データベースシステム 115－8
情 報 学 基 礎 49－8
（1998. 5. 14）

First Experiments on the BMIR-J2 Collection using the NEAT System

Gareth Jones Tetsuya Sakai Masahiro Kajiura Kazuo Sumita
Research and Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

Abstract

The BMIR-J2 Collection is the largest generally available information retrieval test set for Japanese text retrieval. This paper describes first experiments on BMIR-J2 using the Toshiba NEAT information retrieval system. Experimental retrieval results indicate that the probabilistic information retrieval model originally developed for English text transfers well to the Japanese language using both word-based and character-based indexing. Additional experiments report retrieval results comparing various approaches to query processing, and retrieval performance for the different query categories defined within BMIR-J2.

1 Introduction

The recently released BMIR-J2 Collection [1] is the largest generally available information retrieval evaluation test set for Japanese text retrieval. This paper describes initial retrieval experiments using BMIR-J2 with the Toshiba NEAT information retrieval system [2]. The NEAT system has recently been modified to enable us to explore the probabilistic information retrieval model [3] for Japanese text retrieval.

Evaluation is an important stage in the development of information processing systems. The performance of an information retrieval system is usually evaluated in terms of agreed quantitative measures, but depends crucially on the availability of a suitable retrieval test collection for which these parameters can be evaluated. Such collections typically consist of a set of documents available to the retrieval system, a number of user requests describing a user's information needs, and a corresponding set of relevance judgements indicating which documents in the archive are *relevant* to each request, that is these documents satisfy the user's information need.

The need to provide relevance information makes the cost of developing retrieval collections very high. In theory the relevance of all documents in the collection must be assessed for each user search request, even with some approximations this process is expensive. The high development cost means that individual organizations are often unable to develop their own retrieval collections. In addition, even when an organization develops its own collection, this only allows developers to explore the the behaviour of their own system. Publically available collections also enable organisations to compare and contrast the behaviour of their system with that of others. The principle current example of this approach is the US NIST TREC (Text REtrieval Conference) held annually for the last 6 years which concentrates primarily on English language text [4]. Participants must first submit their results on a common retrieval task; later they can compare the performance of their system with others at the conference. Such comparisons are often interesting since it is in the nature of information retrieval that systems are often not better or worse than each other overall, but rather perform differently in response to different queries or for different tasks.

This need for evaluation is no less true for Japanese language information retrieval systems. Although the new BMIR-J2 is still rather smaller than experimental collections used at TREC [4], it nevertheless gives us the opportunity to begin to explore Japanese text retrieval on a standard collection of reasonable size.

The remainder of this paper is organised as follows. Section 2 overviews relevant work in Japanese information retrieval, Section 3 describes the current NEAT system, and Section 4 reviews the probabilistic retrieval model used in the NEAT system. Section 5 gives an overview of the BMIR-J2 collection, Section 6 gives our initial experimental results for BMIR-J2 and finally Section 7 summarises our conclusions and further work.

2 Review of Japanese Information Retrieval

Japanese text, like several other Asian languages, presents two main problems for information retrieval systems. First there is the extensive use of ideographic systems, such as the *kanji* character set, and second Japanese is an *agglutinating* lan-

guage.

In order to perform retrieval, content-information must be extracted from the character strings contained within documents and search requests. To date much previous work on Japanese language retrieval has focussed on the development of effective indexing techniques [5] [6].

For Japanese and other Asian languages such as Chinese and Korean, various methods for indexing have been explored in recent years. These can be broadly classified into two approaches: *word-based* analysis which attempts to perform word level segmentation, and *character-based* techniques which extract character strings from the documents, without seeking to identify component words, and use these as the indexing units. There are various arguments in favour of each approach, a good review of these appears in [7].

2.1 Word-Based Indexing

Ideally we would like to automatically perform a perfect segmentation of the text into its constituent words. Once the words were available, existing retrieval techniques could easily be explored; unfortunately such perfect segmentation is not possible. In segmentation it is often not clear whether compound nouns should be broken up into their constituent words or left as a single indexing unit. Two techniques have been investigated which attempt to produce word-level segmentation. Morphological segmentation and statistical segmentation [6]; in our current work we focus only on morphological segmentation.

Morphological segmentation (often referred to as *dictionary-based* segmentation) divides continuous character strings into words using a morphological analyser. In operation the string of characters is compared against word entries in a dictionary. Character strings which match dictionary entries are then extracted as whole words. The morphological analyser tends to extract the *morphemes* of compound words as separate indexing units. Unfortunately morphological segmentation makes mistakes in segmentation which ultimately degrade retrieval performance. Segmentation errors arise principally from ambiguity of word boundaries in the character string and limitations in the morphological analyser. The main limitation is that the morphological analyser cannot identify words outside its dictionary. Thus ideally the dictionary should be continually updated to add new words as they are encountered, but this is an expensive process which will inevitably often lag behind the appearance of new words.

2.2 Character-Based Indexing

The most simple character-based indexing technique is merely to use all the individual characters as indexing units. This approach has been shown to work successfully for Chinese [8]. A slightly more complex variation is to ignore possible word boundaries and extract character n-grams, usually including overlapping ones, as the indexing units [5].

2.3 Comparison of Indexing Methodologies

Character-based n-gram indexing is simple and computationally cheap, it has also been shown to be better than word-based indexing in various studies [9] [6]. However, the limited size of the test collections used in these studies mean that these results can only be taken as indicative. In Japanese there is often more than one way of writing a word, possibly using a different character set. A good overview of the complex issues of synonymy in Japanese is contained in [9]. Many problems arise due to alternative spellings in *kanji* words, alternative *katakana* transliteration, and the use of different character sets. The most obvious way to deal with these problems is through the use of a synonym dictionary; although such a dictionary is costly both to develop and to maintain. Word-based indexing enables such a dictionary to be used if it is available, unfortunately since it is not based on word-level units character-based indexing does not. A formal investigation of the effectiveness of synonym dictionaries in Japanese retrieval is beyond the scope of this paper, but is an important area for future study.

2.4 Data Fusion

The combination of evidence from multiple information sources has been shown to be useful for improving text retrieval performance in TREC [10]. NEAT includes the option of combining evidence using *data fusion*. In data fusion ranked documents lists produced independently in response to a query are combined by the corresponding query-document matching scores from the lists. A new re-ranked list is formed using the composite scores.

In the experiments reported in this paper we explore morphological segmentation and character-based indexing, both in isolation and combined using data fusion.

3 The NEAT Information Retrieval System

The NEAT Information Retrieval System is being developed for the retrieval of online Japanese text articles [2] [11]. Documents are currently indexed using either or both of morphological segmentation and character-based analysis. In response to a search request a list of articles is returned ranked by request-article matching score.

NEAT contains a large amount of retrieval functionality. It can utilise complex search profiles which may include Boolean filtering and document structure. Document structure can be taken into account since terms in the index file contain information of their presence in entities such as the full document text, the document heading, or its first paragraph. Individual request terms can be entered for each document structure field and the contribution to the matching score of each can be assigned within the profile. Each individual request term can be assigned an individual weight in the profile. In addition NEAT can make use of a multi-level query expansion using thesauri. An individual word may be expanded to alternative spellings, direct synonyms, more general terms or more specific ones with the relative weight of terms from each expansion source set dynamically.

In our current work we are investigating the use of the probabilistic retrieval model for Japanese. The following section contains an introduction to the probabilistic model and a summarised derivation of the model used in our work. This paper describes only experiments using full-text retrieval without Boolean filtering.

4 Probabilistic Retrieval

Probabilistic retrieval models have been successful both in improving experimental information retrieval performance particularly for English, and in providing a theoretical basis for methods which had previously relied on heuristics [12].

The rationale for the introduction of probabilistic concepts into information retrieval lies in the uncertainty of natural language. The language in documents is much too uncertain to state with certainty whether a particular document will be relevant to a particular user request. Probabilistic ideas were first introduced for information retrieval in [13], but it was some time before a practical demonstration of the power of this approach appeared [14]. The basic idea is to make use of data about the distribution of the search words (or *terms*) within a document collection to improve retrieval performance. This data distribution information can be used to calculate *weights* for search terms that define relevance probabilities for previously unjudged documents. The precise form of the resulting weighting scheme depends on assumptions about the nature of the *statistical independence* of terms and the contributions of search terms that are, or are not, present in a document.

This section contains a summarised description of the typical assumptions of the probabilistic retrieval model, and a simplified derivation of one current popular model, BM25 [15], used in our experimental work.

4.1 Basic Model

Assume that each document can be described as a binary vector $x = (x_1, x_2, \ldots, x_v)$ where $x_i = 0$ or 1 indicates whether or not term i is present, and v is the size of the collection vocabulary. Based on this description we can form a decision rule to assign each document as either relevant or non-relevant to a particular query. The obvious rule is to assign the document as relevant if,

$$P(\text{Rel}|x) > P(\text{Non-Rel}|x) \quad (1)$$

A more useful form of this decision rule can be derived using Bayes' theorem to derive a weighting function $g(x)$.

$$g(x) = \log \frac{P(x|\text{Rel})}{P(x|\text{Non-Rel})} + \log \frac{P(\text{Rel})}{P(\text{Non-Rel})} \quad (2)$$

This means that instead of making a binary decision about the relevance of a document, the documents can be ranked by their $g(x)$ value such that the more highly ranked a document is, the more likely it is to be relevant. The second term of Equation 2 is constant for a given query, and thus will not affect the ranking of the documents and can be ignored.

If it is assumed that the index terms occur *independently* in the relevant and non-relevant documents then,

$$P(x|\text{Rel}) = P(x_1|\text{Rel})P(x_2|\text{Rel})\ldots P(x_v|\text{Rel})$$

and similarly,

$$P(x|\text{Non-Rel}) = P(x_1|\text{Non-Rel})P(x_2|\text{Non-Rel}) \ldots P(x_v|\text{Non-Rel})$$

Let,

$$p_i = P(x_i = 1|\text{Rel}) \text{ and } q_i = P(x_i = 1|\text{Non-Rel})$$

where these are the probabilities that an index term occurs in the relevant and non-relevant sets respectively. Then,

$$P(x|\text{Rel}) = \prod_{i=1}^{v} p_i^{x_i}(1-p_i)^{1-x_i}$$

and

$$P(x|\text{Non-Rel}) = \prod_{i=1}^{v} q_i^{x_i}(1-q_i)^{1-x_i}$$

Substituting these into Equation 2 with the assumptions of term independence gives,

$$g(x) = \sum_{i=1}^{v} x_i \log \frac{p_i(1-q_i)}{(1-p_i)q_i} + \sum_{i=1}^{v} \log \frac{1-p_i}{1-q_i} \quad (3)$$

The second term of Equation 3 will be also constant for a given query, and again will not affect the ranking of the documents. In theory the first term involves a summation over all the terms in the document collection, but in practise this summation is usually restricted just to the search terms in the current query.

The evaluation of $g(x)$ thus reduces to a simple summation matching function between the query and document x where term i has weight,

$$w(i) = \log \frac{p_i(1-q_i)}{(1-p_i)q_i} \quad (4)$$

4.2 Robertson/Sparck Jones Relevance Weight

For a particular query let there be R relevant documents, r_i of which contain term i, from a total of N documents n_i of which contain i. Then taking the obvious estimates of p and q,

$$p_i = \frac{r_i}{R} \qquad q_i = \frac{n_i - r_i}{N - R}$$

gives the following weight for i,

$$w(i) = \log \frac{(r_i + 0.5)(N - R + n_i - r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)} \quad (5)$$

where the 0.5's are added for reasons do with estimation from small amounts of data; in particular it should not be assumed that because a term has not appeared in any relevant documents so far, that it will never do so. This weight is often referred to as the Robertson/Sparck Jones relevance weight [14].

In the absence of any relevance information, that is $R = r = 0$, $w(i)$ reduces to,

$$w(i) = \log \frac{N - n + 0.5}{n + 0.5}$$

which is a very close approximation to the standard *collection frequency weight (cfw)* (often called the *inverse document frequency* weight) [12].

$$cfw(i) = \log \frac{N}{n}$$

4.3 Within-Document Term Frequency

The simple binary vector representation of x does include any information of the frequency with which terms occur within individual documents. It is intuitively sensible to suggest that if a search term occurs more than once in a document, then this document is more likely to be relevant. This intuition leads to the popular $w(i,tf) = tf(i,j) \times cfw(i)$ weight, where $tf(i,j)$ is the frequency of term i in document j and $w(i,tf)$ is the corresponding term weight. This intuitive argument is borne out by empirical studies [12].

The probabilistic model can be extended to incorporate term frequency on a more theoretical basis [16]. Thus, Equation 4 can be replaced by,

$$w(i,tf) = \log \frac{p_i^{tf} q_i^0}{q_i^{tf} p_i^0} \quad (6)$$

where

$$p_i^{tf} = P(i \text{ present with frequency } tf|\text{Rel})$$

and

$$q_i^{tf} = P(i \text{ present with frequency } tf|\text{Non-Rel})$$

and p_i^0 and q_i^0 indicate the corresponding probabilities for term absence.

Robertson [16] takes a 2-Possion model of term frequency and by applying a set of approximations derives the following simple enhanced probabilistic model.

$$w(i,tf) = w(i) \times \frac{tf \times (k_1 + 1)}{k_1 + tf}. \quad (7)$$

where the constant k_1 must be determined empirically for a particular retrieval environment.

4.4 Document Length

It can be postulated that there are two reasons why documents may vary in length. Some documents may contain more information either because they are more detailed or because they cover more than one topic; other longer documents could be described as verbose, being longer merely because they use more words to say the same thing. In reality many documents are likely to exhibit a mixture of these effects.

The number of occurrences of a term will depend on dl. Thus longer documents will tend to have higher matching scores; however, it is reasonable to assume that relevance is independent of document length. A simple way to incorporate this

observation is to modify Equation 7 so that tf is normalised with respect to document length. If we assume that k_1 has been set appropriately for documents of average length, this new equation can be expressed as,

$$w(i,j) = w(i) \times \frac{tf \times (k_1 + 1)}{\frac{k_1 \times dl}{\text{average } dl} + tf} \quad (8)$$

Note: Term i now has a unique weight for each document, hence $w(i, tf)$ becomes $w(i, j)$.

4.5 Final BM25 Model

Equations 7 and 8 can be combined to give a more flexible weighting scheme,

$$cw(i,j) = \frac{w(i) \times tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where cw indicates the combined weighting scheme, ndl is the normalised length of document j, and b is an empirically determined constant that controls the degree of length normalisation. This equation is referred to in the literature as BM25. Where there is no relevance information $w(i)$ is simply replaced by the standard $cfw(i)$ weight [3].

5 The BMIR-J2 Collection

The BMIR-J2 collection consists of 5080 articles taken from the Mainichi Newspapers in the fields of economics and engineering, and a total of 60 search requests. Each request consists of a natural language phrase describing a user's information need.

Relevant documents for each query were identified as follows. A broad Boolean expression was used to identify most possible relevant documents. The retrieval documents were manually assessed for relevance to the query and the assessment cross-checked by another assessor. The average number of relevant documents for queries 0101-0150 is 33.62, for queries 0151-0160 is 2.1 and overall is 28.4.

The BMIR-J2 queries are subdivided into 5 categories representing increasing retrieval difficulty:

- (A) *basic function*: retrieval possible by simple term matching;
- (B) *numeric range*: retrieval system should be able to handle numerical comparisons;
- (C) *syntactic function*: analysing a syntactic relationship between words in the request should improve understanding of the query;
- (D) *semantic function*: semantic analysis should improve understanding of the query;
- (E) *world knowledge function*: knowledge beyond that contained in the system or the query is required to understand the query.

Since the NEAT system does not currently perform any form of linguistic analysis we would expect it to perform less well for the latter four categories. Retrieval results are given for queries in each category and averaged over all categories.

5.1 Query Pre-Processing

In the NEAT system natural language search requests are first segmented using morphological analysis. In English language retrieval a request is typically processed to remove frequently occurring common stop words, e.g. function words; and the remaining words then suffix stripped to encourage matching between different forms of the same word. It is not obvious how similar methods might best be employed for Japanese text. Thus we have so far investigated four types of query as follows:

- *(i)* using the request as segmented;
- *(ii)* as above but removing single character *hiragana*;
- *(iii)* as above but additionally removing all single character *kanji*;
- *(iv)* including in the query only terms identified as nouns by the morphological analyser, this includes some single character *kanji*, but obviously excludes others, and includes the extracted base nouns of *suru* verbs.

6 Retrieval Experiments

In this section we present the results of our initial BMIR-J2 retrieval experiments. First, we investigate the effectiveness of term weighting strategies for Japanese text retrieval with BMIR-J2. We next explore the query pre-processing approaches described in Section 5.1, and then examine retrieval performance using data fusion. Finally we present retrieval results for each query type within BMIR-J2.

All results show retrieval *precision* at ranked cutoff of 5, 10, 15 and 20 documents, and standard TREC average precision. Precision is the proportion of retrieved documents at a given rank position which are relevant to the query. Average precision is calculated by averaging the precision values for all relevant documents for a query, and then averaging these values across the query set.

Weight Scheme		*uw*	*cfw*	*cw*
Prec.	5 docs	0.403	0.453	0.513
	10 docs	0.363	0.405	0.438
	15 docs	0.334	0.363	0.396
	20 docs	0.303	0.324	0.358
Av. Precision		0.337	0.390	0.436

cw: $K1 = 0.5$, $b = 0.4$

Table 1: Retrieval precision values for BMIR-J2 using Morphological Indexing.

Weight Scheme		*uw*	*cfw*	*cw*
Prec.	5 docs	0.407	0.440	0.510
	10 docs	0.365	0.398	0.437
	15 docs	0.333	0.369	0.397
	20 docs	0.302	0.335	0.361
Av. Precision		0.337	0.402	0.450

cw: $K1 = 0.5$, $b = 0.4$

Table 2: Retrieval precision values for BMIR-J2 using Character-Based Indexing.

Although BMIR-J2 is the largest collection so far generally available for Japanese text retrieval evaluation, it is still comparatively small, and hence specific figures reported here should not be taken as reliable. Overall we concentrate on the general trends which emerge from our results.

6.1 Effectiveness of Term Weighting

As described earlier we were interested in investigating the effectiveness of the standard term weighting components when applied to Japanese text retrieval. We compare a baseline system using *unweighted uw* terms (term weight is 0 or 1) against *cfw* and *cw* weighting. Table 1 shows retrieval performance for BMIR-J2 with text indexing using morphological analysis; for *cw* weighting the values of $k1$ and b have been optimised for BMIR-J2. In all experiments document length is measured as the number of morphs in the document. Table 2 shows retrieval performance for BMIR-J2 with text indexing using character-based indexing. Again, for *cw* weighting the values of $k1$ and b have been optimised.

Observations From the results in Tables 1 and 2 it can be seen that using *cfw* gives an improvement over the *uw* benchmark in cutoff and average precision. In addition, further improvement is obtained in all cases by using the more complex *cw* weighting scheme.

It can be observed that character-based indexing appears to be more effective for retrieval than morphological segmentation. However, investigation of individual queries showed that the difference is almost entirely due to query 0153. This has only 2 search terms and only 1 relevant document. For *cfw* and *cw* this was ranked at position 8 using morphological indexing, but position 1 using character-based indexing. The divergence in behaviour only appears when *cfw* weighting is applied, further examination showed that for *uw* they both ranked the relevant document at position 8. Examination of the indexed document showed that while both search terms were present, a context-related error by the morphological analyser meant that this term did not appear as an indexing unit. However, for the character-based indexing the term was available as an indexing unit since no hard word boundary decisions are made during indexing. Similar problems are likely to have occurred for other queries, however the presence of more relevant documents may have reduced the effect on overall results through averaging, or longer queries may have reduced the dependence on accurate indexing of individual terms. For query 0153 the absence of restriction on the character strings available as indexing units in character-based indexing produced a positive result. However this may not always be the case since this generalisation in character-based indexing may produce false matches on character strings, which just happen to be the same as an indexing unit; such behaviour may on some occasions degrade retrieval performance.

While behaviour for the individual query 0153 is obviously important, this example illustrates the care which must be taken in analysing retrieval results. Short queries with few relevant documents are particularly prone to volatile retrieval behaviour, and averaging can accentuate the effect of such queries on overall results. Thus we should not draw broad conclusions from behaviour on individual queries.

It is recommended by the designers of BMIR-J2 that queries 0101-0150 are used as the main retrieval collection, and the remaining results in this paper use this reduced query set.

6.2 Query Pre-Processing Experiments

Tables 3 and 4 show retrieval results for queries 0101-0150 using *cw* weighting for the four types of query examined with morphological indexing and character-based indexing respectively.

Observations Although the differences between the performances figures are slight, they suggest that removing hiragana characters is beneficial, but that retaining single character kanji is preferable to removing them. Alternatively, using only

Prec	Query Type			
	(i)	*(ii)*	*(iii)*	*(iv)*
5 docs	0.588	0.580	0.572	0.584
10 docs	0.508	0.506	0.496	0.508
15 docs	0.464	0.464	0.453	0.461
20 docs	0.420	0.423	0.413	0.418
Av. Precision	0.441	0.442	0.435	0.442

cw: $K1 = 0.5$, $b = 0.4$

Table 3: Retrieval precision values for BMIR-J2 using Morphological Indexing.

Prec	Query Type			
	(i)	*(ii)*	*(iii)*	*(iv)*
5 docs	0.580	0.588	0.556	0.576
10 docs	0.506	0.508	0.494	0.510
15 docs	0.465	0.463	0.459	0.469
20 docs	0.423	0.420	0.422	0.427
Av. Precision	0.442	0.443	0.435	0.441

cw: $K1 = 0.5$, $b = 0.4$

Table 4: Retrieval precision values for BMIR-J2 using Character-Based Indexing.

nouns appears to be equally effective. Tests of larger collections with more queries would be required to confirm these initial findings. Queries of type *(ii)* are used for the tests reported in Tables 1 and 2, and all other results reported. Separate tests showed that the behaviour noted already for query 0153 occurred using all 4 query types.

6.3 Data Fusion

Table 5 shows BMIR-J2 retrieval performance for the individual indexing methods and data fusion for queries 0101-0150. All figures are shown for optimal proportional weighting of retrieved document lists from morphological segmentation and character-based indexing.

Observations A slight improvement in retrieval performance is observed when using data fusion. However, there are indexing, storage and retrieval costs associated with utilizing two indexing methods, and it is not clear that the modest performance gain will be sufficient to justify these costs.

6.4 BMIR-J2 Query Types

Table 6 shows retrieval performance broken down by query types defined in section 5. The queries in the final column (F) combine the features of type D and E queries. The table also shows the number of queries in each class.

Morphological Indexing

Weight Scheme		*uw*	*cfw*	*cw*
Prec.	5 docs	0.464	0.496	0.580
	10 docs	20	0.460	0.506
	15 docs	0.388	0.431	0.464
	20 docs	0.353	0.393	0.423
Av. Precision		0.351	0.403	0.442

Character-Based Indexing

Weight Scheme		*uw*	*cfw*	*cw*
Prec.	5 docs	0.460	0.516	0.588
	10 docs	0.418	0.468	0.508
	15 docs	0.389	0.424	0.463
	20 docs	0.355	0.380	0.420
Av. Precision		0.351	0.406	0.443

Data Fusion

Weight Scheme		*uw*	*cfw*	*cw*
Prec.	5 docs	0.460	0.504	0.580
	10 docs	0.418	0.466	0.508
	15 docs	0.389	0.424	0.463
	20 docs	0.355	0.384	0.424
Av. Precision		0.352	0.410	0.449

cw: $K1 = 0.5$, $b = 0.4$

Table 5: Retrieval precision values for BMIR-J2 using Data Fusion.

Observations These results indicate that, as anticipated, retrieval performance does decrease as the query difficulty increases. In particular the simple matching method adopted in NEAT becomes much less effective once semantic understanding of the query is required.

7 Conclusions and Further Work

This paper has described our first experiments with the BMIR-J2 Collection. Our results show similar trends to previous experiments with the much smaller BMIR-J1 Collection [17]; this gives us good reason to hope that the techniques which we are currently using in the NEAT system will extend naturally to larger Japanese language retrieval tasks. It is to be hoped that larger test collections will become available in the future to enable us to test this hypothesis.

Our current work is concentrated on further experimental investigations using BMIR-J2, particularly examining various feedback techniques for query expansion and term reweighting.

References

[1] T. Kitani et al. BMIR-J2 - A Test Collection for Evaluation of Japanese Information

		Query Type					
		A	*B*	*C*	*D*	*E*	*F*
No of Queries		14	3	10	9	4	10
Prec	5 docs	0.714	0.800	0.720	0.378	0.450	0.420
	10 docs	0.593	0.667	0.620	0.422	0.450	0.330
	15 docs	0.567	0.644	0.553	0.378	0.350	0.293
	20 docs	0.546	0.583	0.495	0.344	0.313	0.250
Av. Precision		0.574	0.677	0.580	0.309	0.325	0.252

cw: $K1 = 0.5$, $b = 0.4$

Table 6: Retrieval precision values for BMIR-J2 using Data Fusion.

Retrieval Systems. In *Information Processing Society of Japan National SIG DBS Workshop Notes 98-DBS-3*, Okayama, 1998. IPSJ.

[2] M. Kajiura, S. Miike, T. Sakai, M. Sato, and K. Sumita. Development of the NEAT Information Filtering System. In *Proceedings of the 54th Information Processing Society of Japan National Conference*, pages 3–(299–300), Tokyo, 1997. IPSJ.

[3] S. E. Robertson and K. Sparck Jones. Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, updated May 1997.

[4] D. K. Harman and E. M. Voorhees, editors. *The Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, 1998. NIST.

[5] Y. Ogawa and M. Iwasaki. A New Character-based Indexing Method using Frequency Data for Japanese Documents. In *Proceedings of ACM SIGIR*, pages 121–129, Seattle, 1995. ACM.

[6] Y. Ogawa and T. Matsuda. Overlapping statistical word indexing: A new indexing method for Japanese text. In *Proceedings of ACM SIGIR*, pages 226–234, Philadelphia, 1997. ACM.

[7] J.-Y. Nie, M. Brisebois, and X. Ren. On Chinese Text Retrieval. In *Proceedings of ACM SIGIR*, pages 225–233, Zurich, 1996. ACM.

[8] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams. Okapi at TREC-5. In *Proceedings of the Text REetrieval Conference (TREC) 5*. NIST, 1997.

[9] H. Fujii and W. B. Croft. A Comparison of Indexing Techniques for Japanese Text Retrieval. In *Proceedings of ACM SIGIR*, pages 237–246, Pittsburgh, 1993. ACM.

[10] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31:431–448, 1995.

[11] T. Sakai, M. Kajiura, S. Miike, M. Sato, and K. Sumita. Evaluation of the NEAT Information Filtering System Using the BMIR-J1 Benchmark. In *Proceedings of the 54th IPSJNational Conference*, pages 3–(301–302), Tokyo, 1997. IPSJ.

[12] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[13] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7:214–244, 1960.

[14] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancok-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3*, pages 109–126. NIST, 1995.

[16] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of ACM SIGIR*, pages 232–241, Dublin, 1994. ACM.

[17] G. J. F. Jones, T. Sakai, M. Kajiura, and K. Sumita. Experiments in Japanese Text Retrieval and Routing using the NEAT System. In *Proceedings of ACM SIGIR*, Melbourne, 1998. ACM. To appear.