Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 59-86

Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach

Md. Maruf Hasan* and Yuji Matsumoto *

Abstract

Electronically available multilingual information can be divided into two major categories: (1) alphabetic language information (English-like alphabetic languages) and (2) ideographic language information (Chinese-like ideographic languages). The information available in non-English alphabetic languages as well as in ideographic languages (especially, in Japanese and Chinese) is growing at an incredibly high rate in recent years. Due to the ideographic nature of Japanese and Chinese, complicated with the existence of several encoding standards in use, efficient processing (representation, indexing, retrieval, etc.) of such information became a tedious task. In this paper, we propose a Han Character (Kanji) oriented Interlingua model of indexing and retrieving Japanese and Chinese information. We report the results of mono- and cross- language information retrieval on a Kanji space where documents and queries are represented in terms of Kanji oriented vectors. We also employ a dimensionality reduction technique to compute a *Kanji Conceptual Space* (KCS) from the initial Kanji space, which can facilitate conceptual retrieval of both mono- and cross- language information for these languages. Similar indexing approaches for multiple European languages through term association (e.g., latent semantic indexing) or through conceptual mapping (using lexical ontology such as, WordNet) are being intensively explored. The Interlingua approach investigated here with Japanese and Chinese languages, and the term (or concept) association model investigated with the European languages are similar; and these approaches can be easily integrated. Therefore, the proposed Interlingua model can pave the way for handling multilingual information access and retrieval efficiently and uniformly.

Keywords: Cross-language Information Retrieval; Multilingual Information Processing; Latent Semantic Indexing.

*Nara Institute of Science and Technology 8916-5, Takayama, Ikoma, Nara, 630-0101 Japan
E-Mail: {maruf-h, matsu}@is.aist-nara.ac.jp

1. Introduction

The amount of multilingual information available electronically has escalated in recent years. Lately, the information in non-English European languages and in Chinese, Japanese, Korean and Vietnamese (CJKV) is increasing at an incredibly high rate. Both Japanese and Chinese (also, Korean and Vietnamese, to some extent) are ideographic languages that use thousands of ideographic characters (also known as: Han characters, Kanji, Hanzi or Hanja) in writing. Managing the huge number of characters is no longer a problem in processing Japanese and Chinese language information. However, computer processing of these languages is complicated with the absence of word delimitation and the existence of several national and industrial encoding standards. Word delimitation (also called, *segmentation*) is an extra task to perform to process these languages, because in the written Japanese and Chinese texts, explicit boundaries between words are not available. Due to the existence of several encoding standards, it is also quite common to notice that most Internet search engines provide two different services to search for Chinese information in *Traditional Chinese* (commonly encoded in BIG-5 code) and the *Simplified Chinese* (GB code) form. Technically speaking, the tasks of processing and retrieval of traditional and simplified Chinese can be considered the tasks involving two distinct languages. Similarly, JIS, Shift-JIS and EUC, etc. are common Japanese encoding standards. The existence of several encoding standards complicates the information retrieval (IR) tasks for both Japanese and Chinese [24]. In this paper, we will formulate a unified framework to cope with the above-mentioned problems as well as to facilitate effective multilingual information retrieval.

Electronically available multilingual information can be divided into two major categories: (1) alphabetic language information (English-like alphabetic languages) and (2) ideographic language information (Chinese-like ideographic languages). Unicode, an increasingly popular encoding standard, defines uniform codes for almost all characters (both alphabetic and ideographic) of the world languages [43]. The common CJK ideographs section defined under Unicode is a superset of all ideographic Han characters used across the CJKV languages. This offers us an opportunity to represent Japanese and Chinese documents uniformly in Unicode. By doing so, we can also take advantage of Kanji to index and retrieve information across these languages. Nonetheless, the Kanji-derived semantic units or concepts can also be easily associated with the corresponding terms (stem, word or concept, etc.) of the alphabetic languages, and therefore, a universal multilingual IR framework can also be achieved.

The ubiquity of the Internet, the proliferation of electronic information and the emergence of globalization offer us the challenge of engineering sophisticated techniques to process multilingual and heterogeneous information efficiently. Therefore, in the recent years, the IR community put an exclusive focus on Cross-language Information Retrieval (CLIR) to address this new challenge [33]. CLIR investigates information indexing and retrieval issues across the languages. CLIR is a special case of Monolingual Information Retrieval (MLIR), and addresses the retrieval issues where queries and documents are given in different languages. If either the query or the document collection can be

effectively translated into the target language, the CLIR problems can be reduced to MLIR problems. The commonly used techniques for CLIR include three different approaches: (1) query translation, (2) document translation, and (3) combination of both query and document translation. However, given the fact that the quality of machine translation (MT) is still well below the desired level, CLIR often takes advantage of multilingual dictionaries, thesauri or word or sentence -aligned parallel corpora to circumvent MT. Also, there are CLIR approaches, which tend to bypass MT by making use of multilingual conceptual ontology [8] or multilingual term association [36]. Successful CLIR systems for European languages (English, French, Spanish and German, etc.) are demonstrated using conceptual mapping and term association techniques. In this paper, we investigate mono- and cross- language IR for Japanese and Chinese using Kanji mapping and semantic association of Kanji-derived concepts – a Kanji-based Interlingua CLIR for these ideographic languages.

Precisely speaking, we focus on the semantic information captured in Kanji and attempt to engineer the Kanji for effective mono- and cross- language information retrieval for Japanese and Chinese information. Unlike the characters (letters) of the non-ideographic languages (e.g., English, Arabic or Sanskrit), a single Kanji is capable of capturing significant semantic information within itself. However, single Kanji is ambiguous, and therefore, we attempted to index the Kanji through their explicit and implicit semantic contents. Despite our focus on these ideographic Asian languages, we have also included a discussion towards developing an Interlingua model of multilingual information processing, which is capable of handling other (non-ideographic) languages.

The organization of this paper is as follows. We briefly discuss the special issues of Japanese and Chinese information retrieval (IR) in Section 2. We include a detailed literature review of Japanese and Chinese IR in Section 3. In Section 4, we discuss several encoding standards of Japanese and Chinese texts, and their relationships with the Unicode. Our Kanji oriented retrieval experiments include four different indexing approaches: (1) single Kanji indexing, (2) single Kanji with Kanji bi-gram indexing, (3) single Kanji with correlated Kanji pair indexing (i.e., indexing the Kanji pairs that have high co-occurrence tendencies), and (4) Kanji based semantic indexing (i.e., by extracting latent Kanji concepts after applying the dimensionality reduction techniques). In Section 5, we introduce the vector space IR model in terms of Kanji vectors and Kanji-document matrix. The detail mathematical formalism of Kanji Co-occurrence Tendency (KCT) and Kanji Semantic Indexing (KSI) are outlined in Section 6. Finally, we discuss our mono- and cross- language IR experiments in Section 7, followed by a discussion and analysis in section 8. Throughout the entire article, whenever appropriate, we draw analogical comparisons between our approach and those of others to justify the formalism and the benefit of the proposed Interlingua model for multilingual information indexing and retrieval.

Readers who are familiar with the Japanese and Chinese language processing and vector space IR techniques, may skim through the introductory sections (Section 2, 3 and 4) and focus more on the later sections of this paper. Introductory sections are marked with asterisks.

2*. Special Issues in Japanese and Chinese Information Retrieval

In the following two Subsections, we will analyze the linguistic facets of the Japanese and Chinese languages, respectively, from an IR perspective.

2.1* Japanese IR

An investigation into popular Japanese full-text IR systems (NAMZU and FREYA, etc.) revealed that most Japanese IR systems mimic the IR systems for European languages [12, 28]. The indexing and retrieval usually involve with the four major steps: (1) *segmentation*: to locate word boundaries, (2) *morphological analysis*: to find the word-stems, (3) *representation and indexing of the stems*: e.g., to use inverted file or other data-structures to represent the document collection, and (4) *query-document similarity measurement*: to associate documents with queries using cosine or other similarity measures. The first two steps, segmentation and morphological analysis, are computationally expensive complex tasks, and these preprocessing steps result in a loss of syntactical cues from the documents and the queries. Given also the fact that vector space representation of documents and queries is a flat representation (known as a "bag of words"), which ignores contextual information of the original documents and the queries [37], it makes sense to represent the documents and queries only in terms of Kanji. We will investigate several ways of indexing Kanji (sometimes associated with further processing, such as Kanji mapping and correlation, etc.) straightforwardly to bypass computationally intensive segmentation and morphological analysis. It can be noted that, for Japanese-Chinese CLIR, such an approach can bring us an added advantage because the query or the document translation steps can be easily circumvented with Kanji association.

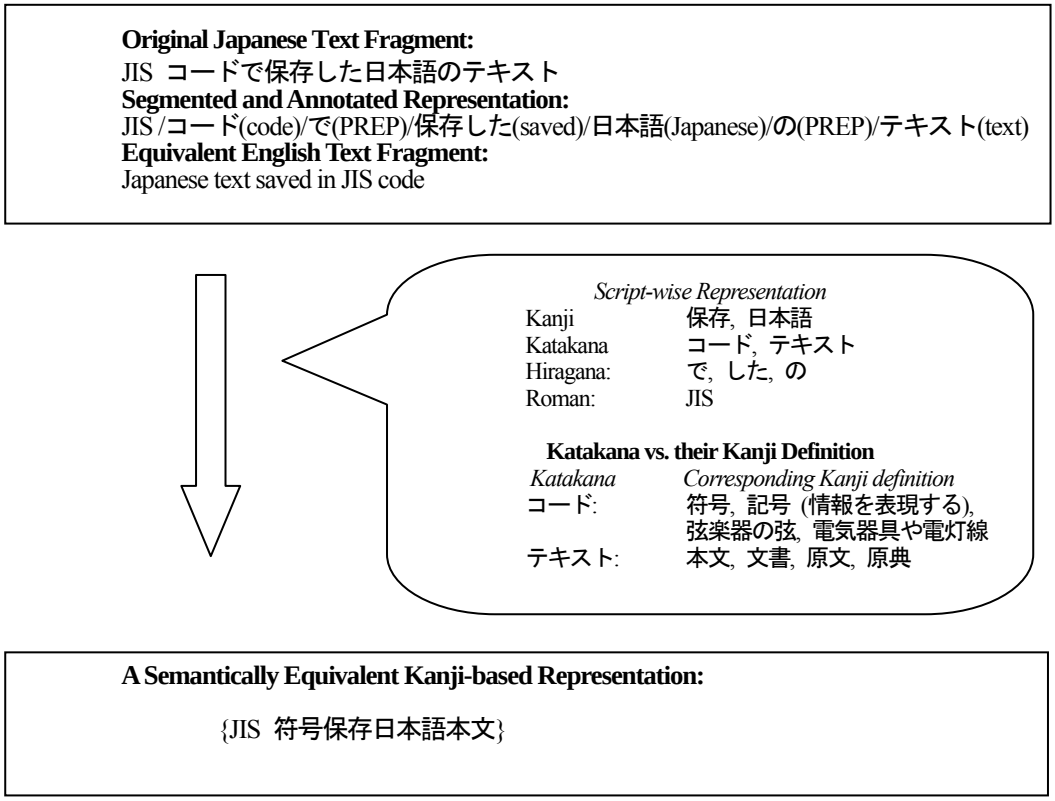

Figure 1 *A Japanese text fragment with four different scripts and its maximum likelihood mapping to corresponding Kanji*

Japanese texts are usually written using four different scripts: Kanji, Katakana, Hiragana and Roman alphabet [39]. The following example of a short Japanese text fragment (in Figure 1) shows the combination of all four scripts. In running Japanese texts, however, Kanji is more dominating than any other scripts. In writing, ideographic Kanji can be replaced with its Hiragana spelling. However, the Hiragana spellings are ambiguous. Kanji expresses the semantic more obviously than its Hiragana spelling; and probably because of this reason, despite a long ongoing debate on replacing Kanji completely with the Hiragana, Kanji still remains a major component in Japanese writing. Katakana is mostly used to transliterate loan words (except those borrowed from Chinese). However, the phonetic scarcity of Japanese, made the Katakana transliteration quite ambiguous. For example, the Katakana string, コード may represent different English words: code, cord and chord, etc. Moreover, Katakana transliteration is often inconsistent; the English word, "digital" can be written as デジタル or ディジタル, for instance. Most Katakana strings are content bearing terms and therefore, in Japanese IR, special care has always been taken to efficiently process the Katakana strings. In the ordinary Japanese-English dictionaries, a Katakana

entry appears with its original foreign word and a Japanese definition (often using Kanji). For example, the Katakana string, テキスト is defined with the relevant Kanji strings 文書, 原文, 原典 or 本文 as well as the English word, "text". In our Kanji based IR approach, we propose mapping the Katakana strings onto the relevant Kanji with the help of dictionary definitions. Hiragana strings are mostly shorter functional words or inflectional components. Short Hiragana strings usually play syntactical and morphological roles, and are usually ignored. However, continuous and long strings of Hiragana are potentially a replacement[1] for a rare and complicated Kanji (or a Kanji string), which, to some extent of accuracy, can also be replaced with the relevant Kanji. The Roman alphabet is usually mapped to the respective ASCII characters and indexed accordingly. For an effective Kanji-based information retrieval, we need to preprocess the Japanese documents and queries and represent them in terms of equivalent Kanji. Although sophisticated algorithms can be developed for Hiragana to Kanji and Katakana to Kanji mapping, we, for simplicity, use maximum likelihood based mapping strategy in this research. In Figure 1, we explain a Kanji based representation of the example text fragment using corresponding Kanji (the acronym, JIS, stands for the Japanese Industrial Standard).

It is worthy to note here that the complexity and the computational costs of such preprocessing are lesser than those of full segmentation and morphological analysis. Another justification for processing Japanese IR in this way is the usability of such indexing for CLIR without machine translation. Nonetheless, mapping a Katakana string to its original language is another feasible option for multilingual IR.

Although a single Kanji captures significant semantic information within itself, such information is highly ambiguous. Therefore, we also derived useful indexing information, such as, Kanji n-grams[2], correlated Kanji pairs and principal components, automatically from the initial Kanji based representation for effective indexing and retrieval.

2.2* Chinese IR

In comparison to Japanese IR, Chinese information retrieval is more straightforward because Chinese texts are mostly written homogeneously using only Kanji. However, like Japanese, Chinese text is also written without explicit word delimiters and therefore, segmentation must be performed to extract words from the string of Kanji for word or phrase level indexing. Since Chinese is a non-inflectional language, morphological analysis is not important. There are three major approaches to indexing Chinese text: (1) *single Kanji indexing,* (2) *Kanji n-gram indexing,* and (3) *word or phrase level indexing* (after segmentation). However, most practical systems incorporate more than one of the above indexing schemes

[1] In Japanese, materials written for the young people often include Hiragana strings to substitute rare Kanji.

[2] 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.*

for effective retrieval [32]. There are also reports on Chinese conceptual IR using a conceptual or semantic hierarchy [5].

The above discussion provides a fairly straightforward view of Chinese IR. However, due to the existence of several incompatible encoding standards, especially the Traditional (Big-5) and Simplified (GB) Chinese character encoding, Chinese IR is suffering a serious drawback. Moreover, from a technical standpoint, the Traditional Chinese IR and the Simplified Chinese IR can be viewed as two individual monolingual IR problems. This argument is further supported with the fact that most Internet search engines process information for simplified and traditional Chinese separately and offer two individual search interfaces for information retrieval in the specific encoding. We will discuss the coding related matters in Section 4 again after reviewing the CJK information retrieval literatures in the following section.

3*. A Collective Review of CJK Information Retrieval Related Work

In this section, we will present a brief review of CJK information retrieval. Although we will not report any experimental results on Korean IR in this paper, we have included a few Korean references in the appropriate contexts.

3.1* CJK Mono- and Cross- Language IR

Several approaches are investigated in CJK text indexing to address monolingual information retrieval (MLIR) - for example, (1) indexing words or phrases after segmentation and morphological analysis, (2) indexing *n*-gram ideographic characters, and (3) indexing single ideographic characters. From the potentially un-delimited sequence of characters, words must be extracted first. For word and phrase level indexing of the *inflectional* ideographic languages (e.g., Japanese and Korean), morphological analysis must also be performed. Sentences are segmented into words with the help of a dictionary using heuristic rules or machine learning techniques. Morphological analysis also needs intensive linguistic knowledge and computer processing. Segmentation and morphological analysis are complex tasks, and the accuracies of automatic segmentation and morphological analysis considerably vary across different domains. For the heterogeneous information sources, like the Internet, the accuracy of segmentation and morphological analysis perform more poorly than that of a particular domain. The computationally expensive word-based indexing of CJK texts, however, can contribute to better retrieval results when compared to the *n*-gram counterpart. Words and phrases are less ambiguous indexing units than the *n*-grams or the correlated *n*-grams, and therefore, can boost retrieval performance. Segmentation and morphological analysis related issues of Chinese, Japanese and Korean are intensively addressed elsewhere [40, 26, 18].

The n-gram (n >1) character-based indexing is computationally expensive as well. The number of indexing terms (n-grams) increases dramatically as n increases. Moreover, not all the n-grams are semantically meaningful words; therefore, smoothing and filtering heuristics must be employed to extract

lexically meaningful n-grams for effective retrieval of information. See [29, 30, 31, 4, 13, 19] for details. For European language CLIR, exciting experimental results are reported in [16] using n-gram character associations across English and French.

For Japanese and Chinese IR, indexing single character (Kanji) is straightforward and less demanding in terms of both space and time than those of n-gram or other indexing schemes. From a CLIR point of view, for single Kanji indexing, there is no need to (1) maintain a multilingual dictionary or thesaurus of words, (2) to extract words and morphemes, and (3) to employ machine learning and smoothing to prune trivial n-grams or to resolve ambiguity in word segmentation [21, 34, 22]. Moreover, for such a single Kanji based approach, there is no translation overhead for both queries and documents. This approach also eliminates some of the typical CLIR related problems discussed in [14].

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 [19, 30, 31, 21]. For MLIR, n-gram and word -based approaches outperformed the single character based approach, at the cost of the extra time and space. Similar comparisons and conclusions for Japanese and Korean MLIR are made in [13] and [22], respectively.

Unlike MLIR, in cross-language information retrieval, a great deal of effort is allocated in maintaining the multilingual dictionary and thesaurus, and translating the queries and documents, and so on. In the CINDOR (**C**onceptual **IN**terlingua **DO**cument **R**etrieval) search from TextWise, LLC [8] uses a multilingual conceptual Interlingua approach for multilingual (English, French, Spanish and other European languages) information retrieval. Chen et al. [5] investigated conceptual CLIR of Chinese and English by mapping the WordNet Synsets to the concept hierarchy of a Chinese thesaurus. 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 [36]. Character tri-gram -based CLIR results [16] are also reported for English and French, which share a similar vocabulary. CLIR using Kanji association is also explored for ideographic languages like Japanese and Chinese [15]. However, no experiments have ever been conducted with a combination of ideographic and alphabetic languages through vocabulary association. This situation probably exists because of our practice of considering alphabetic and ideographic languages from different point of view. Such a practice should not be continued to foster truly multilingual information access and retrieval.

The authors sadly noticed that most of the CLIR research in recent years is focused on European languages. After the opening of the CLIR track in the TREC-6 conference [42], 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-English, Japanese-English, etc.). TREC has not yet initiated any CLIR track 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 [32]. Among the eight papers included in that special issue, only one

paper [19] addressed CLIR on multiple Asian language information retrieval (English, Japanese and Korean CLIR) using multilingual dictionaries and machine translation techniques (to translate both queries and documents within these languages). The recent initiative from the Asian Multimedia Forum (AMF) brought together three research institutes, NTT (Japan), KAIST (Korea) and KRDL (Singapore) to collaborate on CLIR research for CJK languages. However, their main focus is on using machine translation techniques for query and document translation, and thereby, integration of CJK information on the Internet to facilitate CLIR [2]. The series of conferences held in the name of IRAL (Information Retrieval with Asian Languages) addresses CLIR mostly in the traditional framework of query and document translation. For IRAL participants, the choice of languages still remains English and one of the Asian languages. We will investigate an Interlingua framework for Japanese-Chinese CLIR, which can easily be extendable to cover other languages.

3.2* Important Facts about Japanese and Chinese

Tan and Nagao [41] used Kanji correlation to align Japanese-Chinese parallel texts. According to them, the occurrence of common Kanji (in Japanese and Chinese language texts) sometimes is so prevalent that even a monolingual reader could perform a partial alignment of the bilingual texts. This fact can be further verified with the example in Figure 2. The common and visually similar Kanji appeared in news articles written in Chinese and Japanese provides enough cues to correlate the two reports (both describing a political crisis in the Korean Peninsula).

We would like to mention here that the named entities play an important role in IR and there is an intensive research focus on named entity extraction -related research. Not only is the extraction of named entities important, but the IR community must also work towards universal (Interlingual) representation and indexing of named entities for effective multilingual IR.

It should be noted that the pronunciations of the Kanji vary significantly across the CJK languages, but the visual appearances of the Kanji in written texts (across CJK language) have a certain level of similarity. The Unicode Kanji Information Dictionary [38] provides cross-references among all the unified CJK ideographs (encoded by the Unicode Consortium) across the CJK languages including the cross-reference between simplified and traditional forms of a Chinese character. As explained above, we may conclude that effective cross-language information retrieval by indexing and associating the non-trivial Kanji semantics holds promises. We can also bypass complicated segmentation or morphological analysis process using such an approach. At the same time, multilingual dictionaries and thesauri maintenance, as well as query and documents translations can also be avoided. In Section 7, we will report such experimental results.

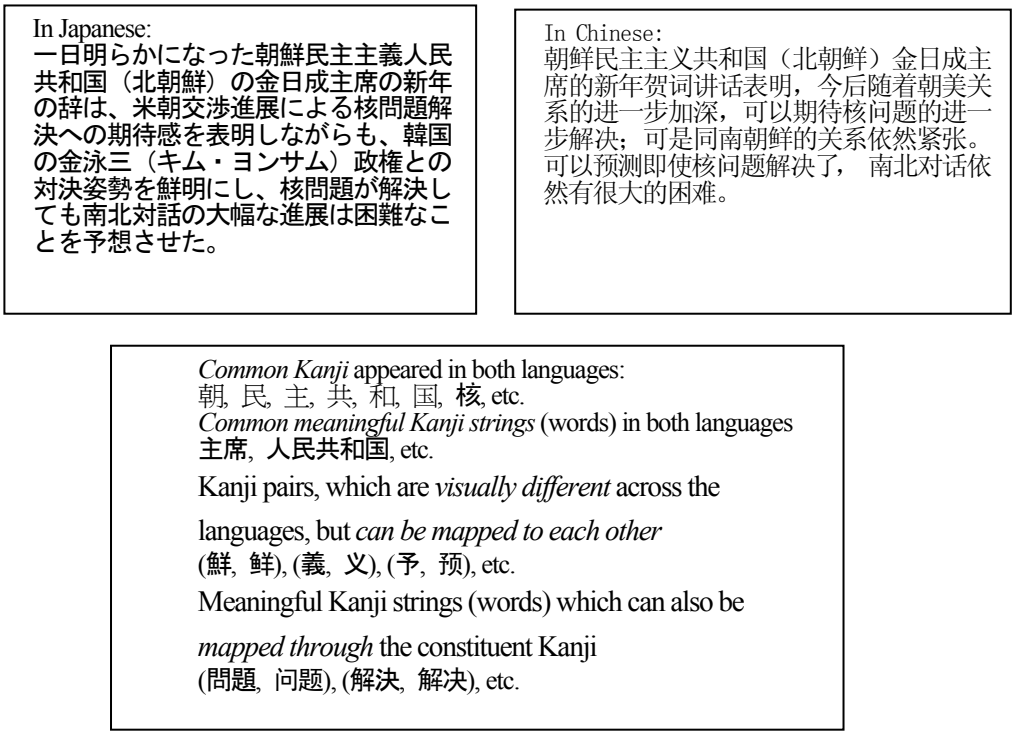

Figure 2 *Examples of common and similar kanji across Japanese and Chinese newswires*

4*. Japanese and Chinese Encoding Standards vs. the Unicode

Character encoding schemes of Japanese and Chinese have several variations: for example, Chinese encoding standards include two major standards- GB and BIG-5; and Japanese encoding standards include JIS, Shift-JIS and EUC, etc. A typical Internet search engine usually asks users to specify not only the language but also the encoding scheme (e.g., simplified (GB) or traditional Chinese (BIG-5)) while searching for information on the same language. For comprehensive details on different encoding standards, readers are referred to [24]. The number of Kanji encoded under a particular encoding scheme of a particular language also varies. Due to the huge difference in the number of Kanji encoded in simplified Chinese (GB) and in traditional Chinese (BIG-5), the retrieval of simplified and traditional Chinese are currently being processed as if they were two different languages. However, the problem with Japanese IR (due to the existence of different encoding standards) is not as severe as that of Chinese.

Unicode, a comprehensive coding scheme of the world languages, paves the way for an alternative and consistent representation of textual information.

Due to the growing acceptance and popularity of the Unicode [43] by the computer industry, we

have a common platform to investigate multiple languages in a unified framework. 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 assigns codes to 27,484 Kanji, a superset of characters encoded under the existing standards. Unicode makes it possible to represent documents uniformly across these languages. By representing Japanese and Chinese documents in Unicode and finding association through Kanji vocabulary it is possible to efficiently address the IR issues of these languages.

Figure 3 *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 [27]. The Unicode standard avoids duplicate encoding of the same character; for example, the character 'a' is encoded only once although it is included in the alphabets of 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 cultural settings. The ideographic characters in Figure 3, although they represent the same word (sword in English), are given six unique codes under the Unicode encoding scheme to satisfy the round-trip criteria[3], that is, to allow round-trip conversion between the source standard (in this case, JIS, which assigns 6 distinct codes) and the Unicode. The 27,484 Kanji encoded in Unicode, therefore, includes semantic redundancy in both single-language and multiple-language perspectives.

[3] A detail description of the *Unicode ideographic character unification rules* can be found in [43],

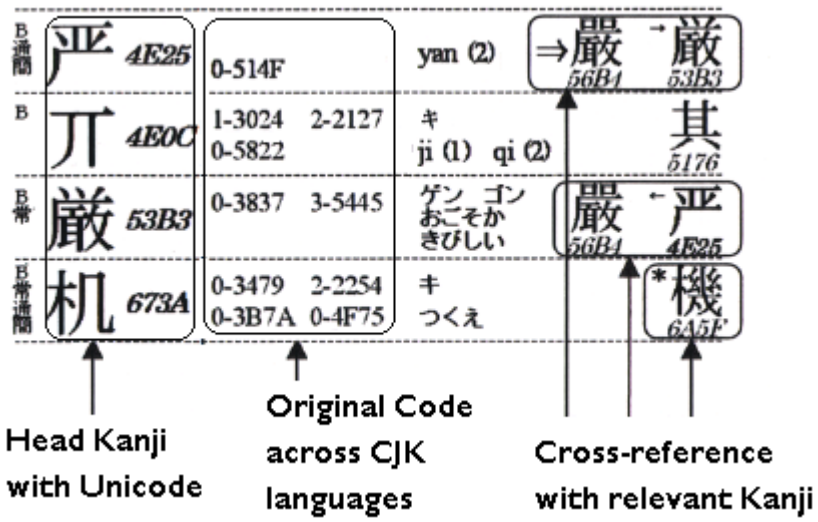

Figure 4 *Entries from the* Unicode Kanji Information Dictionary: *Kanji are annotated with cross-reference information within and across the CJK languages.*

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 Kanji-based CLIR approach, any such redundancy and multiplicity must be identified and resolved to achieve semantic uniformity and association within and across languages. Such tasks are less painstaking than the maintenance of multilingual dictionaries and thesauri of words and concepts. New lexical and ontological references, like the Unicode Kanji Information Dictionary [38] provides substantial co-reference information to assist such tasks. In our experiment, we use a table-lookup mapping approach to locate and associate the semantically related (but visually dissimilar) ideographs within and across CJK languages as a pre-processing task. Nonetheless, we are aware that there are cases where the same Kanji represents totally different concepts across the two languages in question. For the Kanji oriented CLIR, this phenomenon can somehow be considered as Kanji polysemy[4]. Such polysemy resolution can open up a new area of be further research (theoretically similar to word sense disambiguation problems).

5. Kanji-Document Matrix in Japanese and Chinese Information Retrieval

As justified above, in this research, we represent Japanese and Chinese documents uniformly using

[4] Kanji polysemy across the languages

Unicode. For Japanese information retrieval, since disregarding the Hiragana and Katakana could cause a partial loss in document and query semantics, we also perform further preprocessing by locating and replacing the lexically meaningful Hiragana and Katakana strings with the relevant Kanji (or Kanji string).

The next task is to index the Japanese and Chinese documents in terms of Kanji. Single Kanji indexing is the simplest among all types of indexing. Indexing Kanji n-grams (specially, bi-grams) is another possible alternative. Other correlation measures [10], used for calculating word co-occurrence of a language, are equally applicable to calculate Kanji correlation. We compute term frequencies (tf) and inverse document frequencies (idf) for single Kanji and Kanji bi-grams for indexing. We only consider bi-grams with medium frequency and exclude the most frequent and rare bi-grams based on empirically decided cut-offs. We also identify the correlated Kanji pairs based on the co-occurrence tendency measured using mutual information (c.f., Section 6.1). We only compute tf and idf of highly correlated Kanji pairs (decided by their mutual information measure) for indexing.

In the vector space IR model, a term-document matrix is computed from a collection of documents. Each column of the term-document matrix represents a document, and each row represents a term (e.g., a word, a phrase, a Kanji or a Kanji string). Each element of the matrix represents the weight of a term. For the simplest case, weights can be binary values, representing the presence or absence of a particular term in the document. The frequency of a term in a particular document normalized with the same term's inverse frequency with respect to the entire collection (generically known as, tf.idf) is also often used [37] as weights. A Kanji-document matrix is similar to a term-document matrix when we consider Kanji (one or more) as terms. We compute three different Kanji-document matrices using the tf.idf weighting scheme in three different ways: single Kanji (K_A for short), single Kanji with the Kanji Bi-grams (K_B), and single Kanji with the correlated Kanji pairs (K_C). These matrices are essentially the Kanji Space Representations of our document collection. Each column vector of the Kanji-document matrix is the Kanji Vector Representation of a particular document. Queries can also be represented as Kanji vectors. Relevance is computed by calculating the vector similarity between the query and the document collection.

5.1 Monolingual Kanji Conceptual Space (KCS)

The three different Kanji-document matrices introduced above are Kanji-vector representations of a document collection. *Kanji Conceptual Space* (KCS) is a conceptual representation of Kanji-concepts after projecting the original high dimensional Kanji vectors to a lower dimensional conceptual space. Although theoretically, we can compute three different KCSs from the three Kanji-document matrices, we will restrict us in computing only one KCS from the single Kanji based Kanji-document matrix (using the *log-entropy* weighting scheme). This restriction is due to the constrain that with a small collection of documents, if we include the Kanji bi-grams or the Kanji correlated pairs as terms, the total number of terms (including single Kanji) exceeds the total number of documents in the collection. This situation violates the assumption behind SVD as explained in Section 6.2. Moreover, using SVD to reduce the

dimension of a heterogeneous space (estimated using single Kanji along with the correlated Kanji pairs or Kanji bi-grams estimated from a small collection of documents) into a reduced space may not achieve a proper conceptual mapping.

We used the log-entropy weighting, (log (tf + 1) . entropy) as described in [9], where tf and the entropy are the term frequency and entropy of a Kanji. We chose the log-entropy weighting scheme for three reasons, (1) it is faster to compute, (2) other LSI-based experiments intensively use log-entropy measure, and (3) we verified (using non-parametric Wilcoxon matched-pair sign test) that the difference between the tf.idf weighting scheme and the log-entropy weighting for single Kanji indexing is insignificant[5]. Other weighting schemes that incorporate a local and a global factor (such as, tf.idf variants) may also be applicable.

Latent Semantic Indexing (LSI), a well-known vector space model of IR, is capable of performing conceptual information retrieval. LSI uses the singular value decomposition (SVD) technique to reduce the rank of the original term-document matrix. Theoretically, SVD, a principal component analysis technique, performs a term-to-concept mapping and therefore, conceptual indexing and retrieval is made possible [7]. Considering the computational overhead of SVD, we chose the log-entropy weighting scheme of single Kanji, which can be computed faster [9], and computed single-Kanji based Kanji-document matrices for both Japanese and Chinese documents. By applying SVD to these Kanji-document matrices, we can derive conceptual representations of a text object (a document or a query) on the Kanji conceptual space in the respective language. For convenience, we will refer to this dimensionality reduction based approach as K_D.

5.2 Cross-language Kanji Conceptual Space: An Interlingua Representation

For CLIR experiments of European languages, Rehder et al. [36] experimented with English-French-German CLIR using a multilingual parallel corpus of these languages. They decomposed the multilingual term-document matrix using SVD to find associations of words (e.g., vocabulary mapping) among these languages. Interlingual conceptual representation of a document or a query can be computed from decomposed multilingual term-document representation since such a representation captures significant information about cross-language vocabulary mapping [11].

As described in Section 5.1, the rank-reduced Kanji-document representation for Japanese (or Chinese) documents can be used to represent a Japanese (or Chinese) document or a query in the Japanese (or Chinese) Kanji conceptual space. Similarly, with a collection of properly aligned Japanese and Chinese parallel documents, it is possible to compute a reduced rank Kanji-document matrix on the unified Kanji space, where each column represents an aligned pair of bilingual documents (in terms of all the Kanji that appeared in the Japanese documents and in its corresponding Chinese document). A moderate size of such

[5] Wilcoxon tests can measure whether the difference in retrieval results between two experiments is significant [17].

a parallel corpus can capture the bilingual Kanji Conceptual Space across the two languages. Decomposing this bilingual Kanji-document matrix into a conceptual space theoretically enables us to compute an Interlingual Kanji Conceptual Space. Both mono- and cross- language information retrieval can be efficiently performed on this unified KCS. We discussed the mathematical formalism of such an approach in Section 6.2.

Since we did not find any suitable parallel corpora of Japanese and Chinese documents, we used a commercial MT system to translate the Japanese document collection into Chinese, and the Chinese collection into Japanese. Assuming that the quality of the machine translation has a trivial impact (since we are only interested in finding the Kanji association), the original documents and their translations provide us an alternative opportunity to roughly estimate the bilingual KCS. First, we computed a bilingual Kanji-document matrix using the log-entropy of each Kanji. By reducing the rank of this bilingual matrix using SVD, we can derive an Interlingual representation of our bilingual document collection on the unified KCS.

We want to conclude this section by pointing out that such an Interlingua representation, although derived from ideographic Kanji, is easily extendable to non-ideographic languages because of the flexibility of representation in the vector space model. Rather than associating Kanji with the word-stems of each alphabetic language, we consider mapping the words to their respective word-roots. Since the vocabulary of many European languages can be mapped back to their original roots (Latin, and Greek, etc.), word-roots associated with the Kanji can provide a multilingual conceptual space for effective representation and retrieval of multilingual information.

6. Kanji Co-occurrence Tendency and Kanji Semantic Indexing

6.1 Calculation of Kanji Co-occurrence Tendency (KCT)

Counting the weights for a single Kanji and that of bi-grams are straightforward and we skip the details for brevity. Here, we will define the *Kanji Co-occurrence Tendency* (KCT) and explain how we computed KCT and choose the correlated Kanji pair.

Mutual Information (MI) is one of the metrics that can be used for calculating the significance of word co-occurrence associations [6, 25]. We extended the idea to estimate KCT. The mutual information MI between two events x and y is defined as follows:

$$MI_2(x, y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where $p(x,y)$ is the joint probability that two events, x and y co-occur, and $p(x)$ and $p(y)$ are the probabilities that event x or y occurs independently.

MI can be applied to the Kanji in the documents and can be used to calculate the correlation between those Kanji. To calculate the co-occurrence tendency for two Kanji, k_1 and k_2, we define n_{ij} (i, j = 1,2) in a 2-by-2 table shown below. In the table, n_{11} indicates the number of times two Kanji, k_1 and k_2 co-occur within a text window, n_{12} indicates the number of times k_1 occurs, but k_2 does not occur within a text window, and so on.

Table 1. *Auxiliary Table for defining Kanji Co-occurrence Tendency*

	$\mathbf{k_2}$	$\mathbf{\sim k_2}$
$\mathbf{k_1}$	n_{11}	n_{12}
$\mathbf{\sim k_1}$	n_{21}	n_{22}

$n_{i.}$, $n_{.j}$ and N are defined as follows:

$$n_{i.} = n_{i1} + n_{i2}$$

$$n_{.j} = n_{1j} + n_{2j}$$

$$N = \sum_{i,j} n_{ij}$$

That is, $n_{i.}$ indicates the number of times k_1 occurs (i = 1) or does not occur (i = 2) regardless of the occurrence of k_2, and N indicates the total number of co-occurrence windows in the corpus.

The co-occurrence tendency of a Kanji pair k_1 and k_2 in a corpus, KCT_{MI2} is defined as follows:

$$KCT_{MI2}(k_1, k_2) = \log_2 \frac{\frac{n_{11}}{N}}{\frac{n_{.1}}{N}\frac{n_{1.}}{N}} \quad (1)$$

Note that we use one entire document as the window of co-occurrence instead of a fixed number of words. Usually, co-occurrences are measured between two Kanji mainly because of computational and storage costs. We can use co-occurrence frequencies among any *n* Kanji. *n* Kanji co-occurrence tendency KCT_{MIn} among Kanji, k_1 , k_2 ...k_n is defined, as an extension of KCT_{MI2}, as follows:

$$KCT_{MIn}(k_1, k_{2,\ldots}k_n) = \frac{1}{n-1}\log_2 \frac{\frac{f(k_1,k_{2,\ldots}k_n)}{N}}{\frac{f(k_1)}{N}\frac{f(k_2)}{N}\ldots\frac{f(k_n)}{N}}$$

where N is the total number of documents in a document collection (corpus), f(k) is the number of documents where the Kanji k occurs, and f(k_1, k_2, ...k_n) is the number of documents where all Kanji, k_1 , k_2, ... k_n appear. Note that *MI* is essentially a measure between two events, so this is an ad hoc extension only for the purpose of calculating *n*-Kanji co-occurrence tendency. We will only calculate *MI* for Kanji pairs to select highly correlated Kanji pairs for indexing.

We would like to add that other co-occurrence tendency measures, for example, dice co-efficient, log likelihood ratio and Chi-square test [10, 25] are also applicable to calculate KCT-like measures.

6.2 Latent Semantic Indexing (LSI) and Kanji Semantic Indexing (KSI)

Like Latent Semantic Indexing (LSI), KSI begins with a collection of *m* documents containing *n* unique Kanji (i.e., terms) and KSI forms an *n x m* sparse matrix *A*, with A_{ij} containing a value related to the number of times Kanji *i* appears in document *j*. Various weighting schemes can be applied to the raw occurrence counts. In this work, we used log-entropy weighting.

Once the Kanji-document matrix *A* has been created, KSI computes the similarity between two text objects (a query and a document, for example) as follows. First, a text object *q* is represented by an *n x 1* vector, much like a column of the *A* matrix and with the same sorts of term weighting applied. Next, the similarity between text objects, q_1 and q_2 can be computed, typically by cosine scoring in the vector space model. This similarity can be represented as, **sim** $(q_1, q_2) = q_1^T \quad q_2 / \sqrt{(q_1^T q_1 \,.\, q_2^T q_2)}$.

A mathematically useful way of viewing the process of computing text-object similarity scores in the vector space model is: (1) Each of the *n* Kanji in the collection has a vector representation. Specifically speaking, Kanji *i* is an *n x 1* vector of zeros with a *1* in component *i*; (2) The representation of a text object, *q* is the weighted sum of the Kanji vectors of all the Kanji that appear in the text object. Thus, the similarity between text objects, q_1 and q_2 is:

$$\mathbf{sim}\,(I_n q_1\,, I_n q_2\,) \qquad (2)$$

where I_n is the *n x n* identity matrix. Here, I_n plays the role of a *vector lexicon*, in the sense that it assigns each Kanji a vector definition. Of course, pre-multiplying by the identity matrix in the Eq. 2 does not change the comparison in any way. On the other hand, by using other vector lexicons, we can substantially change the way similarities are computed. In addition, the only role played by the Kanji-document matrix *A* in the vector space model is in the computation of weighting factors for the components (i.e., Kanji or terms) of text objects.

KSI, like LSI, is a vector space IR formalism. KSI begins with the formation of the Kanji-document matrix *A*. Then, the *A* matrix is analyzed using singular value decomposition (SVD) to extract structure concerning document-document and Kanji-Kanji correlations. Mathematically, an SVD of *A* can be written as,

$$A = \mathrm{U}(A)\,\Sigma(A)\,\mathrm{V}(A)^{\mathrm{T}} \qquad (3)$$

where U(*A*) is an *n x n* matrix such that $\mathrm{U}(A)^{\mathrm{T}}\,\mathrm{U}(A) = I_n$, $\Sigma(A)$ is an *n x n* diagonal matrix of singular values, and V(*A*) is an *n x m* matrix such that $\mathrm{V}(A)^{\mathrm{T}}\,\mathrm{V}(A) = I_m$. This assumes for simplicity of exposition that *A* has fewer terms than documents, $n < m$.

This SVD analysis can be used to construct lower rank approximations of *A*, and this is how it is

typically used in the context of LSI. Reducing the rank of the approximation results in a synonym collapsing effect in practice. Such a reduction also lessens the total amount of processing and storage overheads associated with preprocessing and retrieval. We use A_k to denote the components of the k-dimensional SVD, and the rank-k reconstruction of A as follows:

$$A_k = \mathrm{U_k}(A)\, \Sigma_k(A)\, \mathrm{V_k}(\mathrm{A})^{\mathrm{T}} \qquad (4)$$

The $\mathrm{U_k}(A)$ matrix in Eq. 4 can be used as an alternative vector lexicon to the I_n in Eq. 2 in that it assigns a vector representation to every Kanji in the Kanji-document matrix A. Thus, in KSI, the k-dimensional similarity between text object q_1 and text object q_2 in the context of A is,

$$\mathbf{sim}\,(\mathrm{U_k}(\mathrm{A})^{\mathrm{T}} q_1\,,\, \mathrm{U_k}(A)^{\mathrm{T}} q_2) \qquad (5)$$

In the LSI literature [7, 9, 11, 3, 36] justifications for the use of the matrix of left singular vectors $\mathrm{U_k}(\mathrm{A})$ as a vector lexicon are intensively studied.

Cross-language LSI (CL-LSI)

The techniques of monolingual LSI can be extended easily to the cross-language case simply by using a different notion of the term-document matrix. For concreteness, let E be a term-document matrix of m English documents and n^E English terms, and let F be a term-document matrix of m semantically equivalent French documents and n^F French terms. These documents are aligned pair-wise, in the sense that document $1 \le i \le m$ in the English collection is directly related to document i in the French collection. The multi-language term-document matrix M can be written as follows:

$$M = \begin{bmatrix} E \\ F \end{bmatrix}$$

M is an $(n^E + n^F)$ *x* m matrix in which column i is a vector representing the English and French terms appeared in the *union* of document i written in both languages. Cross-language LSI (CL-LSI) begins with the matrix M and performs an SVD,

$$M = \begin{bmatrix} U_k^E(M) \\ U_k^F(M) \end{bmatrix} \Sigma_k(M)\ V_k(M)$$

where $\mathrm{U_k^E}\,(M)$ and $\mathrm{U_k^F}\,(M)$ are k-dimensional vector-lexicons for English and French, respectively. Empirically, similar English and French words are given similar definitions, so this vector lexicon can be used for cross-language retrieval. In particular, consider an English text object $\mathrm{q^E}$ and a French text object q^F. They can be compared using the obvious generalization of Eq. 5,

$$\mathbf{sim}\,(\mathrm{U_k^E}(M)^{\mathrm{T}}\, \mathrm{q_E}\,,\, \mathrm{U_k^F}(M)^{\mathrm{T}}\, \mathrm{q_F}) \qquad (6)$$

In our experiments, we chose Japanese and Chinese, and represented the Japanese and Chinese document and query using Unicode, where each Kanji is equivalent to a term. Common or semantically

similar Kanji are considered as cross-language homonyms. Differences in Kanji usage across the languages are captured through SVD decomposition. We refer to such representation and indexing as *cross-language Kanji Semantic Indexing*.

7. Experimental Setups and Results

7.1. Document Collection and Queries

The most popular IR Test Collections in Japan are the NTCIR collection and the BMIR-J2 collection. The NTCIR collection consists of a collection of 330,000 English and Japanese scientific articles. Half of the collection (187,081 documents) consists of bilingually aligned English-Japanese document pairs. This collection was not suitable for our experiments since we want to experiment with Chinese as well. The BMIR-J2 Test Collection is a Japanese text collection of 5,080 newspaper articles chosen from the Mainichi newspaper. However, due to the participants' interests, this collection is restricted to the Engineering and Economics domains only. We could use this collection along with a set of corresponding Chinese document collection (not strictly parallel, but comparable counterparts). However, locating corresponding Chinese articles in the Engineering and Economics domains using Search Engines or Internet Robots is a tedious task.

Another international test collection, the TREC test collection includes English-Chinese and English-Japanese parallel corpora but no test collection of Japanese-Chinese parallel documents is yet available. Since we want to experiment with Japanese and Chinese IR and since there is no Japanese-Chinese bilingual test collection available so far, we took the initiative to prepare a bilingual test collection for our use. Nevertheless, we adhered to the TREC and NTCIR guidelines as strictly as possible. For the ease of locating corresponding Chinese documents, we restricted ourselves to current international affairs. We will explain the details of our collection preparation procedures below.

Mainichi newspaper is publishing their newswire archive on the CD-ROM on a yearly basis since early 90s. First, we used full-text search engines to index the most recent archives of the Mainichi Shimbun newspaper articles. Initially, we constructed 50 initial queries by selectively examining the collection. We used the freely available full-text search tools (NAMAZU and FREYA) to retrieve documents (likely to include both relevant and non-relevant) from a selected portion of the most recent Mainichi Newspaper archives [1995-1999]. For each query, we retained the top 20 documents retrieved by each search engine. By merging a few hundred documents from each search engine, we obtained a collection of about 1,600 documents. We carefully investigated the 224 articles retrieved by both the search tools against our 50 initial queries. Finally, we settled on a set of 33 *revised* queries and 1,000 news articles. After deciding on a set of documents and queries, we re-indexed this collection of 1,000 articles

and used the polling method[6] to prepare the query-relevance matrix. Our polling process is highly approximated because we used the output results of only two systems validated by a single human evaluator. The accuracy of the query-relevance matrix can be further improved by employing more human evaluators.

After deciding on the set of Japanese queries and documents, we translated the queries into Chinese and used Internet search engines to retrieve and choose a collection of 1,000 Chinese documents using advanced search features provided by AltaVista [1]. Because our Japanese document collection mostly consists of articles about current and International affairs in recent years, it was easier to locate similar Chinese articles on the Internet. Moreover, the AltaVista search engine facilitates restricted-search within a particular domain. By restricting us to news sites, we could quickly extract a collection of Chinese documents comparable to the Japanese document collection. We again used the polling strategy to compute a query-relevance matrix for this collection with the help of two search tools and one human evaluator.

7.2. Retrieval Results for Monolingual IR

We convert the entire bilingual collections (including queries) into Unicode. Necessary preprocessing (e.g., Kanji mapping) of the document collection is also done prior to indexing. We use a modified version of the publicly available *mg* System [45] developed as part of the New Zealand digital library (NZDL) project to index our document collection in 3 different ways:

1. K_A, Single Kanji indexing
2. K_B, Kanji bi-gram indexing, and
3. K_C, Correlated Kanji pair indexing

We also use the LSI++ [23] package for singular value decomposition of the Kanji-document matrices (computed with the log-entropy weighting of single Kanji), and investigate the latent Kanji semantic retrieval.

4. K_D, reducing the dimension of the Kanji-document matrix using SVD

We use the TREC evaluation scripts (TREC-EVAL) to compute the non-interpolated average precision for the 33 queries. The monolingual retrieval results are listed in Table 2, for both Chinese and Japanese information retrieval.

In Table 2, the average precision is very low for single character indexing and bi-gram indexing. In general, the average precision for our Japanese and Chinese monolingual IR are far below the average precision level achieved by the TREC, NTCIR and BMIR-J2 participants. The poor average precision is

[6] The polling method is described in [44] and used by TREC, NTCIR and other test administering authorities.

because of the fact that we do not incorporate classical IR enhancement mechanisms such as, query expansion and relevance feedback. The single character indexing and bi-gram indexing approaches are the simplest approaches compared to the extensive linguistic and computational techniques employed by the NTCIR and BMIR-J2 participants. Since our focus is to investigate the effectiveness of a Kanji-based representation, we refrain from enhancing the retrieval procedures through complex mechanisms so that we can investigate the true effect of such representation and indexing. For the same reason, direct comparison of the results of our experiments with those of others is not possible.

Table 2. *Average precision for Japanese and Chinese Monolingual IR*

Indexing Method	**Non-interpolated Average Precision**	
	1,000 Japanese documents & 33 Queries	1,000 Chinese documents & 33 Queries
K_A (Single Kanji)	.1435	.1838
K_B (+Kanji Bi-gram)	.1626	.2253
K_C (+co-occurrence, MI)	.1757	.2482
K_{D30} (log +SVD, k=30)	.1505	.2037
K_{D100} (log +SVD, k = 100)	.1870	.2528

By incorporating extra computation efforts for the singular value computation (i.e., without linguistic analysis), we achieve a significant boost in the average precision for the case of a relatively larger value of k (k = 100 to 300). A single Kanji itself is ambiguous but when groups of Kanji are mapped into a reasonable number of concepts, the latent concept of the query and the documents matches efficiently.

For a smaller value of k (k = 30), we perform some error analysis and discover that the performance degradation is severe for the short-queries, for which a large number of non-relevant documents are also retrieved with high ranking. Kanji semantic indexing (KSI) may perform better with a proper mechanism of query expansion for the short queries. Another potential problem with KSI is that the parameter, k, may have to be adjusted depending on the nature of the collection. For our case, the retrieval results with k=30 and 100 are chosen empirically. For k value within 100 to 300, we obtain stable retrieval performance.

The overall Chinese retrieval results are better than those for Japanese, perhaps due to the homogeneous Chinese scripts. For Japanese IR, a plausible way of improving the retrieval efficiency is to put more effort in Katakana disambiguation and Hiragana to Kanji conversion.

7.3. Retrieval Results for Cross-language IR

In this section, we discuss three different retrieval results. First, we use Japanese queries to retrieve Chinese documents. Second, we use Chinese queries to perform retrieval from the Japanese collection. Note that in both experiments neither query translation nor document translation is performed. Documents are retrieved in terms of Kanji correspondence. The third experiment is performed under a special condition where

commercial MT system is used to translate the Japanese and Chinese documents and a pseudo-query expansion mechanism (described later in this section) is employed.

From the CLIR results shown in the 2nd and 3rd columns of the Table below (Table 3), it can be concluded that the CLIR results using only Kanji mapping and associations are not prohibitive for the K_A and K_B, where single Kanji and the Kanji bi-grams are the basis of indexing and retrieval. However, the retrieval results are very promising for the K_C and K_D, where Kanji correlation and Kanji association are the basis of indexing and retrieval. Theoretically, it can be said that the retrieval results can be further improved if the Kanji correlation and the Kanji association are estimated from a large collection of documents.

The 4th column of Table 3 shows the retrieval result under a special situation. We use a commercial MT system (Chinese-Japanese/Japanese-Chinese) from Kodensha [20] to translate the Chinese document collection into Japanese, and vice versa. This MT system uses a basic bilingual dictionary of 120,000 Japanese-Chinese and 220,000 Chinese-Japanese entries. After the translation, we append the translated documents with the respective originals. In this way, we have a bilingual document collection of 2,000 documents. Since we are only considering Kanji and Kanji derived information in our indexing process, we assume that the quality of the machine translation (in terms of readability, syntax, etc.) has trivial effects. We also merge the corresponding queries in Chinese and Japanese to obtain the pseudo- *query translation* and pseudo- *query expansion* effects. Please note that the Kanji semantic retrieval, K_D is based on log-entropy weights and the other three approaches are based on the *tf.idf* weighting scheme.

Table 3. *Average Precision for Japanese and Chinese Cross-Language IR*

Indexing Method	**Non-interpolated Average Precision**		
	1,000 Japanese documents & 33 Chinese queries	1,000 Chinese documents with 33 Japanese queries	2,000 bilingual documents with 33 merged queries
K_A (Single Kanji)	.1033	.1398	.1241
K_B (+Kanji Bi-gram)	.1244	.1569	.1348
K_C (+co-occurrence, MI)	.1656	.1972	.2024
K_{D30} (+SVD, k = 30)	.1547	.1780	.2326
K_{D100} (+SVD, k = 100)	.1622	.2016	.2537

The non-interpolated average precisions of document retrieval using this approach with 33 merged queries and 2000 bilingual documents are listed in the 4th column of Table 3. The bi-gram based method (K_B) suffers from low precision. This is possibly due to MT-related errors. We assume that a properly aligned Japanese-Chinese parallel or comparable corpus may boost the bi-gram based retrieval results as well as the single Kanji based retrieval results (K_A). The co-occurrence based method (K_C) and Kanji association based method (K_D) perform better with the translated bilingual documents and merged queries.

From the above scenario of Kanji-based monolingual and cross-language information retrieval of Japanese and Chinese, we can safely conclude that by estimating Kanji correlation and Kanji association

from a large parallel (or comparable) corpus, it is possible to formulate effective Japanese-Chinese mono- and cross- language IR.

8. Discussions

In this paper, we explored one of the few possibilities of cross-language IR research with Asian languages. We experimented with Japanese and Chinese, where Kanji play an important semantic role; and we demonstrated that mono- and cross- language IR can be performed effectively through Kanji associations. In our experiments, we deliberately used Kanji and Kanji-derived semantics to address Japanese and Chinese IR (including CLIR). It is worthy to mention here that we do not advocate abandoning linguistic enhancements (e.g., segmentation, morphological analysis, etc.) and classical IR enhancements (e.g., query expansion, relevance feedback, etc.) techniques for IR tasks. Our exclusive attention to Kanji in our experiments is to identify the role of Kanji semantics in IR and CLIR. This approach can easily accommodate other linguistic and IR enhancements, and with such enhancements, the proposed approach will eventually give birth to practical CLIR systems.

Several types of ambiguities with Kanji usage across the Japanese and Chinese languages [15] exist. Such ambiguities contribute highly to the lower precision in single Kanji oriented indexing and retrieval. For bi-gram based indexing, we cannot conclude anything with high confidence due to the small document collections used in our experiments because of the data sparseness problem. However, the average precision of mono- and cross-language IR with KCT and with SVD shows that Kanji based indexing and retrieval of these ideographic languages is effective. Unlike other IR research reports where the IR task is comprehensively addressed, our experiments involves with only a straightforward hypotheses. Because of our exclusive focus on Kanji semantics and Japanese-Chinese language pair, we could not make direct comparison of our experimental results with those of the others'.

For Japanese-Chinese CLIR, this is one of the very first reports. The indexing methods we tried inherently bypass the complicated segmentation and morphological analysis phases, which would otherwise be necessary. Nonetheless, incorporating such linguistic analyses with the proposed approach will certainly improve the retrieval results. We are also aware that query expansion and relevance feedback -based enhancements can also be easily incorporated with the proposed Interlingua framework since this framework uses a flexible vector space representation. Moreover, Kanji association makes cross-language IR simpler than the traditional MT-based approach. We mapped Katakana and Hiragana strings to their relevant Kanji using an ad-hoc approach. During the error analysis, we noticed that such an approximated mapping significantly contributed to erroneous retrieval. Accurate mapping of Katakana and Hiragana strings to Kanji can further boost the retrieval effectiveness.

Since words in the non-ideographic languages can also be mapped to their original roots, the proposed Kanji-based Interlingua framework can be extended to deal with multilingual information indexing and retrieval of any combination of languages as far as parallel (or comparable) corpora and

sufficient computing power are available. Effective processing of multilingual heterogeneous information in this Internet age is inevitable. Revolution in storage capacity, abundance in computing power and invention of sophisticated mathematical methods for dimensionality reduction and projection (e.g., [23]) may present us with a better opportunity to integrate alphabetic and ideographic languages equally effectively under a uniform Interlingua representation. Indexing and retrieving heterogeneous multilingual information in a unified manner will therefore be possible. Traditional lexical (or Boolean) IR techniques will continuously be less effective as the number of documents grows. Multilingual IR is inevitable because of the global connectivity and the proliferation of electronic information. Automated and conceptual IR techniques will therefore dominate the future IR research.

Acknowledgements

We would like to extend our thanks to the people involved with NZDL, NAMAZU and FREYA projects. Their tools helped us to speed up our research. Thanks to Dr. Akira Maeda for allowing us to use his correlation calculation tool and Dr. Michael Berry for the LSI++ and SVDPACK packages. We thank the anonymous reviewers for their valuable comments.

References

[1] ALTAVISTA, "Altavista Advanced Search Tutorial", http://doc.altavista.com/adv_search/ast_toc.html

[2] AMF, "Cross-Language Information Retrieval at AMF - For Overcoming the Language Barrier in the Use of Internet", Asian Multimedia Forum, http://www.ntt.co.jp/news/news99e/9902/990224a.html

[3] M. Berry and P. Young, "Using Latent Semantic Indexing for Multi-Language Information Retrieval", *Computers and the Humanities*, 29(6), pp. 413-429, 1995.

[4] A. Chen, J. He, L. Xu, F.C. Gey and J. Meggs, "Chinese Text Retrieval Without Using a Dictionary", *In Proceedings of the Conference on Research and Development in Information Rertrieval*, ACM SIGIR-97, pp. 42-49, 1997.

[5] H.H. Chen, C.C. Lin and W.C. Lin, "Construction of a Chinese-English WordNet and its application to CLIR", *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, IRAL-2000, pp. 189-196, 2000.

[6] K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics*, Vol. 16(1), pp. 22-29, 1990.

[7] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas and R.A. Harshman, "Indexing by latent semantic analysis," *In Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[8] A. Diekema, F. Oroumchian, P. Sheridan and E.D. Liddy, "TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French", *In Proceedings of the Seventh Text REtrieval Conference* (TREC-7), pp. 169-180, 1999. http://www.cindorsearch.com/

[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources", Behavior Research Methods, Instruments, & Computers, Vol. 23, pp. 229-236, 1991.

[10] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, 19 (1), pp. 61-74, 1993.

[11] D.A. Evans, S.K. Handerson, I.A. Monarch, J. Pereiro, L. Delon and W.R. Hersh, "Mapping Vocabularies Using Latent Semantics", In G. Grefenstette Ed., *Cross-Language Information Retrieval*, Kluwer Academic Publisher, pp. 63-80, 1998.

[12] FREYA, *Full-text Retrieval Engine for Your Archive,* http://www.ingrid.org/ja/project/freya/ (in Japanese).

[13] H. Fujii and W.B. Croft, "A comparison of Indexing for Japanese Text Retrieval", *In Proceedings of the ACM SIGIR-93*, pp. 237-246, 1993.

[14] G. Grefenstette, "The Problem of Cross-Language Information Retrieval", In G. Grefenstette Ed., Cross-Language Information Retrieval, Kluwer Academic Publisher, pp. 1-10, 1998.

[15] M.M. Hasan and Y. Matsumoto, "Chinese-Japanese Cross Language Information Retrieval: A Han Character Based Approach", *In Proceedings of the SIGLEX Workshop on Word Senses and Multi-linguality*, pp.19-26, ACL-2000, Hong Kong, 2000.

[16] J. Hochberg and D. Nix, "Vector Mapping CLIR with Character Trigrams", *Los Alamos National Laboratory (LANL) CLIR Project Notebook Paper*, 1999. http://citeseer.nj.nec.com/134994.html

[17] D. Hull, "Using Statistical Testing in the Evaluation of Retrieval Experiments", *In Proceedings of the ACM SIGIR-93*, pp.329-338, 1993.

[18] D.B. Kim, K.S. Choi and K.H. Lee, "A Computational Model of Korean Morphological Analysis: A Prediction-based Approach", *Journal of East Asian Linguistics*, Vol. 5(2), pp. 183-215, 1996.

[19] T. Kim, C.M. Sim, S. Yuh, H. Jung, Y.K. Kim, S.K. Choi, D.I. Park and K.S. Choi, "FromTo-CLIR™: web-based natural language interface for cross-language information retrieval", *Journal of Information Processing and Management*, Pergamon, Vol. 35(4), pp. 559-586, 1999.

[20] KODENSHA, *J-Pekin 2000: Japanese-Chinese*, Chinese-Japanese Twin Translation Software, Kodensha, Japan, 2000.

[21] K.L. Kwok, "Comparing Representation in Chinese Information Retrieval", *In Proceedings of the ACM SIGIR-97*, pp. 34-41, 1997.

[22] J.H. Lee, H.Y. Cho and H.R. Park, *"n-Gram-based Indexing for Korean Text Retrieval", Journal of Information Processing and Management*, Pergamon, Vol. 35(4), pp. 427-441, 1999.

[23] T.A. Letsche and M.W. Berry, "Large Scale Information Retrieval with Latent Semantic Indexing", *Information Sciences – Applications*, Vol. 100, pp. 105-137, 1997.

[24] K. Lunde, *CJKV Information Processing: Chinese, Japanese and Korean Computing*, O'Reilly & Associates, Inc., 1999.

[25] A. Maeda, "Studies on Multilingual Information Processing on the Internet", *PhD Thesis*, NAIST-IS-DT-9761021, Nara Institute of Science and Technology (NAIST), Japan, 2000.

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

[27] M.F. Mudawwar, "Multicode: A Truly Multilingual Approach to Text Encoding", *IEEE Computer*, Vol. 30(4), pp. 37-43, 1997.

[28] NAMZU, Namazu, *A Full Text Search Engine*, http://www.namazu.org/

[29] J.Y. Nie, M. Brisebois and X. Ren, "On Chinese Text Retrieval", *In Proceedings of the ACM SIGIR-96*, pp. 225-233, 1996.

[30] J.Y. Nie, J.P. Chevallet and M.F. Bruandet, "Between terms and Words for European Language IR and Between Words and Bigrams for Chinese IR", *In Proceedings of the Sixth Text REtrieval Conference* (TREC-6), pp. 697-710, 1998.

[31] J.Y. Nie and F. Ren, "Chinese Information Retrieval: using character or words?", *Journal of Information Processing and Management, Pergamon*, Vol. 35(4), pp. 443-462, 1999.

[32] J.Y. Nie, J. Gao, J. Zhang and M. Zhou, "On the Use of Words and N-grams for Chinese Information Retrieval", *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, IRAL-2000, pp. 141-148, 2000.

[33] D.W. Oard and B.J. Dorr, "A Survey of Multilingual Text Retrieval", University of Maryland, *Technical Report*, UMIACS-TR-96-19, CS-TR-3615, 1996.

[34] Y. Ogawa and T. Matsuda, "Overlapping Statistical Word Indexing: A New Indexing Method for Japanese Text", *In Proceedings of the ACM SIGIR-97*, pp. 226-234, 1997.

[35] PERGAMON, "Special issue on Information Retrieval with Asian languages", *Journal of Information Processing and Management*, Vol 35. No.4. Pergamon, London, 1999.

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

[37] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[38] SANSEIDO, Sanseido's *The Unicode Kanji Information Dictionary*, Sanseido, Japan, 2000.

[39] M. Shibatani, *The Languages of Japan,* Cambridge Languages Surveys, Cambridge University Press, 1990.

[40] R. Sproat, C. Shih, W. Gale and N. Chang, "A Statistic Finite State Word-Segmentation Algorithm for Chinese", *Computational Linguistics*, Vol. 22 No. 2, pp. 377-404, 1996.

[41] C.L. Tan and M. Nagao, "Automatic Alignment of Japanese-Chinese Bilingual Texts", *In IEICE Transactions of Information and Systems*, Japan. Vol. E78-D. No. 1, pp. 68-76, 1995.

[42] TREC-6, *Proceedings of Text REtrieval Conference (TREC-6)*. National Institute of Science and Technology (NIST), 1998. http://trec.nist.gov/pubs/trec6/t6_proceedings.html

[43] UNICODE, *The Unicode Standard, Version 3.0,* Addison Wesley, Reading, MA, 2000. http://www.unicode.org/

[44] E.M. Voorhees, "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," *In Proceedings of the ACM SIGIR-98*, pp. 315-323, 1998.

[45] I.H. Witten, A. Moffat and T.C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Second Edition, Morgan Kaufmann Publishers, 1999.