Phrase Processing Methods for Japanese Text Retrieval

Noriko Kando Kyo Kageura Masaharu Yoshioka Keizo Oyama

Research and Development Department

National Center for Science Information Systems (NACSIS)

Tokyo, Japan

Email: {kando,kyo,yoshioka,oyama}@rd.nacsis.ac.jp

URL: http://www.rd.nacsis.ac.jp/~{kando,kyo,yoshioka,oyama}/

ABSTRACT

This paper examines the effectiveness of different phrase identification and weighting methods for Japanese text retrieval in an operational information retrieval (IR) system, called *NACSIS-IR*. Based on our previous experiments, we used character-based indexing with positional information and word- or phrase-based query processing, which allowed us to implement sophisticated linguistic analysis on large-scale databases while maintaining adequate efficiency. The results of retrieval experiments on a large-scale Japanese test collection showed that the combination of enhanced phrase identification using patterns defined over part-of-speech tags and our algorithm *Weight5* made a significant positive contribution to retrieval effectiveness. The paper also discusses indexing and phrase processing of Japanese or East Asian languages.

Keywords

Phrase processing, Japanese text retrieval, character-based indexing, word-based query segmentation

1. Introduction

This paper reports the results of retrieval experiments evaluating several phrase processing methods over a large-scale Japanese test collection called *NACSIS Test Collection 1*, which contains more than 330,000 documents with more than half of them English-Japanese paired (Kando et al. 1998). Our main purpose is to determine what phrase processing methods are usable in the new version of *NACSIS-IR*, which is an operational IR service system providing fee-based access to a wide range of Japanese and English scientific bibliographic or fulltext databases of various subject domains for academic users (NACSIS, 1997).

1.1 Phrases in Information Retrieval

As a unit of information, a phrase is more content-bearing and precise than the sum of its constituent words. Thus, phrase-based representations of document and query content can lead to improvements in retrieval effectiveness. Although phrase-based representations may cause matching difficulties during retrieval, for the most part a phrase-based representation, coupled with a word-based representation, is seen as a worthwhile feature for inclusion in an IR system (Smeaton 1998).

However, the effects of phrase-based representation on retrieval effectiveness as reported in recent studies are not consistent: linguistic phrase identification has been noted as improving retrieval effectiveness (Fagan 1987); for a large collection, proper phrase weighting improved retrieval effectiveness (Jing & Croft 1994; Beaulieu et al. 1996; Xiangji & Robertson 1997); on the other hand, Mitra et al. (1997) reported little improvement; Smeaton (1998) reported that the benefit of phrase-based representation varied with users. Fewer experiments have been reported with Japanese texts. It is partly due to the well-known difficulties of automatic indexing with Japanese texts where there is no explicit separator between words, and partly because of the lack of large-scale standard test collections in the case of Japanese texts.

1.2 Problems of Japanese Operational IR systems

Addition to the above, Japanese operational IR service systems face problems as follows;

1. **A trade-off between effectiveness and efficiency**: both retrieval effectiveness and efficiency are critical, and retrieval must be done in a reasonable time,
2. **Database size** can be several tens of gigabytes or more,
3. **Multiple-file searches**: users often search multiple files at a time,
4. **Heterogeneity of documents**: a file may contain various types of documents and a record contains various fields,
5. **Word segmentation of Japanese texts**: there is no explicit word boundaries in a sentence, and
6. **Language-specific indexing**: Japanese scientific databases are substantially multilingual: in Japanese scientific documents, technical terms or important concepts are often represented as English or other

foreign language terms, in their original spelling, and many document collections contain both English and Japanese texts (Kando 1997). Indexing, stopping, and stemming are language-dependent procedures. Therefore language-specific indexing procedures are needed.

In order to cope with these problems, after examining the indexing and query processing methods suitable for Japanese scientific documents, we adopted language-specific strategy of indexing (*i.e.*, character-based indexing with Japanese text and word-based with English text, including English terms appearing in Japanese texts. The index retains positional information) and word-based query segmentation with phrase processing for both English and Japanese queries, and tested its effectiveness on a large-scale test collection. This strategy takes advantages of the efficiency and simplicity of character-based indexing, and also allows sophisticated query weighting based on linguistic analysis of the query, which is expected to lead to higher precision with reasonable computational cost on practical large-scale databases. Moreover the query processing method can be modified to be more effective ones without re-indexing the databases.

Although several English-based operational IR service systems like *Westlaw*, *Dialog, Nexis-Lexis* and Internet search engines provide the function of relevance ranking and some of them accept natural language sentences as queries, the Boolean-based exact match retrieval model is still predominant in Japanese operational IR service systems. This is mostly because of the problems of Japanese operational IR systems mentioned above.

Among them, the new version of *NACSIS-IR* will be the first Japanese operational IR service system which provides the function of *relevance ranking* from *natural language sentence queries* (NLQ). Automatic phrase identification and proper weighting of phrases is one of the integral part of our algorithm of NLQ processing.

1.3 Organization of this Paper

In the following Section 2 is a brief review of studies regarding indexing and query segmentation of Japanese texts. Section3 describes the method of indexing and query processing and the collection we used. In section 4, we present the results of experiments comparing the effectiveness of several phrase-based approaches on a large-scale collection. Further directions for investigation are discussed in section 5.

2. Review: Indexing of Japanese Texts

There are three approaches to segmenting Japanese or East Asian language texts into indexing units: (1) *n*-gram indexing and *n*-gram query segmentation (Fujii & Croft 1993; Ogawa & Iwasaki 1995; Chien 1995; Lee & Ahn 1996; Buckley, Singhal & Mitra 1996); (2) word-based indexing and word-based query segmentation using morphological analysers (Fujii & Croft 1993) or a large-scale dictionary (Nie et al. 1996; Xiangji & Robertson 1997); and (3) *n*-gram indexing and word-based query segmentation (Fujii & Croft 1993; Xiangji & Robertson 1997).

2.1 Word-based Approaches:

The unit of "word" in the Japanese texts are substantially ambiguous. Formation of Japanese compound nouns have been investigated using both statistic and symbolic methods (Kageura 1996) and it is still challenging issue. The index unit can be set small as a morpheme, or large as a whole compound term(Kageura et al. 1997; Kando et al. 1998; Koyama et al. 1998).

For example, in Japanese, a term “university librarian” is expressed as a string of 6 characters: the first two characters represent “university”, next three characters are “library”, and the last character is “stuff”. Moreover the string consists of the last two characters means “stuff of institutes like libraries or museum”. Which segmentation is sementically or lexically correct? How IR systems should segment the term? And how users are expected to input the query terms?

In spite of such problem, word-based indexing has been predominant in the Japanese traditional operational IR service systems. To avoid the word mismatch caused by the difference in word segmentation of systems and users' input, many systems adopts duplicated, overlapping indexing, where a compound term is segmented as various as possible and all the possible constituent words are collected in the index. This is rather expensive and problematic both in effectiveness and efficiency. It also makes the procedure of statistical weighting complex. To utilize the phrase longer than a compound term, positional information is needed.

Example of duplicated word-based indexing often used in the Japanese operational IR service systems:

In Japanese, a term “university librarian” is represented as a string of 6 Kanji characters, $C_1\,C_2\,C_3\,C_4\,C_5\,C_6$.

--> Index as: $C_1\,C_2\,C_3\,C_4\,C_5\,C_6$ (university librarian), $C_1\,C_2\,C_3\,C_4\,C_5$ (university library), $C_3\,C_4\,C_5\,C_6$ (librarian), $C_1\,C_2$ (university), $C_5\,C_6$ (stuff (of an institute like library or a museum)), C_6 (stuff)

2.2 *N*-gram approach:

N-gram indexing can avoid the such problem of segmentation and promising in Japanese text indexing.

However, uni-gram indexing and uni-gram query segmentation is less representative and tends to show the low precision. Bi-gram or tri-gram index can be very large since more than 6000 types of characters are used in Japanese language and the combination of each of them makes huge number of index entries (Ogawa&Matsuda, 1995). To achieve both effectiveness and efficiency in uni-gram indexing and query segmentation, the approach of incremental n-gram is proposed. The ordinary type of n-gram indexing is without positional information, phrase weighting can not be incorporated

2.3 Combination approach:

There is another type of *n*-gram indexing ; *n*-gram index with positional information. It enables us to match against strings longer than *n* in the texts.

Since regarding Japanese text retrieval, experimental environments reported in the literature vary, we cannot easily decide which approach is superior to the others. However, it has been reported that a combination of uni-gram indexing and word-based query segmentation is more effective than *n*-gram indexing and *n*-gram query or word indexing and word query in experiments on large collections (Fujii & Croft 1993; Xiangji & Robertson 1997).

Our previous experiments also showed that *n*-gram query segmentation is less effective and less efficient than word-based query segmentation. Moreover, a cross-linguistic approach is critical in Japanese scientific or scholarly information retrieval (Kando 1997), and segmenting into words is necessary to allow utilization of a multilingual dictionary or other multilingual lexical resources that are necessary for cross-language IR. We therefore chose the approach of ***character-based indexing,*** and ***word- and phrase-based query processing***..

3. Indexing, Query Processing and Collection Used

3.1 Indexing and overall strategies

In this section, we briefly describe the method used in the experiments reported here. As stated above, feature of our approach is language-oriented strategies of indexing and stopping: *i.e*., character-based, or uni-gram, indexing with Japanese text and word-based with English text, including English terms appearing in Japanese texts. The index retains positional information (offset) of each index entry in the texts.

It is combined with word- and phrase-based query term weighting strategy for both English and Japanese queries. It is expected to lead to higher precision with reasonable computational cost on practical large-scale databases. The overhead of morphological or syntactical analysis of queries is much smaller than that for the overall texts.

3.2 Query Processing

When queries are natural language sentences, they were processed in four steps: (1) segmenting into words, (2) selecting words and phrases which are used as query terms, (3) constructing a structural representation of phrases using operators, and (4) weighting. In the following, we describe the methods used in this paper.

3.2.1 Segmentation:

Natural language queries are firstly segmented into words using a Japanese morphological analyzer, *Juman v.3.5* (Nagao & Kurohashi 1997). In order to improve efficiency, we developed a server mode version of *Juman* with permission from and in collaboration with Dr. Kurohashi of Kyoto University. The minimum units used are constituent words of compound nouns.

3.2.2 Word selection and phrase identification:

In this paper, a phrase-based representation is used coupled with a word-based representation for query processing. After discarding the stop phrases (e.g. article, document, I'd like to have, I'm interested in, What is …, etc.), words and phrases are selected based on Basic or Enhanced procedure as follows. Both of them use the patterns defined over part-of-speech tags and syntactic heuristics:

Basic:
{prefix.adjective..noun+.suffix}

Enhanced:
{prefix.noun. noun+.suffix}

{(prefix|adverb).(noun.functionword|.noun|adjective.suffix) +.noun+.suffix}

where words enclosed in a '{ }' are constituents of a phrase, '+' indicates the component should be occurred at least once, and any component without '+' can be omitted. 'noun+' indicates a concatenation of one or more nouns, and '\' indicates either component divided by '|' in the parentheses should be matched. Unknown terms are treated as nouns since such terms are mainly proper nouns. For Enhanced procedure, verb stems are also treated as nouns The underlined elements selected only as constitutes of phrases and are not selected as words.

"Basic" is one of the common automatic indexing strategies of Japanese texts, and "Enhanced" uses the enhanced method of selecting words and phrases based on the linguistic consideration which is proposed in this paper.

3.2.3 Structural representation of phrases

A phrase can be represented in various ways. For example, a phrase consisting of the words t_1, t_2, t_3 is matched to strings in the texts such as:

(a) ' t_1 t_2 t_3 ', all the words are directly adjacent in Japanese or adjacent but separated only by spaces in English;

(b) ' t_1 (3) t_2 (3) t_3', words t_1, t_2, t_3 appear in this order but the maximum distance between words is 3 characters in Japanese or 3 words in English;

(c) ' t_1 (3) t_2', ' t_2 (3) t_3', pairs of words occur in the specified order within a maximum distance of 3;

(d) ' t_1 and t_2 and t_3', all three appear in the same paragraph but the order is unspecified, and

(e) ' t_1 or t_2 or t_3', at least one of the three appears in the text and the order is unspecified.

3.2.4 Weighting

Phrase processing methods used here are described as combination of methods of phrase identification (*3.2.2*) and structural representation of phrase construction (*3.2.3*).

Table 1 Phrase processing methods

	identification	word/phrase	phrase construction
BASE:	Basic	word only	
Weight0	Basic	word + phrase	(b) (d) (e)
Weight1	Enhanced	word + phrase	(a)
Weight2	Enhanced	word + phrase	(b)
Weight3	Enhanced	word + phrase	(c)
Weight4	Enhanced	word + phrase	(b) (e)
Weight5	Enhanced	word + phrase	(b) (d) (e)

Weights were firstly calculated for each query unit, *i.e.* word or phrase, in the query, and the weight of the query was sum of weights of every constituent query units. Examples of phrase-construction and weighting are shown below;

For example, a query sentence consists of 4 words of t_1, t_2, t_3, t_4, the string of " t_2._t_3. t_4" are identified as a phrase, where t_3 is a term only used in a phrase, and weighting scores are calculated for each query unit as representing using w( ).

BASE: w(t_1)+w(t_2)+w(t_4)

Weight0: w(t_1)+w(t_2)+w(t_4)+w(t_2 (3) t_3 (3) t_4)+w(t_2 or t_3 or t_4)+w(t_2 and t_3 and t_4)

Weight1: w(t_1)+w(t_2)+w(t_4)+w(t_2 t_3 t_4)

Weight2: w(t_1)+w(t_2)+w(t_4)+w(t_2 (3) t_3 (3) t_4)

Weight3: w(t_1)+w(t_2)+w(t_4)+w(t_2 (3) t_3)+w(t_3 (3) t_4)

Weight4: w(t_1)+w(t_2)+w(t_4)+w(t_2 (3) t_3 (3) t_4])+w(t_2 or t_3 or t_4)

Weight5: w(t_1)+w(t_2)+w(t_4)+w(t_2 (3) t_3 (3) t_4)+w(t_2 or t_3 or t_4)+w(t_2 and t_3 and t_4)

3.3 Collection

We used a test version of the *NACSIS Test Collection 1* (Kando et al. 1998). The collection contains more than 330,000 documents, 30 queries, and a relevance assessment for each query. Documents are abstracts of conference papers in various subject domains selected from the *NACSIS Academic Conference Database* (NACSIS, 1997). More than half of them are English-Japanese paired documents.

Queries are collected from researchers in the domains. Their formats are similar to the ones used in the TREC Collection (especially, the ones used for TREC2). A query consists of a title (1-3 words), a description (1 sentence), a detailed narrative and a list of concepts. A narrative may contain a detailed explanation of the topic, term definitions, background knowledge, purpose of the search, expected number of relevant documents, preference in text types, criteria of relevance judgement, and so on. We used only the description fields in the experiments reported here.

Relevance assessments have three grades: relevant, partially relevant, and non-relevant. 'Relevant' level assessments were used in this paper.

The search engine was OpenText 6 (OpenText Corp., Canada), which can handle both English and Japanese characters. The documents in the returned set were ranked using OpenText's 'RankMode Relevance1', which ranks the members of the returned set based on the frequency of the string in a region of the document, the total number of index tokens in the document, and the number of the documents which contain the string (OpenText, 1997). This is a variant of *tf-idf*, which is used by most IR systems.

4. Retrieval Results

The results are shown in Figure 1 and Table 2.

Regarding term selection from natural language queries, the enhanced methods (Weight1-5) gave improved search effectiveness over the baseline. Weight0 and Weight5 used the same phrase construction but different term selection. Comparison suggests the improvement occurred because of the term selection method. Weight5 showed a 9.4% improvement in search effectiveness over Weight0 on average.

On comparison of phrase structuring methods, Weight3, Weight5 and Weight2 showed improvements of 9.3%, 9.0%, 8.0% in the average of 11 points precision over the baseline. Of these, Weight3 did not show improvement in the low recall ranks. Thus, Weight5 and Weight2 are good candidates for the operational system since higher precision in upper rank documents is important in the operational setting.

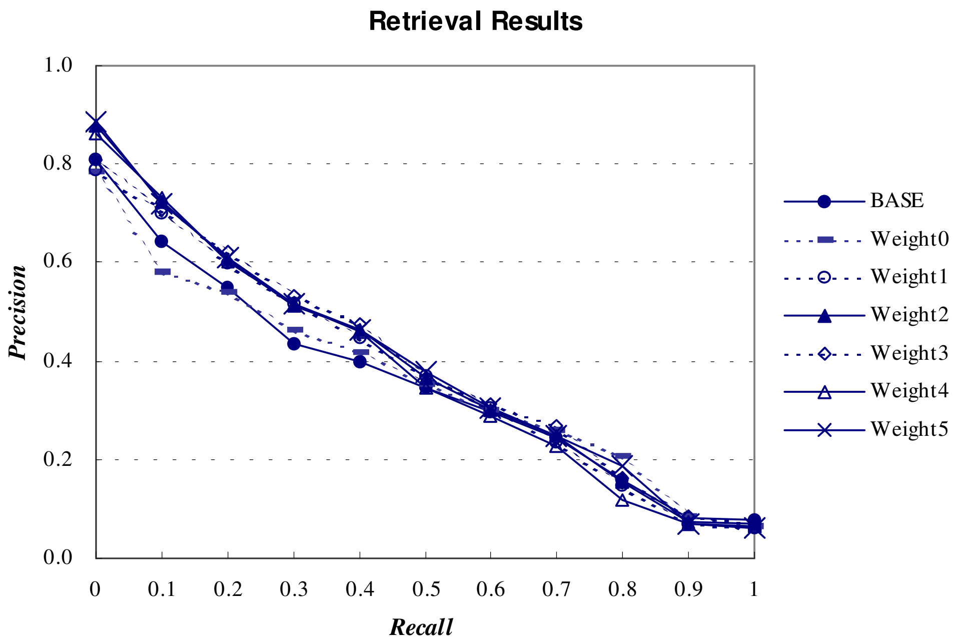

Figure 1: Retrieval Results with Relevance Assessment Level of 'Relevant'

Table 2: Retrieval Results

recall	**Base**	*weight0*	(%changed)	*weight1*	(%changed)	*weight2*	(%changed)	*weight3*	(%changed)	*weight4*	(%changed)	*weight5*	(%changed)
0	0.823	0.797	(-3.2)	0.809	(-1.7)	0.866	(5.2)	0.823	(0.0)	0.857	(4.2)	0.869	(5.6)
0.1	0.641	0.583	(-9.1)	0.699	(9.1)	0.715	(11.6)	0.707	(10.2)	0.719	(12.2)	0.706	(10.1)
0.2	0.547	0.543	(-0.9)	0.598	(9.3)	0.605	(10.5)	0.622	(13.6)	0.600	(9.7)	0.610	(11.5)
0.3	0.435	0.465	(6.8)	0.518	(18.9)	0.513	(17.9)	0.534	(22.7)	0.517	(18.7)	0.518	(19.0)
0.4	0.398	0.419	(5.0)	0.448	(12.4)	0.463	(16.1)	0.475	(19.3)	0.461	(15.8)	0.462	(16.0)
0.5	0.345	0.354	(2.4)	0.368	(6.6)	0.365	(5.7)	0.366	(5.8)	0.346	(0.1)	0.380	(10.0)
0.6	0.298	0.302	(1.4)	0.303	(1.8)	0.300	(0.8)	0.313	(5.3)	0.288	(-3.2)	0.305	(2.5)
0.7	0.245	0.260	(6.2)	0.237	(-3.1)	0.248	(1.3)	0.268	(9.4)	0.230	(-6.2)	0.248	(1.3)
0.8	0.158	0.206	(30.5)	0.148	(-6.2)	0.155	(-1.6)	0.165	(4.4)	0.117	(-25.6)	0.186	(18.3)
0.9	0.082	0.086	(4.7)	0.067	(-18.0)	0.073	(-10.7)	0.084	(1.9)	0.070	(-14.8)	0.069	(-16.3)
1	0.077	0.064	(-17.7)	0.063	(-18.7)	0.069	(-11.1)	0.069	(-11.5)	0.063	(-18.1)	0.062	(-20.4)
average	0.368	0.369	(0.3)	0.387	(5.2)	0.397	(8.0)	0.402	(9.3)	0.388	(5.4)	0.401	(9.0)
#query units	4.00	7.62		6.42		6.42		7.50		7.71		9.00	

Values enclosed in parenthesis indicate the percentage improvement of retrieval effectiveness over the baseline.

Moreover the average number of query units per a query is a good source to estimate the efficiency of the retrieval. Though the average precision of Weight2 is slightly less than the one of Weight5, the precision in higher ranks are almost same and average number of query units per a query in Weight2 is less than in Weight5. Therefore Weight2 is considered more promising.

5. Discussion

The experiments showed about 8.0 - 9.0% of improvement in retrieval performance in average by using automatic phrase identification using several patterns defined over part-of-speech tags and phrase weighting which are applied to the queries only. Though the improvements are moderate, the results indicate a promising direction. Generally, detailed linguistic analysis on the document collection is rather expensive, but detailed analysis and weighting based on queries only can be done with reasonable computational cost and is feasible in an operational setting.

Moreover our strategy of simple indexing with positional information and enhanced query processing takes an advantage that the query processing methods can be modified without re-indexing the whole databases. Therefore we can rather easily enhance the system's functionality continuously.

Natural language sentence queries and relevance ranking on operational large-scale Japanese databases are just beginning in an operational setting. In an operational database retrieval service system, there are other issues to be considered besides automatic phrase identification and weighting. For example, a record consists of various different elements such as title, keywords, abstract, etc. The difficulty lies in combining the evidence from such different elements. Cross-language IR (CLIR) is also vital in scientific information retrieval in Japan, since a data collection contains both English and Japanese documents, and a Japanese document contains many English technical terms in the original spelling or transcribed forms. A concept may be represented in Japanese or English or transcribed forms, and word mismatches may be caused by these variations in concept representation.

Regarding CLIR, several approaches including CLIR using automatically constructed mulilingual keyword clusters (Aizawa et al, 1998; Kando at al, 1998b), examination of correspondence between components of English and Japanese compound terms have been investigated (Tsuji et al 1997). Based on the users' responses and usage of the systems, we will investigate these problems in a continual process of improvement of the system.

In addition to this, we intend to open up part of our databases of Japanese scientific information as a test collection for the purpose of IR and related NLP researches (Kageura et al, 1997; Kando et al, 1998) since there is so far no large-scale resource of Japanese scientific documents available to the general IR research community. We hope this will facilitate further investigation and discussion of various issues related to IR with Japanese documents, especially Japanese scientific or academic documents, and also will facilitate cross-language IR, taking advantage of the practical scale of the database.

ACKNOWLEDGMENTS

Our thanks to Dr Kurohashi, Kyoto University, for permitting us to modify *Juman* they had developed.

This research is partly supported by "Research for the Future" Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science.

REFERENCES

Aizawa, A. ; Kageura, K. 1998a, "An approach to the automatic generation of multilingual keyword clusters". COMPUTERM '98, Aug. 1988. Montreal, Canada

Beaulieu, MM, Gatford, M; Huang, X.; Robertson, SE; Williams, P. 1996, "Okapi at TREC-5" In Proceedings of the Fifth Text Retrieval Conference (TREC-5). NIST Special Publication 1996.

Buckley, C; Singhal, A; Mitra, M. 1996, "Using query zoning and correlation within SMART: TREC-5". Paper presented at the Fifth Text Retrieval Conference (TREC-5). NIST Special Publication

Chien, LF., 1995, "Fast and quasi-natural lnaguage search for gigabits of Chinese texts." Paper presented at the 18th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Seatle, WA

Fagan, JL., 1987, "Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods". In Proceedings of the 10th ACM-SIGIR Conference on Research and Development in Information Retrieval, New Orleans, June 1987, p.91-108.

Fujii,H; Croft, WB., 1993, "Comparison of Indexing techniques for Japanese Text Retrieval." Paper presented at the 16th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh,PA

Gonnet, GH, Baeza-Yates, RA, Snider, T., 1992 "5. New indices for text : PAT trees and PAT arrays". Information retrieval : data structures & algorithms. Edited by Frakes, WB, and Baeza-Yates, R. Englewood Cliffs, N.J., Prentice Hall, p.66-82.

Jing, Y; Croft, WB., 1994, "An association thsaurus for informatin retireval." Paper presented at RIAO '94, Kageura,K., 1996, “Bigram statistics revisited: a comparative examination of some statistical measure in morphological analysis of Japanese kanji sequences”. available at http://www.dcs.shef.ac.uk/kyo/llc1.ps.

Kageura, K. et al. 1997, “NACSIS Corpus Project for IR and Terminological Research”. Natural Language Processing Pacific Rim Symposium '97, 2-5 December, 1997 at Phuket, Thailand, p. 493-496. (available from http://www.rd.nacsis.ac.jp/~kyo/)

Kando, N., 1997, “Cross-linguistic scholarly information transfer and database services in Japan”. a paper presented at the Annual Meeting of the American Society for Information Science, Nov. 2-7, 1997, Washington, D.C., U.S.A.

Kando, N.; Koyama, T.; Oyama, K.; Kageura, K.; Yoshioka, M.; Nozue, T.; Matsumura, A.; Kuriyama, K., 1998, “NTCIR: NACSIS Test Collection Project" [poster]. the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research, March 25-27, 1998, Autrans, France.

Kando, N.; Aizawa, A.; Tsuji, K.; Kageura,K.; Kuriyama, K., 1998b, “Cross-Language Information Retrieval and Automatic construction of multilingual lexicons”. 1998 Anuual Meeting of the American Society for Information Science. Oct.24-29, 1998, Pittsburgh, PA, U.S.A.

Koyama, T.; Yoshioka, M.; Kageura, K., 1998, “The Construction of a Lexically Motivated Corpus : The Problem with Defining Lexical Units.” The First International Conference on Language Resources and Evaluation. May, 1998, Granada, Spain.

Lee, JH; Ahn, JS., 1996, “Using n-grams for Korean Text Retrieval.” Paper presented at the 19th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich.

Mitra, M., et al., 1997, “An analysis of statistical and syntatic phrases”. In proceedings of RIAO97, Computer Assisted Information Searching on the Internet, Montreal, Canada, June 1997, p.200-214.

NACSIS, 1997, “Introduction to the National Center for Science Information Systems”, NACSIS. (http://www.nacsis.ac.jp/),

Nagao, M.; Kurohashi,S. 1997, “Japanese Morphological Analyzer *Juman, v.3.5”*, Univ. of Kyoto 1997.

Nie, JY; Brisebois, M; Ren, X., 1996, “On Chinese text retrieval.” Paper presented at the 19th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval Zurich,.

Ogawa, Y; Iwasaki, M., 1995, “A new character-based indexing method using frequency data for Japanese documents.” Paper presented at the 18th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Seatle, WA

Ogawa, Y; Matsuda, Toru., 1997, “Overlapping statistical word indexing : a new indexing method for Japanese text.” Paper presented at the 20th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, 1997.

Smeaton.AF., 1998, “User-Chosen Phrases in Interactive Query Formulation for Information Retrieval”. In Proceedings of the 20th BCS-IRSG Colloqium on IR Research, Autrans, France, March 1998, p.128-136.

Vines, P.et al., 1996, “Indexing for Chinese text retrieval.” Proceedings of the 1st Australian Document Computing Conference, pp.85-89

Xiangji Huang, Robertson, SE., 1997, “Experiments on large test collections with probabilistic approaches to Chinese text retrieval.” Paper presented at the 2nd International workshop on Information retrieval with Asian Languages 1997 (IRAL 97), Tsukuba, Japan,.