Chinese-Japanese Cross Language Information Retrieval: A Han Character Based Approach

Md Maruf HASAN
Computational Linguistic Laboratory
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma,
Nara, 630-0101 Japan
maruf-h@is.aist-nara.ac.jp

Yuji MATSUMOTO
Computational Linguistic Laboratory
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma,
Nara, 630-0101 Japan
matsu@is.aist-nara.ac.jp

Abstract

In this paper, we investigate cross language information retrieval (CLIR) for Chinese and Japanese texts utilizing the Han characters – common ideographs used in writing Chinese, Japanese and Korean (CJK) languages. The Unicode encoding scheme, which encodes the superset of Han characters, is used as a common encoding platform to deal with the multilingual collection in a uniform manner. We discuss the importance of Han character semantics in document indexing and retrieval of the ideographic languages. We also analyse the baseline results of the cross language information retrieval using the common Han characters appeared in both Chinese and Japanese texts.

Keywords: Cross Language Information Retrieval, Multilingual Information Processing, Chinese, Japanese and Korean (CJK) Languages

Introduction

After the opening of the Cross Language Information Retrieval (CLIR) track in the TREC-6 conference (TREC-1998), several reports have been published on cross language information retrieval in European languages, and sometimes, European languages along with one of the Asian languages (e.g., Chinese, Japanese or Korean). However, no report is found in cross language IR that *focuses* on the Asian languages *exclusively*. In 1999, Pergamon published a special issue of the journal, *Information Processing and Management* focusing on Information Retrieval with Asian Languages (Pergamon-1999). Among the eight papers included in that special issue, only *one* paper addressed CLIR (Kim et al., 1999). Kim et al. reported on multiple Asian language information retrieval (English, Japanese and Korean CLIR) using multilingual dictionaries and machine translation techniques (to translate both queries and documents).

In TREC, intensive research efforts are made for the European languages, for example, English, German, French, Spanish, etc. Historically, these languages share many similar linguistic properties. However, *exclusive* focus on Asian languages, for example, Chinese, Japanese and Korean (CJK) - which also share significantly similar linguistic properties, has not been given. Enormous amount of CJK information is currently on the Internet. The combined growth rate of the CJK electronic information is also predicted to be growing at a faster rate. Cross language IR focusing on these Asian languages is therefore inevitable.

In this paper, we investigate the potential of indexing the semantically correlated Han characters appear in both Chinese and Japanese documents and queries to facilitate a cross language information retrieval. Using Han character oriented document and query vectors, within the framework of the vector space information retrieval, we then evaluate the effectiveness of the cross language IR with respect to their monolingual counterparts. We conclude with a discussion about further research possibilities and potentials of Han character oriented cross language information retrieval for the CJK languages.

1 Related Research and Motivation

Several approaches are investigated in CJK text indexing to address monolingual information retrieval (MLIR) - for example, (1) indexing **single** ideographic character, (2) indexing **n-gram**[1] ideographic characters and (3) indexing words or phrases after segmentation and morphological analysis. Monolingual information retrieval (MLIR) of CJK languages is further complicated with the fact that CJK texts do not contain word delimiters (e.g., a blank space after each word in English) to separate words. From the un-delimited sequence of characters, words must be extracted first (this process is known as *segmentation*). For inflectional ideographic language like Japanese, morphological analysis must also be performed. Sentences are segmented into words with the help of a dictionary and using some machine learning techniques. Morphological analysis also needs intensive linguistic knowledge and computer processing. Segmentation and morphological analysis are tedious tasks and the accuracy of the automatic segmentation and morphological analysis drastically vary in different domains. The word based indexing of CJK texts is therefore computationally expensive. Segmentation and morphological analysis related issues of both Chinese and Japanese are intensively addressed elsewhere (Sproat et al., 1996; Matsumoto et al., 1997 and many others).

The n-gram (**n** >1) character based indexing is computationally expensive as well. The number of indexing terms (n-grams) increases drastically as **n** increases. Moreover, not all the n-grams are semantically meaningful words; therefore, smoothing and filtering heuristics must be employed to extract linguistically meaningful n-grams for effective retrieval of information. See Nie et al. (1996, 1998, 1999), Chen et al. (1997), Fujii et al. (1993), Kim et al. (1999) for details.

In contrast, indexing single characters is straightforward and less demanding in terms of both space and time. In single character indexing, there is no need to (1) maintain a multilingual dictionary or thesaurus of words, (2) to extract word and morphemes, and (3) to employ machine learning and smoothing to prune the less important n-grams or ambiguity resolution in word segmentation (Kwok, 1997; Ogawa et al., 1997; Lee et al., 1999; etc.). Moreover, a CLIR system, based on Han character semantics, incurs no translation overhead for both queries and documents. In a single character based CLIR approach for CJK languages, some of the CLIR related problems discussed in (Grefenstette, 1998) can also be circumvented.

Comparison of experimental results in monolingual IR using single character indexing, n-gram character indexing and (segmented) word indexing in Chinese information retrieval is reported in Nie et al. (1996, 1998, 1999) and Kwok (1997). For the case of *monolingual* information retrieval (MLIR) task, in comparison to the single character based indexing approach, n-gram based and word based approaches obtained better retrieval at the cost of the extra time and space complexity. Similar comparison and conclusion for Japanese and Korean MLIR are made in Fujii et al. (1993) and Lee et al. (1999), respectively.

Cross language information retrieval (CLIR, Oard and Dorr, 1996) refers to the retrieval when the query and the document collection are in different languages. Unlike MLIR, in cross language information retrieval, a great deal of efforts is allocated in maintaining the multilingual dictionary and thesaurus, and translating the queries and documents, and so on. There are other approaches to CLIR where techniques like latent semantic indexing (LSI) are used to automatically establish associations between queries and documents independent of language differences (Rehder et al., 1998).

Due to the special nature (ideographic, un-delimited, etc.) of the CJK languages, the cross language information retrieval of these languages is extremely complicated. Probably, this is the reason why only a few reports are available so far in Cross Asian Language Information Retrieval (**CALIR**).

---

[1] In this paper, we use the term, *n-gram* to refer to (**n** >1) cases. When **n** =1, we use the term, *single character indexing*.

Tan and Nagao (1995) used correlated Han characters to align Japanese-Chinese bilingual texts. According to them, *the occurrence of common Han characters (in Japanese and Chinese language texts) sometimes is so prevalent that even a monolingual reader could perform a partial alignment of the bilingual texts.*

One of the authors of this paper is not a native speaker of Chinese or Japanese but has the intermediate level proficiency in both languages *now*. However, before learning Japanese, based on the familiar Han characters (their visual similarity and therefore, the semantic relation) appeared in the Japanese texts, the author could roughly comprehend the *theme* of the articles written in Japanese. This is due to the fact that unlike Latin alphabets, Han characters capture significant semantic information in them. Since document retrieval is inherently a task of semantic distinction between queries and documents, Han character based CLIR approach can therefore be justified. It is worthy to mention here that the pronunciation of the Han characters varies significantly across the CJK languages, but the visual appearance of the Han characters in written texts (across CJK language) retains certain level of similarity.

As discussed above, we can make use of the non-trivial semantic information encoded within the ideographic characters to find associations between queries and documents across the languages and perform cross language information retrieval. By doing so, we can avoid complicated segmentation or morphological analysis process. At the same time, multilingual dictionary and thesaurus lookup, and query-documents translations can also be circumvented.

In our research, we index single Han characters (common and/or semantically related) appeared in both Japanese and Chinese texts to model a new simplistic CLIR for Japanese and Chinese cross language information retrieval. CJK languages use a significant number of common (or similar) Han characters in writing. Although some ambiguities[2] exist in the usage of Han characters across the languages, there are obvious contextual and semantic associations in the usage of Han characters in the written texts across the CJK languages (Tan and Nagao, 1995).

2 Encoding scenarios of CJK languages

Character encoding schemes of CJK languages have several variations (e.g., Chinese: GB and BIG-5, etc.; Japanese: JIS, EUC, etc.)[3]. The number of Han characters encoded under a particular encoding scheme also varies. However, due to the continuous acceptance and popularity of the Unicode (Unicode-2000) by the computer industry, we have a way to investigate these languages comprehensively. The *Common CJK Ideograph* section of the Unicode encoding scheme includes all characters encoded in each individual language and encoding scheme. Unicode version 3.0 assigned codes to 27,484 Han characters, a superset of characters encoded in other popular standards.

Figure 1: Different ideographs represent the same concept, *sword*

However, Unicode encoding is not a linguistically based encoding scheme; it is rather an initiative to cope with the variants of different local standards. A critical analysis of Unicode and a proposal of Multicode can be found in Mudawwar (1997). Unicode standard avoids duplicate encoding of the same character; for example, the character '**a**' is encoded only once although it is being used in several western languages. However, for ideographic characters, such efforts failed to a certain extent due to the variation of typeface used under different situations and cultures. The characters in Figure 1, although they represent the same word (*sword* in English), is given a unique code under Unicode encoding scheme to satisfy the *round-*

[2] Ambiguities also exist in word or phrase level.

[3] A typical Internet search engine (like Yahoo) sometimes asks users to specify not only the language but also the encoding scheme (e.g., simplified (GB) or traditional Chinese (BIG-5)) for a single language search.

trip criteria[4], that is, to allow round-trip conversion between the source standard (in this case, JIS) and the Unicode. The 27,484 Han characters encoded in Unicode, therefore, includes semantic redundancy in both single-language and multiple-language perspectives.

In the unified CJK ideograph section, Unicode maintains redundancy to accommodate typographical or cultural compatibility because the design goal of Unicode is mainly to attain compatibility with the existing corporate and national encoding standards. In a Han character based CLIR approach, such redundancy and multiplicity must be identified and resolved to achieve semantic uniformity and association. Such multiplicity resolution tasks, with compare to maintaining multilingual (word) dictionaries, are less painstaking. In our Han character based CLIR, we use a table lookup mapping approach to resolve semantic ambiguities of the Han characters and associate the semantically related ideographs within and across CJK languages, as a preprocessing task.

3 Comparative analysis of Japanese and Chinese language for Han character based CLIR

Chinese text is written homogeneously using only Han characters. There are no word delimiters and therefore, segmentation must be performed to extract words from the string of Han characters. Chinese is a non-inflectional language and therefore morphological analysis is not essential.

In contrast, Japanese text is written usually as a mixture of Han characters, Hiragana and Katakana. Katakana is usually used to write non-Japanese words (except those borrowed from Chinese). Hiragana is mostly used to represent the inflectional part of a word and to substitute complicated (and less common) Han characters in modern Japanese. Japanese texts are also written without word delimiters and therefore, must be segmented. Prior to any word based indexing, due to the inflectional nature of Japanese, text must be morphologically analyzed and the root words should be indexed (equivalent to the *stemming* in western languages) to cope with the inflectional variations.

Due to the historical evolution and cultural differences, Han character itself become ambiguous across the CJK languages. We will discuss the semantic irregularities of Han characters in Japanese and Chinese below with examples.

Han Characters*:* In Japanese, the ideographic character-string, 切手 means *postal stamp*. The constituent characters, if used independently in other contexts, represent "to cut" and "hand", respectively. However, in Chinese, 郵票 represents *postal stamp* and the constituent characters represent "postal" and "ticket", respectively. Interestingly, both in Japanese and in Chinese, the character string, 郵便局, represents *post office*. However, majority of the postal service related words, in both Chinese and Japanese, consist of the Han character, 郵 as a component. Although there are some idiosyncrasies, there are significant regularities in the usage of Han characters across the CJK languages. Like word sense disambiguation (WSD), Kanji Sense Disambiguation (**KSD**) within and across the CJK languages is an interesting area of research by itself. Lua (1995) reported an interesting neural network based experiment to predict the meaning of Han character based words using their constituent characters' semantics.

For effective CLIR, we need to analyze the irregular Han characters and work out relevant mapping algorithm to augment the query and document vectors. A simplistic approach (with binary weight) is illustrated in Table 1. For the partial co-occurrences of the characters like, 切, 手 and 郵, etc. in a particular document or a query requires adjustments of the document or the query vector. We are aware that such manual modification is not feasible for a large heterogeneous document collection. Dimensionality reduction techniques, like LSI (Evans at al., 1998; Rehder et al, 1998) or Han character clustering are the potential solutions to automatically discover associations among Han characters.

[4] A detail description of the *Unicode ideographic character unification rules* can be found in Unicode-2000, pp. 258-271.

Table 1: Enhancement of query or document vectors to create semantic association (an example)

Document or Query	**Vector Representation** (partial)
Han Characters appeared in a Japanese or a Chinese document or a query: [..切.. 手.. 郵.. 票..]'	
Possible binary vectors representing a query or a document (before enhancement)	[..1.. 1.. *.. *..]' [..*.. *.. 1.. 1..]' etc.
Mapped binary vector representing a query or a document (after enhancement)	[..1.. 1.. 1.. 1..]'

Asterisk (*) represents 0 or 1.

Katakana Strings*:* In Japanese, especially in the technological domain, Katakana is predominantly used to transliterate foreign words. For example, in modern Japanese, the words, ツール and テクノロジー, etc. (*tool* and *technology*, respectively) are very common. Their Han character equivalents are 道具 and 技術, etc., and they are similar to those used in Chinese. A Katakana to Kanji (Han character) mapping table is created to transfer the semantics of Katakana in the form of Han characters (relative positions of the document or query vector need to be adjusted) to help our Chinese-Japanese CLIR task. In this purpose, the definition part of a Japanese monolingual dictionary is used to find the relevant Han characters for a particular Katakana string. Manual correction is then conducted to retain the meaningful Han character(s).

Proper Names*:* In Japanese, foreign proper names are consistently written in Katakana. However, in Chinese, they are written in Han characters. For a usable CLIR system for Chinese and Japanese, a mapping table is therefore inevitable. In our experiment, due to the nature of the text collection, we manually edited the small number of proper names to establish association. We are aware that such manual approach is not feasible for large scale CLIR task. However, since proper name detection and manipulation is itself a major research issue for natural language processing, we will not address it here.

Hiragana Strings: Continuous long strings of Hiragana need to be located and replaced[5] with the respective Han characters, and the document and the query vectors must be adjusted accordingly. Shorter hiragana strings can be ignored as stop word since such hiragana strings are mostly functional words or inflectional attributes.

4 Vector Space Model: Western and Asian language perspective

The most popular IR model, the Vector Space Model, uses vectors to represent documents and queries. Each element of a document or a query vector represents the presence or absence of a particular term (binary), or the weight (entropy, frequency, etc.). Functional words are eliminated; stemming and other preprocessing are also done prior to the vectorization. As a result, syntactic information is lost. The vector simply consists of an ordered list of terms, and therefore, the contextual cues have also disappeared. The document and the query vectors are gross approximation of the original document or query (Salton et al., 1983). In vector space information retrieval, we sacrifice syntactic, contextual and other information for representational and computational simplicity. For western languages, sometimes phrase indexing is proposed to offset such losses and to achieve better retrieval quality. In vector space model, a *term* usually refers to a *word*. For western languages, a document or a query vector constructed from the letters of the alphabets would not yield any effective retrieval. However, representing CJK documents and query in terms of Han character vectorization yields reasonably effective retrieval. This is due to the fact that a Han character encodes non-trivial semantics information within itself, which is crucial for information retrieval. Han Character based document and query representation is therefore justified. For CLIR,

[5] In Japan, it is common that materials written for young people uses Hiragana extensively to bypass complex Han characters.

considering the inherent complexity in query and document translation, multilingual dictionary and thesaurus maintenance, etc., Han character based (both single character or n-gram characters) approaches under the vector space framework, despite of being a gross approximation, provide significant semantic cues for effective retrieval due to the same reason.

5 Experimental Setup

We collected the translated versions of the Lewis Carroll's *"Alice's Adventure in the Wonderland"* in Japanese and in Chinese. The original Chinese version (in GB code) and the original Japanese version (in S-JIS code) are then converted into Unicode. Preprocessing is also conducted to correlate the proper names, to resolve the semantic multiplicity of coding and to associate the language specific irregularities, etc. as described in **S**ection **2** and **3**.

The *mg* system (a public domain indexing system from the New Zealand Digital Library project, Witten et al., 1999) is adapted to handle Unicode and used to index the Unicode files. We consider each paragraph of the book as a single document. There are 835 paragraphs in the original book and the translated versions in both Japanese and Chinese also preserve the total number of paragraphs. In this way, we have a collection of 1670 paragraphs (hereafter, we refer to each *paragraph* as a *document* of our bilingual text collection) in both Chinese and Japanese. We used the *mg* system to index the collection based on TF.IDF weighting. For a particular query the *mg* system is used to retrieve documents in order of relevance.

We asked 2 native Japanese who have an intermediate level understanding of Chinese language and who are the frequent users of the Internet search engines, to formulate 5 queries each in natural Japanese. Similarly, we also asked 2 native Chinese who have the intermediate level understanding of Japanese and who are the frequent users of the Internet, to formulate 5 queries each in Chinese. Therefore, 4 bilingual human subjects formulated a total of 20 queries in their respective native tongue (10 queries in Chinese and 10 queries in Japanese). The subjects were initially not told about the cross language issues involved in the experimental process, that is, the subjects formulated the queries as how they would usually do for monolingual information retrieval.

All the 4 subjects are familiar with the story of the *Alice's Adventure in the Wonderland*. However, we asked them to take a quick look at the electronic version of the book in their own language to help them to formulate 5 different queries in their own native language.

Table 2: Comparison of mono- and cross- language information retrieval

	Number of **Chinese** documents judged relevant (a total of 10 documents are retrieved for each query) *Out of 100 retrieved docs*	Number of **Japanese** documents judged relevant (a total of 10 documents are retrieved for each query) *Out of 100 retrieved docs*	**CLIR to MLIR ratio**
Queries in Chinese (total 10 queries from 2 native Chinese subjects)	35	26	**74 %**
Queries in Japanese (total 10 queries from 2 native Japanese subjects	19	30	**63 %**

Documents are retrieved with the queries from both the Japanese and the Chinese versions of the book. Top 10 documents in Chinese and top 10 documents in Japanese language are then

retrieved for each query. Each subject is then presented with the 20 extracted documents for each of his/her own original query. Therefore, for the total 5 queries formulated by a subject, a total of 100 documents (50 documents in his/her mother tongue and 50 documents in the other language) are given back to each subject for evaluation. Subjects are asked to evaluate the documents extracted in their native language first and then similarly the documents extracted in the other language.

As shown in Table 2, it can be concluded that the cross language information retrieval in this experimental framework performed about 63-74% as good as their monolingual counterparts. Cross language information retrieval of European languages, with the help of multilingual thesaurus enhancement reaches about 75% performance of their monolingual counterparts (Eichman et al., 1998). The effectiveness of Han character based CLIR for CJK languages is therefore promising. It is important to note here that in business, political and natural science domains, Han characters are prevalently correlated across Japanese and Chinese documents. Our approach should perform even better if applied in those domains.

6 Further Research

In our experiment, we represent Chinese and Japanese documents and queries as weighted vectors of Han Characters. Before the vectorisation, necessary preprocessing is done to cope with the *multiplicity of coding* problem of semantically similar ideographs and to cope with some obvious language specific issues. Same as the monolingual vector space information retrieval approach, we measured cosine similarity between a query and a document to retrieve relevant documents in order of relevance. Similarity is measured for both cases; that is, (1) monolingual: the query and the document are in the same language, and (2) cross-language: the query and the document are of different languages. The comparative result shows that the effectiveness of cross language information retrieval between Chinese and Japanese in this way is comparable to that of other CLIR experiments conducted mainly with multiple western languages with the help of thesauri and machine translation techniques.

One of the promising applications of this approach can be in identifying and aligning Chinese and Japanese documents online. For example, retrieving relevant news articles published in both languages from the Internet. It is understood that several mathematical techniques, like Han character clustering and dimensionality reduction techniques (Evans et al., 1998) can augment and automate the process of finding associations among the Han characters within and across the CJK languages. The vector space model is also flexible for the adjustment of weighting scheme. Therefore, we can flexibly augment the Han character based query vectors (a pseudo- query expansion techniques) and document vectors (a pseudo-relevance feedback technique) for effective CLIR. We left these parts as our immediate future work.

As done with the MLIR, n-gram characters based indexing can also be experimented. However, due to the small document collection and the number of queries we had, n-gram based indexing suffers from data sparseness problem. We, therefore, left out the n-gram character based CLIR evaluation until a huge collection of documents and queries are ready.

Conclusion

In this paper, we experimented on a small collection of homogeneous bilingual texts and a small set of queries. The result obtained supports the promising aspect of using Han characters for cross language information retrieval of CJK languages. Such an approach has its own advantage since no translation of query or documents are needed. In comparison to maintaining multilingual dictionaries or thesauri, maintaining Han characters mapping table is more effective because the mapping table needs not to be updated so often. Sophisticated mathematical analysis of Han characters can bring a new dimension in retrieving cross Asian language information. Kanji Sense Disambiguation (KSD) techniques using advanced machine learning techniques can make the proposed CLIR method more effective. KSD is a long neglected area of research. Dimensionality reduction techniques, clustering, independent component analysis (ICA) and other mathematical methods can be exploited to

enhance Han character based processing of CJK languages.

References

Chen, A., Jianzhang He, Liangjie Xu, Fredric C. Gey and Jason Meggs (1997). Chinese Text Retrieval Without Using a Dictionary. In Proceeding of the Conference on Research and Development in Information Rertrieval, ACM SIGIR-97, pp. 42-49.

Eichmann, D., M.E. Ruiz and P. Srinivasan (1998). Cross-language Information Retrieval with the UMLS Metathesaurus. In Proceeding of the Conference on Research and Development in Information Rertrieval, ACM SIGIR-98, pp. 72-80.

Evans, D.A., S.K. Handerson, I.A. Monarch, J. Pereiro, L. Delon, W.R. Hersh (1998). Mapping Vocabularies Using Latent Semantics. In Gregory Grefenstette Edited, *Cross-Language Information Retrieval*, Kluwer Academic Publisher.

Grefenstette, G. (1998) The Problem of Cross-Language Information Retrieval. In Gregory Grefenstette Edited, *Cross-Language Information Retrieval*, Kluwer Academic Publisher, pp. 1-10.

Fujii, H. and W.B. Croft (1993). A comparison of Indexing for Japanese Text Retrieval. In Proceeding of the ACM SIGIR-93, pp. 237-246.

Kim T., Sim C.-M., Yuh S., Jung H., Kim Y.-K., Choi S.-K., Park D.-I., Choi K.S. (1999). FromTo-CLIR™: web-based natural language interface for cross-language information retrieval. Journal of Information Processing and Management, Pergamon, Vol. 35. No.4. pp. 559-586.

Kwok, K.L. (1997). Comparing Representation in Chinese Information Retrieval, In Proceeding of the ACM SIGIR-97, pp. 34-41.

Lee, J.H. Hyun Yang Cho, Hyouk Ro Park (1999). n-Gram-based Indexing for Korean Text Retrieval. Journal of Information Processing and Management, Pergamon, Vol. 35. No.4. pp. 427-441.

Lua K.T. (1995) Predication of Meaning of Bisyllabic Chinese Words Using Back Propagation Neural Network. In Communications of COLIPS, An International Journal of Chinese and Oriental Languages Information Processing Society, Vol.5, Singapore. URL:

http://www.comp.nus.edu.sg/~colips/commcolips/paper/p95.html

Matsumoto, Y., H. Kitauchi, T. Yamashita (1997). User's Manual of Japanese Morphological Analyzer, ChaSen version 1.0 (in Japanese). Technical Report IS-TR97007, Nara Institute of Science and Technology (NAIST), Japan.

Mudawwar, M.F. (1997). Multicode: A Truly Multilingual Approach to Text Encoding. IEEE *Computer*, Vol. 30. No. 4, pp. 37-43.

Nie, J.Y., Martin Brisebois and Xiaobo Ren (1996). On Chinese Text Retrieval. In Proceeding of the ACM SIGIR-96, pp. 225-233.

Nie, J.Y., Jean-Pierre Chevallet and Marie-France Bruandet (1998). Between terms and Words for European Language IR and Between Words and Bigrams for Chinese IR. In Proceeding of Text REtrieval Conference (TREC-6), pp. 697-710.

Nie, J.Y. and Fuji Ren (1999). Chinese Information Retrieval:using character or words? Journal of Information Processing and Management, Pergamon, Vol. 35. No.4. pp. 443-462.

Oard, D.W. and Bonnie J. Dorr (1996). A Aurvey of Multilingual Text Retrieval. University of Maryland, Technical Report, UMIACS-TR-96-19, CS-TR-3615.

Ogawa, Y. and Toru Matsuda (1997). Overlapping Statistical Word Indexing: A New Indexing Method for Japanese Text. In Proceeding of the ACM SIGIR-97, pp. 226-234.

Pergamon-1999 (1999) Special issue on Information Retrieval with Asian languages, Journal of Information Processing and Management, Vol 35. No.4. Pergamon Press, London.

Rehder, B., M.L. Littman, Susan Dumais and T.K. Landauer (1998). Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In Proceeding of Text REtrieval Conference (TREC-6), pp. 233-240.

Salton, G. and M.J. McGill (1983). Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

Sproat, R., Chilin Shih, William Gale and Nancy Chang. A Statistic Finite State Word-Segmentation Algorithm for Chinese, Computational Linguistics, Vol. 22 No. 2, pp. 377-404.

Tan C.L. and Makoto Nagao (1995) Automatic Alignment of Japanese-Chinese Bilingual Texts, In IEICE Transactions of Information and Systems, Japan. Vol. E78-D. No. 1. pp. 68-76.

TREC-6 (1998). Proceeding of Text REtrieval Conference (TREC-6). National Institute of Sciece and Technology (NIST). URL:

http://trec.nist.gov/pubs/trec6/

Unicode-2000 (2000). The Unicode Standard, Version 3.0, Addison Wesley, Reading, MA, URL: http://www.unicode.org/

Witten I.H., Alistair Moffat and T.C. Bell (1999). Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition, Morgan Kaufmann Publishers.