translation, "나고야" (Na go ya). In this example, we can see that, because the translation rules in Chinese and Korean are different, it is ineffective to utilize phonetic similarity to find the Chinese translation equivalent to the Korean translation.

2.4 Non-CJK NET

In both Korean and Chinese, transliteration methods are mostly used to translate non-CJK NEs. Korean uses the Hangul alphabet for transliteration. Because of the phonology of Korean, some phonemes are changed during translation because the language lacks these phonemes. (Oh, 2003; Lee, 2003) In contrast, Chinese transliterates each syllable in a NE into Chinese characters with similar pronunciation. Although there are some conventions for selecting the transliteration characters, there are still many possible transliterations since so many Chinese characters have the same pronunciation. For instance, the name "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 NE transliterated from Korean with its Chinese transliteration due to the latter's variations.

3 Our Method

In this section, we describe our Korean-Chinese 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. Our method is a hybrid of two components: extended bilingual dictionaries and web-based NET.

3.1 Named Entity Candidate 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 Daum Korean-Chinese bilingual dictionary[1]. If the length of any token segmented from a term is 1, the term is regarded as an NE to be translated.

3.2 Extension of Bilingual Dictionaries

Most NEs are not included in general bilingual dictionaries. We adopt two online databases to translate NEs: Wikipedia and Naver people search.

3.2.1 Wikipedia

In Wikipedia, each article has an interlanguage link to other language editions, which we exploit to translate NEs. Each NE candidate is first sent 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 to acquire the NE's translation in English. The English translation is then transliterated into Chinese by the method described in Section 3.3.3.

3.2.2 Naver People Search Engine

Most NEs are person names that cannot all be covered by the encyclopedia. We use Naver people search engine to extend the coverage of person names. Naver people search is a translation tool that maintains a database of famous people's basic profiles. If the person is from CJK, the search engine returns his/her name in Chinese; otherwise, it returns the name in English. In the former case, we can adopt the returned name directly, but in the latter, we need to translate the name into Chinese. The translation method is described in Section 3.3.3.

[1] http://cndic.daum.net

3.3 Translation Pattern from the Web

Obviously, the above methods cannot cover all possible translations of NEs. Therefore, we propose a pattern-based method to find the translation from the Web. Since the Chinese translations of some NEs cannot be found by patterns, we find their Chinese translations indirectly by first finding their English translations and then finding the Chinese translations. Therefore, we must generate K-C patterns to extract K-C translation pairs, as well as K-E and E-C patterns to extract K-E and E-C pairs, respectively.

3.3.1 Translation Pattern Learning

Our motivation is to learn patterns for extracting NEs written in the source language and their equivalents in the target language from the Web. First, we need to prepare the training set. To generate K-C and K-E patterns, we collect thousands of NEs that originated in Korean, Chinese, Japanese, or non-CJK languages from Dong-A Ilbo (a South Korean newspaper). Then, all the Korean NEs are translated into Chinese manually. NEs from non-CJK languages are also translated into English. To generate E-C patterns, we collect English NEs from the MUC-6 and MUC-7 datasets and translate them into Chinese manually.

We submit each NE in the source language (source NE) and its translation in the target language as a query to Google search engine. For instance, the Korean NE "메이저리그" and its translation "Major League" are first composed as a query "+메 이 저 리 그 + Major League", which is then sent to Google. The search engine will return the relevant web documents with their snippets. We collect the snippets in the top 20 pages and we break them into sentences. Only the sentences that contain at least one source NE and its translation are retained.

For each pair of retained sentences, we apply the Smith-Waterman local alignment algorithm to find the longest common string, which is then added to the candidate pattern pool. During the alignment process, positions where the two input sequences share the same word are counted as a match. The following is an example of a pair of sentences that contains "메이저리그" and its English translation, "Major League":

- "**메이저리그**(**Major League**)는 수많은 산고 끝에 탄생한 산물입니다"
- "미국 **메이저리그**(**Major League**)는,"

After alignment, the pattern is generated as:

<*Korean NE*>(<*English Translation*>)는

This pattern generation process is repeated for each NE-translation pair.

3.3.2 Translation Pattern Filtering

After learning the patterns, we have to filter out some ineffective patterns. First, we send a Korean NE, such as "메 이 저 리 그", to retrieve the snippets in the top 50 pages. Then, we apply all the patterns to extract the translations from the snippets. The correct rate of each translation pattern is calculated as follows: $CorrectRate = C_{correct}/C_{all}$, where $C_{correct}$ is the total number of correct translations extracted by the pattern and C_{all} is the total number of translations extracted by the pattern. If the correct rate of the pattern is below the threshold τ, the pattern will be dropped.

3.3.3 Pattern-Based NET

The translations of some NEs, especially from CJK, can be found comparatively easily from the Web. However, for other NEs, especially from non-CJK, this is not the case. Therefore, we split the translation process into two stages: the first translates the NE into its English equivalent, and the second translates the English equivalent into Chinese.

To find an NE's Chinese translation, we first apply the translation patterns to extract possible Chinese translations. If its Chinese translation cannot be found, the K-E patterns are used to find its English translation instead. If its English translation can be found, the E-C patterns are then used to find its 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. The following sections describe the four stages of our KCIR system. We use an example query, "코스보의 사태, 나토, 유엔" (Kosovo's situation, NATO, UN), to demonstrate the work flow of our system.

4.1 Query Processing

Unlike English, Korean written texts do not have word delimiters. Spaces in Korean sentences separate eojeols. First, the postposition or verb ending in each eojeol is removed. In our example query, we remove the possessive postposition "의" at the end of the first eojeol. Then, NE candidates are selected using the method described in Section 3.1. "코스보" (Kosovo) is recognized as an NE, and other terms "사태" (situation), "나토" (NATO), and "유엔" (UN) are general terms because they can be found in the bilingual dictionary.

4.1.1 Query Translation

Terms not selected as NE candidates are sent to the online Daum Korean-Chinese dictionary and Naver Korean-Chinese dictionary[2] to get their Chinese translations. In our example, the terms "사태" (situation), "나토" (NATO), and "유엔" (UN) can be correctly translated into Chinese by the bilingual dictionaries as "事態" (situation), "北大西洋公約組織" (NATO), and "聯合國" (UN), respectively.

We employ Wikipedia, Naver people search, and the pattern-based method simultaneously to translate the NE candidate "코스보" (Kosovo). Up to now, there is no article about Kosovo in Korean Wikipedia. Naver people search does not contain an article either because it is not a person name. Meanwhile, since the K-C translation patterns cannot extract any Chinese translations, the K-E patterns are used to get the English translations, such as "Kosovo", "Cosbo", and "Kosobo". The E-C patterns are then employed to get the Chinese translation from the three English translations. Among them, only Chinese translations for "Kosovo" can be found because the other two are either wrong or rarely used translations. The Chinese translations extracted by our patterns are "科索夫" (Ke-suo-fu), "科索伏" (Ke-suo-fu), and "科索沃" (Ke-suo-wuo). They are all correct transliterations.

[2] http://cndic.naver.com

4.2 Term Disambiguation

A Hangul word might have many meanings. Besides, sometimes the translation patterns might extract wrong translations of the NE. This phenomenon causes ambiguities during information retrieval and influence the performance of IR significantly. To solve this 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:

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

where $Z(qt_x)$ is the number of translation candidates of the x-th query term qt_x; tc_{xy} is y-th translation candidate for 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}. Next, we compute the ratio of the each candidate's score over the highest candidate's score as follows: $ScoreRatio(tc_{ij}) =$ MI score$(tc_{ij}|Q)$/MI score$(tc_{ih}|Q)$, where tc_{ih} is the candidate with highest MI score from the qt_i. If the candidate's score ratio is below the threshold τ_{MI}, the candidate will be discarded.

Here, we use the above example to illustrate the term disambiguation mechanism. For the given English term "Kosovo", the MI scores of "科索夫", "科索伏", and "科索沃" are computed; "科索伏" achieves the highest score, while the score ratio of the other two candidates are much lower than the threshold. Thus, only "科索伏" is treated as Kosovo's translation and used to build the final Chinese query to perform the IR.

4.3 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 (Robertson et al., 1996) is used to score

a retrieved document's relevance. In addition, we employ the following document re-ranking function (Yang et al., 2007):

$$\sqrt{\frac{(\sum_{i=1}^{K} df(t,d_i) \times f(i))/K}{DF(t,C)/R}} \times \sqrt{|t|}$$

$$df(t,d_i) = \begin{cases} 1 & t \in d_i \\ 0 & t \notin d_i \end{cases},$$

where d_i is the ith 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, $f(i) = \frac{1}{sqrt(i)}$.

5 Evaluation and Analysis

To evaluate our KCIR system, we use the topic and document collections of the NTCIR-5 CLIR tasks (Kishida et al., 2005). 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 use 50 topics provided by NTCIR-5 and use the title field as the input query because it is similar to queries input to search engines.

We construct five runs as follows:

- **Baseline**: using a Korean-Chinese dictionary-based translation.
- **Baseline+Extended Dictionaries only**: the baseline system plus the extended dictionaries translation.
- **Baseline+NET Methods**: the baseline system plus our NET methods, namely, Wikipedia, Naver people search, and the pattern-based method.
- **Google Translation**: using the Google translation tool.
- **Chinese monolingual**: using the Chinese versions of the topics given by NTCIR.

We use the Mean Average Precision (MAP) and Recall (Saracevic et al., 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.

Table 1: Evaluation Results

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
Baseline	0.0553	0.0611	0.2202	0.2141
Baseline+extended dictionaries	0.1573	0.1751	0.5706	0.5489
Baseline+NET	**0.2576**	**0.2946**	**0.7255**	**0.7103**
Google translation	0.1340	0.1521	0.5254	0.5149
Chinese mono	0.2622	0.3019	0.7705	0.7452

Table 1 shows that our method improves KCIR substantially. Our method's performance is about five times better than that of the baseline system and very close to that of Chinese monolingual IR. Wikipedia translation improves the performance, but not markedly because Wikipedia cannot cover some NEs. Google translation is not very satisfactory either, since many NEs cannot be translated correctly.

To evaluate our NE translation method, we create two additional datasets. The first dataset contains all the 30 topics with NEs in NTCIR-5. To further investigate the effectiveness of our method for queries containing person names, which are the most frequent NEs, we construct a second dataset containing 16 topics with person names in NTCIR-5. We compare the performance of our method on KCIR with that of Chinese monolingual IR on these two datasets. The results are shown in Tables 2 and 3.

5.1 Effectiveness of Extended Dict

We adopt two online dictionaries to extend our bilingual dictionaries: Wikipedia and Naver people search engine. Wikipedia is an effective tool for translating well-known NEs. In the test topics, NEs like "김정일"(Kim Jong-il, North Korea's leader), "탈리반"(Taliban), "해리포터"(Harry Potter) and "한나라당"(Great National Party in South Korea) are all translated correctly by Wikipedia.

We observe that the most difficult cases in Korean-Chinese person name translation, especially Japanese and 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 the U.S.) and "이치로"(Ichiro Suzuki, a Japanese baseball player). The major advantage of the Naver people search engine is it can can provide the original names written in Chinese characters.

According to our evaluation, the extended dictionaries improve the IR performance of the baseline system about threefold. It shows that the extended dictionaries can translate part of Korean NEs into Chinese. However, there are still many NEs that the extended dictionaries cannot cover.

5.2 Effectiveness of Patterns

In our method, we employ automatically learned patterns to extract translations for the remaining NEs not covered by the offline or online dictionaries. For example, we can extract Chinese translations for "오키나와"(Okinawa, in Japan) by using K-C translation patterns. Most non-CJK NEs can be translated correctly by using the K-E translation patterns. For example, "제니퍼 카프리아티"(Jennifer Capriati), "탄저"(anthrax), and "광우병"(mad cow disease) can be extracted from Google snippets effectively by our translation patterns.

Although our method translates some NEs into English first and then into Chinese in an indirect manner, it is very effective because the non-CJK NEs in Korean are mainly from English. In fact, 16 of the 17 NEs can be successfully translated by the two stage translation method that employs two types of translation patterns: K-E and E-C.

5.3 Effectiveness Analysis of NET

As shown in Table 2, for topics with NEs, the rigid MAP of our method is very close to that of Chinese monolingual IR, while the relax MAP of our method is even better than that of Chinese monolingual IR. We observe that 26 of the 31 NEs in the topics are successfully translated into Chinese. These results demonstrate that our hybrid method comprising the extended dictionaries and translation patterns can deal with Korean-Chinese NE translation effectively and improve the performance of IR substantially.

Table 2: Results on Topics with NEs

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
NET	0.2700	0.3385	0.7565	0.7578
Chinese	0.2746	0.3273	0.7922	0.7846

Note that, our method can extract more possible Chinese translations, which is similar to query expansion. For non-CJK NEs, there may exist several Chinese transliterations that are actually used in Chinese, especially for the person names. Take "Tito"for example; its six common Chinese transliterations, namely, "迪托"(di-tuo), "蒂托"(di-tuo), "帝托"(di-tuo), "提托"(ti-tuo), and "狄托"(di-tuo) can be extracted. With 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). This is the reason that our method outperforms the Chinese monolingual run in topics with NEs.

5.4 Error Analysis

NEs that cannot be translated correctly can be divided into two categories. The first contains names not selected as NE candidates. The Japanese person name "후지모리" (Alberto Fujimori, Peru's ex-president) is in this category. For the name "후지모리" (Fujimori), the first two characters "후지" (hind legs) and the last two characters "모리" (profiting) are all Sino-Korean words, so it is regarded as a compound word, not an NE. The other category contains names with few relevant web pages, like the non-CJK names "안토니오 토디" (Antonio Toddy).

The other problem is that our method can translate the Korean NEs into correct Chinese translations, but not the translation used in the CIRB 4.0 news collection. For example, "쿠르스크" (Kursk) is translated into "庫爾斯克" (Ku-er-si-ke) correctly, but only the transliteration "科斯克" (Ke-si-ke) is used in CIRB 4.0. In this situation, the extracted translation cannot improve the performance of the KCIR.

Table 3: Results on Topics with Person Names

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
NET	0.2730	0.3274	0.7146	0.7299
Chinese	0.2575	0.3169	0.7513	0.7708

6 Conclusion

In this paper, we have considered the difficulties that arise in translating NEs from Korean to Chinese for IR. We propose a hybrid method for K-C NET that exploits an extended dictionarie containing Wikipedia and the Naver people search engine, combined with the translation patterns automatically learned from the search results of the Google search engine. To evaluate our method, we use the topics and document collection of the NTCIR-5 CLIR task. Our method's performance on KCIR is over five times better than that of the baseline configuration with only an offline dictionary-based translation module. Moreover, its overall MAP score is up to 0.2986, and its MAP on the NE topics is up to 0.3385 which is even better than that of the Chinese monolingual IR system. The proposed method can translate NEs that originated in the Chinese, Japanese, Korean, and non-CJK languages and improve the performance of KCIR substantially. Our NET method is not language-specific; therefore, it can be applied to the other CLIR systems beside K-C IR.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 400–408.

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Cbung Tsai. 1998. Proper name translation in cross-language information retrieval. *Proceedings of 17th COLING and 36th ACL*, pages 232–236.

Kazuaki Kishida, Kuang hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng. 2005. Overview of clir task at the fifth ntcir workshop. *Proceedings of the Fifth NTCIR Workshop*.

Juhee Lee. 2003. Loadword phonology revisted: Implications of richness of the base for the analysis of loanwords input. *Explorations in Korean Language and Linguistics*, pages 361–375.

Mira Oh. 2003. English fricatives in loanword adaption. *Explorations in Korean Language and Linguistics*, pages 471–487.

S.E. Robertson, S. Walker, MM Beaulieu, M. Gatford, and A. Payne. 1996. Okapi at trec-4. *Proceedings of the Fourth Text Retrieval Conference*, pages 73–97.

Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison. 1988. A study of information seeking and retrieving. *Journal of the American Society for Information Science*, 39(3):161–176.

L. Yang, D. Ji, and M. Leong. 2007. Document reranking by term distribution and maximal marginal relevance for chinese information retrieval. *Information Processing and Management: an International Journal*, 43(2):315–326.