Japanese-Chinese Information Retrieval with An Iterative Weight-tuning Scheme

Chu-Cheng Lin[ab]
chu.cheng.lin@gmail.com

Yu-Chun Wang[ac]
albyu@iis.sinica.edu.tw

Richard Tzong-Han Tsai[d*]
thtsai@saturn.yzu.edu.tw

Wen-Lian Hsu[a]
hsu@iis.sinica.edu.tw

[a]Institute of Informtaion Science, Academia Sinica, Taiwan
[b]Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan
[c]Dept. of Electrical Engineering, National Taiwan Univeristy, Taiwan
[d]Dept. of Computer Science and Engineering, Yuan Ze University, Taiwan
[*]corresponding author

Abstract

This paper describes our Japanese-Chinese cross language information retrieval system. We adopt "query-translation" approach and employ both a conventional Japanese-Chinese bilingual dictionary and Wikipedia to translate query terms. We propose that Wikipedia can be regarded as a good dictionary for named entity translation. According to the nature of Japanese writing system, we propose that query terms should be processed differently based on their written forms. We use an iterative method for weight-tuning and term disambiguation, which is based on the PageRank algorithm. When evaluating on the NTCIR-5 test set, our system acheives as high as 0.2217 and 0.2276 in relax MAP (Mean Average Precision) measurement of T-runs and D-runs.

Keywords: Japanese-Chinese cross language information retrieval, query disambiguation, iterative term weighting

1 Introduction

Cross-Language Information Retrieval (CLIR) has become an active research area in recent years, because it helps people overcome language barriers to retrieve information written in other languages by using their own languages. CLIR for European languages has been studied for several decades and achieved satisfying results. However, CLIR for East Asian languages has not been studied extensively and a number of problems must still be resolved.

The cultures of China, Japan and Taiwan have been interwinded for several centuries, and culture exchange between China and Japan have become warmer recently. Besides, because of its economic growth, China has become Japan's largest trading partner. Therefore, the demand for Japanese-Chinese CLIR systems has grown.

Even though the origin of Japanese is unclear [10], Chinese and Japanese differ markedly in many linguistic features, such as grammar and phonology. However, in spite of these distinctions, Japanese has adopted a lot

of Chinese vocabulary, and Chinese characters (Kanji) in their writing system.

There are two main CLIR approaches for translation: one translates query terms into the language of the documents; the other translates all the documents into the language of the queries. Translating the query terms is more practical, since the entire collection of documents may be very large, and it can not be updated regularly. However, with the query-translation approach, ambiguity is a serious problem. Words may have several different meanings, as this is the nature of natural languages.

Named entity (NE) recognition is also an issue. Bilingual dictionaries often have few entries for NEs. Moreover, if NEs are wrongly segmented as ordinary words and translated with a bilingual dictionary, the result will be poor.

We therefore propose a Japanese-Chinese information retrieval system, in which the IR performance is improved substantially by exploiting the nature of Japanese vocabulary and the Japanese writing system.

2 Related Works

Hasan et al. [6] were the first to exploit the high co-occurrence of Kanji in Chinese and Japanese texts for CLIR. However the ambiguity problem was not addressed.

Though not directly related to Japanese-English CLIR, Buckley et al. [2] took a very interesting approach in French-English CLIR. They did not take the dictionary approach; instead, they treated English words as misspelled French words, and then used lexicographically close French words for monolingual runs. The concept is very similar to our approach.

In NTCIR proceedings, some researchers have attempted Japanese-Chinese CLIR. Nakagawa et al. [12] took English as the pivot language, since they could not find satisfactory Japanese-Chinese language resources. Gey [5] took a "no-translation" approach, using only Chinese characters in Japanese-Chinese and Chinese-Japanese CLIR. The results are mixed, for some topics the performance is good, while on some the performance is rather poor.

Quite a lot of attention has been paid to NE translation, an important subproblem of CLIR. Cheng et al. [3] used web corpora for automatic NE translation. Their paper took snippets from the search engine, and obtained translations using n-grams. This is not feasible with Japanese-Chinese IR because, while there are many texts mixed with Chinese and English, those mixed with Japanese and Chinese are much more rare. Lee et al. [9] used the EM method to automatically extract NEs and their translations from English-Chinese parallel corpora. Kuo et al. [8] experiemented with supervised and unsupervised learning in extracting English-Chinese NEs from web corpora. Wu et al. [16] used the initial and final syllables of English-Chinese translation pairs for automatic extraction.

3 Nature of Japanese Writing System

Japanese vocabulary has adopted many Chinese words. The Chinese characters have also been incoporated into the writing system as *Kanji*. Chinese has thus exerted an everlasting impact on Japanese, in speech, and more profoundly, in writing. [15]

The Japanese writing system consists of *Kanji* characters, and *kana* syllabaries, namely *Hiragana* and *Katakana*. [14]

3.1 Kanji

Kanji refers to Chinese characters used in Japanese. Since the postwar era, the usage of Kanji has been limited, but it is still common in Japanese. Most Japanese people's names are written in Kanji; and many Japanese places, organizations, and many other entities have Kanji names.

Many Kanji characters were borrowed from Middle Chinese. Kanji has two kinds

of pronunciation in Japanese. One is from Chinese, which is called Sino-Japanese; the other is *Kunyomi*. Kunyomi associates a native Japanese morpheme with a Chinese character that has a close meaning to the native Japanese morpheme. For example, the Chinese character 山 has two types of pronunciation: *san* (Sino-Japanese) and *yama* (Kunyomi).

Besides, a Kanji character may have different pronunciation due to the usage in different periods of history and the varied usages of Kunyomi. Despite its complexity, pronunciation of Kanji can be looked up in a dictionary.

広大に無限に広がる宇宙
在廣闊無垠的宇宙中

Figure 1: Bilingual text

We observe that the written forms and meanings of most Kanji words are close to those in modern Chinese, as shown in Figure 1. The reason behind this phenomenon is that Japanese has adopted a lot of Chinese vocabulary and Chinese also adopted a lot of Japanese-made Kanji words back during 19th and 20th centuries.

3.2 Hiragana

Hiragana, which originated from the cursive writing of Chinese characters, serve to represent agglutinative affixes, functional words, and many adverbs. Since Hiragana usually serve grammatical roles, and sometimes are used to write adverbs, we do no use any special method to process Hiragana. Agglutinative affixes and adverbs can be recognized easily by the word segmenter and translated by a dictionary.

3.3 Katakana

The Katakana syllabary in modern Japanese is mainly used to transliterate loanwords from foreign languages, such as English or Chinese. Besides direct transliteration, some words are truncated before transliteration. This problem has discouraged the development of a transliteration-mapping approach based on phonetic similarities.

Katakana transliterations and their equivalent Chinese transliterations of foreign terms have little correlation. In practice, the translations can be found only by a dictionary-based translation approach.

4 System Description

We construct a Japanese-Chinese cross-language information retrieval system incoporating the knowledge mentioned in Section 3. The workflow is depicted in Figure 2. We describe the components as follows.

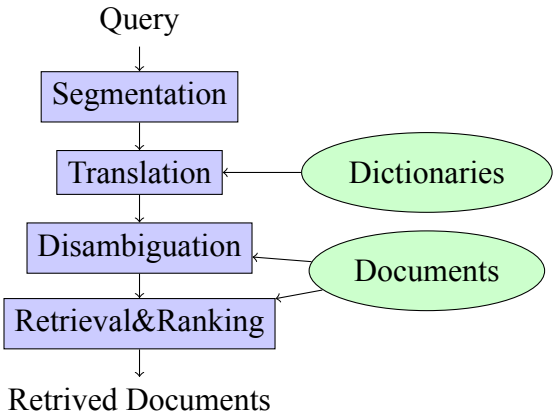

Figure 2: The workflow

4.1 Word Segmentation

Our system accepts both natural language sentences and keywords as queries. We use MeCab[1] for Japanese word segmentation. We prepare a list of stop words, which are ignored if they occur in queries.

MeCab has the capability of tagging parts-of-speech (POS). Words tagged as nouns are then extracted for further translation. Because of the dictionary that MeCab uses for training, MeCab may treat some morphemes as suffixes and separate the suffix from a noun. However, sometimes nouns without

[1] http://mecab.sourceforge.net/

suffixes cannot match the correct entries of the Japanese-Chinese dictionary that we use. Therefore, in addition to the words tagged as nouns by MeCab, we also glue the word and the suffix behind it together for further translation.

Additionally, *Katakana* words are likely to be transliterations of foreign named entities. They are recognized from the query using regular expressions and then separated by interpuncts (dots in the middle used to seperate words) to generate all the possible combinations.

4.2 Translation

The terms segmented in Section 4.1 are looked up in dictionaries. Dictionary-based translation is effective and simple to implement. We adopt two dictionaries: the Sanseido Japanese-Chinese dictionary and Wikipedia, a free online encyclopedia.

The Sanseido Japanese-Chinese dictionary[2] is used for word-to-word translations. It accepts Japanese words as input and provides Chinese meanings. The dictionary is for Japanese learners learning simplified Chinese. Thus, our system converts the simplified Chinese characters to the corresponding traditional Chinese ones. All the translations from the dictionary are regarded as the translation candidates.

The Sanseido dictionary does not contain enough NEs, and is somewhat out-of-date. To resolve the problem, we use Wikipedia as an alternative dictionary. Many articles in Wikipedia may have inter-language links to the editions describing the same topics but written in other languages. We exploit the titles of articles the inter-language links point to as translations. The terms are submitted as a query to the Japanese Wikipedia. If Wikipedia has a matched article, our system checks whether there is an inter-language link to the Chinese edition. If there is, the title of its corresponding Chinese is taken as a translation candidate. If there is no link to Chinese, but one to English exists, the English title is used instead. Although these terms are kept in English, they still improve IR performance, as it is not uncommon for Chinese texts to contain foreign terms written in Latin alphabets.

As discussed in Section 3.1, if a term is entirely in Kanji, our system converts the Kanji characters into their corresponding traditional Chinese forms. If a term is not written in Kanji, but has Kanji forms provided by the dictionaries, the Kanji forms will also be used. Then, all these terms are gathered as translation candidates.

4.3 Query Disambiguation

A Japanese Hiragana or Katakana word may refer to different Kanji words, as discussed in Section 3. Even a Kanji word, or a word composed entirely in Hiragana or Katakana, may have several meanings. It is common for a term to have up to a dozen of different translations. The translation component in Section 4.2 gathers all possible translation candidates for a query term, but many of them will not be correct. The ambiguity problem degrades the IR performance seriously. Therefore, we must adopt a disambiguation scheme to solve the problem.

We follow the iterative method described in [11], which is basically the PageRank algorithm [1]. Consider two Japanese terms, J_i and J_k and their respective Chinese translation candidates $C_{i,1}\cdots C_{i,n_i}$ and $C_{k,1}\cdots C_{k,n_k}$. It is reasonable that suitable translations of J_i and J_k would have a high co-occurrence frequency. Thus one may choose $C_{i,a}$ and $C_{k,b}$ as translations of J_i and J_k that

$$C_{i,a}, C_{k,b} = \arg\max_{x,y} \frac{freq(C_{i,x}, C_{k,y})}{freq(C_{i,x})freq(C_{k,y})}, \quad (1)$$

where $x = 1\cdots n_i$ and $y = 1\cdots n_k$.

However, this trivial pairwise approach presents a serious problem. First, there might be no translations co-occurring among translations of J_i and J_k. Though J_i and J_k may both have co-occurring translation pairs with J_ℓ;

[2] http://www.excite.co.jp/dictionary/japanese_chinese

yet still the problem of inconsistency persists. For example, assume that $C_{i,1}$ and $C_{k,1}$ have the highest co-occurrence frequency among all translations of J_i and J_k, $C_{k,2}$ and $C_{\ell,3}$ between J_k and J_ℓ, and no co-occurrences between any translations of J_i and J_ℓ. One can not justify that $C_{i,1}$ is the best match for $C_{\ell,3}$, as shown in Figure 3. We need to *propagate* the suitability of translations not directly connected through their links to other candidates.

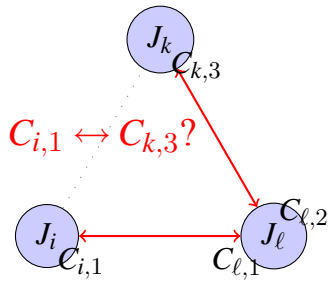

Figure 3: No co-occurrence between J_i and J_k

We adopt an iterative approach. For each translation candidate $C_{i,t}$ of a Japanese term J_i, we assign a weight $W_{i,t}$.

Initially, for every candidate $C_{i,t}$ of J_i

$$W_{i,t}^0 = \frac{1}{n_i}, \tag{2}$$

and n_i is the count of translation candidates of J_i.

After iteration step k,

$$W_{i,t}^k = W_{i,t}^{k-1} + \sum_{\ell \in Q} W_\ell^{k-1} \cdot link(C_{i,t}, \ell), \tag{3}$$

where $Q = \{$ all translation candidates except those of $J_i\}$, and *link* is a weight-link scoring function. The pairwise scoring method in Equation 1 can be used as the *link* function. We use an alternative scoring function in our system, which is a likelihood ratio test between two hypotheses, H_1 and H_2:

$$\begin{array}{lccccc} H_1: & p(t_2|t_1) & = & p & = & p(t_2|\neg t_1) \\ H_2: & p(t_2|t_1) & = & p_1 \neq p_2 & = & p(t_2|\neg t_1). \end{array} \tag{4}$$

H_1 states that the occurrence of the two terms t_2 and t_1 are independent, and the probability of t_2 co-occurring with t_1 is the same as that of t_2 occurring without t_1. In contrast, H_2 states that the probability of t_2 co-occurring with t_1 is not the same as t_2 occurring without t_1.

The number of all documents is N; the number of documents containing term t_1 is n_{t_1}; the number of documents containing term t_2 is n_{t_2}; the number of documents containing both t_1 and t_2 is $n_{t2 \cap t_1}$, and the number of documents containing t_2, but not containing t_1, is $n_{t_2 \cap \neg t_1}$. With the log-likelihood defined by Dunning [4],

$$H(p,k,n) = p^k (1-p)^{n-k},$$

and

$$\begin{array}{rcl} L(H_1) & = & H(P_{t_2}, n_{t_2}, N), \\ L(H_2) & = & H(P_{t_2|t_1}, n_{t_2 \cap t_1}, n_{t_1}) \cdot \\ & & H(P_{t_2|\neg t_1}, n_{t_2 \cap \neg t_1}, N - n_{t_1}). \end{array} \tag{5}$$

We have

$$\begin{array}{rcl} -\log\lambda & = & -\log \dfrac{L(H_1)}{L(H_2)}, \\ & = & \log L(H_2) - \log L(H_1). \end{array} \tag{6}$$

We use $-\log\lambda$ as the weight-link scoring function.

Using either Equation 1 or Equation 6 as the *link* function, the weight of each translation candidate is updated per iteration. We repeat the iterations until for some translation t that $W_t^k - W_t^{k-1} < \delta$, where δ is a threshold.

4.4 Document Retrieval and Reranking

Our system uses the Lucene information retrieval engine for document indexing and retrieval. The Okapi BM25 function is used to measure the relevance of documents. [13]

We filter out translation terms whose weights are too low before submitting them to the Lucene engine. Lucene returns the highest scoring 1000 documents. We then employ the following document reranking function: [17]

$$\sqrt{\frac{(\sum_{i=1}^{K} df(t,d_i) \cdot f(i))/K}{DF(t,C)/R}} \cdot \sqrt{|\,t\,|} \cdot W_t, \tag{7}$$

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

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 term t in C; $|\,t\,|$ is the length of t; $f(i) = \frac{1}{\sqrt{i}}$; and W_t is the weight of t, as calculated in Section 4.3.

5 Evaluation

To evaluate of our Japanese-Chinese information retrieval system, we use the topics and document collection of the NCTIR-5 CLIR tasks. The document collection is the Chinese Information Retrieval Benchmark (CIRB) 4.0, containing news articles from four Taiwanese newspapers published from 2000 to 2001. NTCIR-5 provides 50 topics, each of which contains four fields: title, description, narration, and concentrate words. Our configuration is the same as the original NTCIR-5 CLIR task. The T-runs take the title field of each topic as a query. The D-runs use the description field.

For comparison, we include the results of the NTCIR-5 CLIR Japanese-Chinese task. The group names of "OKI" and "BRKLY" refer to the Oki Eletric Industry and Berkeley Text Retrieval Research Group respectively. In the overview of the NTCIR-5 CLIR task [7], only the rigid Mean Average Precision (MAP) of D-run is provided. Gey [5], however, provided the best MAP of T-runs in his report. The evaluation results are shown in Tables 1 and 2.

We construct four T-runs and four D-runs with following set-ups:

Baseline Using only the Japanese-Chinese bilingual dictionary for translation.

B + W Besides the bilingual dictionary, Wikipedia is also used for translations.

B + K Besides using the bilingual dictionary, Kanji words are treated as translation candidates.

B + K + W Besides using the bilingual dictionary and Wikipedia for translations, Kanji words are treated as translation candidates.

In every run, necessary substitutions of Japanese and Simplified Chinese-exclusive Kanji into corresponding Traditional Chinese characters are executed. Also, in every run we use the disambiguating scheme introduced in Section 4.3.

The Mean Average Precision (MAP) and Recall are used to evaluate the performance. NTCIR provides two kinds of relevance judgments: Rigid and Relax. A document is rigid-relevant if it is highly relevant to the topic; if it is highly relevant or partially relevant to the topic, then it is relax-relevant.

The results are shown in Tables 1 and 2. The relative improvements are shown in parentheses.

6 Discussion

6.1 Effect of Treating Kanji as Translation

Our observation in Section 3 indicates that Kanji words in the original query can be used for translation directly. The results in Tables 1 and 2 support this observation: the **B + K** run shows significant improvements in MAP, and moderate improvements in Recall. In the T-run of topic 22, where the Title field has only one word:

狂牛病

and topic 28

ブブカ，鳥人，引退

the bilingual dictionary failed to find any translations for the query terms. However, the Kanji approach still managed to retrieve relevant documents, and with good results in topic 22, achieving 0.2816 and 0.4641 in AP rigid and relax-relevance, respectively.

Table 1: Evaluation Results: T-runs

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
Baseline	0.1198	0.1240	0.4978	0.4666
B + W	0.1747 (45.8%)	0.1956 (57.8%)	0.6330 (27.2%)	0.6100 (30.8%)
B + K	0.1562 (30.4%)	0.1709 (37.8%)	0.5392 (8.3%)	0.5094 (9.2%)
B + K + W	**0.1952 (62.9%)**	**0.2217 (78.9%)**	**0.6818 (37.0%)**	**0.6652 (42.6%)**
BRKLY	0.0925			

Table 2: Evaluation Results: D-runs

Run	MAP		Recall	
	Rigid	Relax	Rigid	Relax
Baseline	0.1129	0.1275	0.5291	0.4982
B + W	0.1688 (49.5%)	0.2049 (60.7%)	0.6602 (24.8%)	0.6373 (27.9%)
B + K	0.1402 (24.2%)	0.1674 (31.3%)	0.5930 (12.1%)	0.5576 (11.9%)
B + K + W	**0.1866 (65.2%)**	**0.2276 (78.6%)**	**0.7245 (36.9%)**	**0.6963 (39.7%)**
BRKLY	0.1568			
OKI	0.0779			

6.2 Performance of Dictionaries

We have two dictionaries in our system, the Sanseido Japanese-Chinese bilingual dictionary and Wikipedia. The bilingual dictionary is used for Baseline set-up. We propose that Wikipedia be used for NEs, especially those written in Katakana. The results are impressive. Compared to Baseline, with improvements of approximately 60% and 80%, respectively for Rigid and Relax MAP evaluation; and about 40% in Recall evaluations of both. Wikipedia has successfully translated many NEs, such as "コソボ" (Kosovo), "グリーンスパン" (Allen Greenspan), and "東ティモール" (East Timor). When encountering a Kanji term, Wikipedia sometimes translates it into an identical Chinese term — examples are "金大中" (former South Korean president), "韓国" (South Korea), which is similar to using the Kanji approach. Yet using Wikipedia cannot be regarded as an improved Kanji-only approach; in the evaluation, some Kanji terms are not recognized by Wikipedia. Moreover, some translations given by Wikipedia yield better results than original Kanji do, while some do not.

6.3 Effectiveness of Disambiguation

The disambiguation scheme described in Section 4.3 filters out unsuitable translation candidates found in dictionaries.

For example, the Title field of topic 4 of NTCIR-5 is:

> 米国防長官，ウィリアム・セバスチャン・コーエン，北京

Without disambiguation, the result would be

> 美國 大米 稻米 米國 米 稻 威廉 Sebastian 北京 北京市 國防 National security 武備 長官

The Japanese abbreviation of United States is "米"; however, it also means "rice" if it is written in Kanji. The dictionaries respond with both the correct translation "美國" and the dubious ones "大米 稻米 米 稻."

The disambiguated result is

美國 威廉 Sebastian 北京 國防 長官

IR performance is greatly improved. The AP of this topic increased from 0.0072 to 0.0235 in rigid-relevance, and 0.0084 to 0.0249 in relax-relevance.

6.4 Error Analysis

There are cases where dictionaries respond with several meanings, of which some are desirable, but they are filtered out by disambiguation.

For example, topic 20 of NTCIR-5 is

性転換，カエル，魚

The Katakana “カエル” is the transliteration of English “frog,” but in Japanese this indigenous morpheme “カエル” means “to change; to substitute; to transform,” and can be written in various Kanji forms, such as “換,” “代,” and “変.” Besides, “カエル” is also a Sino-Japanese reading of “孵” (to hatch.)

The translation candidates of “カエル” consist of

蛙 青蛙 田雞 換 倒 調動 回來 回去 歸回 重返 替代 變 改 變更 改變 轉換 轉移 孵 孵化 無尾目;

and the disambiguated result is

改變^1 轉移^0.1040576 轉變^1 轉換^0.1331546

The number following the hat denotes the term’s weight, ranging from 0 to 1. Clearly, the correct translations related to frogs are all stripped. Thus, the outcome is very poor.

Another problem arises with transliterations. Some NEs, especially person names, which would traditionally be written in Kanji, are transcribed in Katakana in the NTCIR-5 test topics. For example, topic 11 is about the Japanese baseball player Ichiro Suzuki (鈴木一朗). The Title field of this topic is

イチロー，新人王，大リーグ

“イチロー” is “Ichiro,” the player’s first name, written in Katakana.

And topic 7:

ウェン・ホー・リー，機密情報，国家安全保障

“ウェン・ホー・リー” is the Taiwanese-born American scientist 李文和.

We managed to get “鈴木一朗” as a translation for “イチロー” using Wikipedia in topic 11, but we could not find appropiate translations for “ウェン・ホー・リー” in topic 7. This kind of failure indicates that using Wikipedia cannot solve the NE problem totally.

7 Conclusion

We have described the construction of a Japanese-Chinese cross-language information retrieval system, and adopted the “query-translation” approach to the Japanese-Chinese CLIR problem. For query translation, we have used the Sanseido Japanese-Chinese bilingual dictionary. We also have exploited Wikipedia for translations.

We exploited the nature of Japanese vocabulary and the Japanese writing system for better translations. Using Kanji for translation yields significant improvements in our evaluation. The results of the evaluation confirm that foreign terms are widely transcribed in Katakana.

To cope with ambiguity, we have adopted an iterative disambiguating scheme. The current implementation of this scheme, which uses the likelihood function as its weight function, proved to be effective in the evaluation.

Our system has acheived MAP as high as 0.2276, and outperforms the previous NTCIR-5 CLIR Japanese-Chinese T-runs’ best rigid MAP by 111%, and D-runs’ by 19%.

References

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998. [Online]. Available: citeseer.ist.psu.edu/brin98anatomy.html

[2] C. Buckley, M. Mitra, J. Walz, and C. Cardiey, "Using clustering and SuperConcepts within SMART : TREC 6," in *The Sixth Text Retrieval Conference (TREC-6), NIST Special publication 500-240*, E. M. Voorhees and D. K. Harman, Eds. Department of Commerce, National Institute of Standards and Technology, 2000, pp. 107–124.

[3] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien, "Translating unknown queries with web corpora for cross-language information retrieval," in *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2004, pp. 146–153.

[4] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Comput. Linguist.*, vol. 19, no. 1, pp. 61–74, 1993.

[5] F. C. Gey, "How similar are Chinese and Japanese for cross-language information retrieval?" in *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, 2005.

[6] M. M. Hasan and Y. Matsumoto, "Chinese-Japanese cross language information retrieval: a Han character based approach," in *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*. Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 19–26.

[7] K. Kishida, K.-h. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng, "Overview of CLIR task at the fifth NTCIR workshop," *Proceedings of the Fifth NTCIR Workshop*, 2005.

[8] J.-S. Kuo, H. Li, and Y.-K. Yang, "Learning transliteration lexicons from the web," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 1129–1136. [Online]. Available: http://www.aclweb.org/anthology/P/P06/P06-1142

[9] C.-J. Lee and J. S. Chang, "Acquisition of English-Chinese transliterated word pairs from parallel-aligned texts using a statistical machine transliteration model," in *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 96–103.

[10] B. Lewin, "Japanese and Korean: The problems and history of a linguistic comparison," *Journal of Japanese Studies*, vol. 2, no. 2, pp. 389–412, 1976. [Online]. Available: http://links.jstor.org/sici?sici=0095-6848%28197622%292%3A2%3C389%3AJAKTPA%3E2.0.CO%3B2-D

[11] C. Monz and B. J. Dorr, "Iterative translation disambiguation for cross-language information retrieval," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2005, pp. 520–527.

[12] T. Nakagawa and M. Kitamura, "NTCIR-4 CLIR experiments at Oki," in *Working Notes of the Fourth NTCIR Workshop Meeting*, 2004.

[13] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," *Proceedings of the Fourth Text Retrieval Conference*, pp. 73–97, 1996.

[14] C. Seeley, *A history of writing in Japan*. University of Hawaii Press, 2000.

[15] I. Taylor and M. M. Taylor, *Writing and literacy in Chinese, Korean, and Japanese*. Benjamins, John Publishing Company, 1995.

[16] J.-C. Wu and J. S. Chang, "Learning to find English to Chinese transliterations on the web," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 996–1004. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1106

[17] L. Yang, D. Ji, and M. Leong, "Document reranking by term distribution and maximal marginal relevance for Chinese information retrieval," *Information Processing and Management: an International Journal*, vol. 43, no. 2, pp. 315–326, 2007.