Experiments in Japanese Text Retrieval and Routing using the NEAT System

Gareth J. F. Jones Tetsuya Sakai Masahiro Kajiura Kazuo Sumita

Human Interface Technology Center
Communication and Information Systems Research Laboratories
Research and Development Center, Toshiba Corporation
1, Komokai Toshiba-cho, Saiwai-ku, Kawasaki 210-8582, Japan
email: {jones,tets1,kajiura,sumita}@eel.rdc.toshiba.co.jp

Abstract This paper describes a structured investigation into the retrieval of Japanese text. The study includes a comparison of different indexing strategies for documents and queries, investigation of term weighting strategies principally derived for use with English texts, and the application of relevance feedback for query expansion. Results on the standard BMIR-J1 and BMIR-J2 Japanese retrieval collections indicate that term weighting transfers well to Japanese text. Indexing using dictionary based morphological analysis and character strings are both shown to be individually effective, but marginally better in combination. We also demonstrate that relevance feedback can be used effectively for query expansion in Japanese routing applications.

1 Introduction

It is widely acknowledged that the increased availability of electronic information sources has greatly increased general interest in information retrieval. While much research effort has focussed on retrieval of English text, demand for effective retrieval systems is emerging in many languages. This paper describes experiments from the development of the NEAT system for the retrieval of articles from online Japanese newspapers.

There has been much interesting research work in information retrieval in recent years. However, largely motivated by the TREC programme, see for example [8] [9], a large amount of this is currently focussed on very large English document archives. Although separate tracks within TREC have examined various other languages, such as Spanish and Chinese, and applications, such as OCR and spoken document retrieval, for smaller tasks; there are many interesting research problems outside the scope of the current TREC programme. One interesting observation from the existing additional tracks is that methods originally developed for English text often transfer well to other languages and domains. In this paper we provide a careful examination of the transfer of some techniques to Japanese, while taking into account the particular problems inherent in the processing of Japanese text.

Our experiments explore the effectiveness of term weighting in Japanese text retrieval; in particular examining the utility of the three main components generally used to affect the value of a term weight: collection frequency weighting, within document term frequency and document length normalization.

Further, the daily retrieval of news texts can be viewed as an iterative information routing task. News articles from today can be viewed as a routing collection and all previous material as a training set which can be used to build the most effective possible query. In this application we are interested not only in effective retrieval from the daily news archive, but also in being responsive to relevance information provided each day by the reader. In a first attempt to make effective use of incremental relevance information we investigate the modification of standing Japanese search profiles via query expansion from relevance feedback.

The remainder of this paper is organized as follows. Section 2 provides an overview of the problems in Japanese text retrieval and reviews previous work in this area. Section 3 outlines the NEAT system for Japanese text retrieval. Section 4 describes the retrieval collections used for our current experimental evaluations, and Section 5 describes the particular retrieval features and techniques that are explored in this paper, experimental results are then presented in Section 6. Finally Section 7 summarizes the conclusions from this work and outlines directions for our future research.

2 Overview of Japanese Text Retrieval

The retrieval of texts in various Asian languages such as Japanese, Chinese and Korean presents two problems. First there is the extensive use of ideographic systems, such as the use of the Chinese *kanji* character set in Chinese and Japanese. Second they are *agglutinating* languages, that is sentences have no spaces between words.

In order to perform retrieval content-information must be extracted from the character strings contained within documents and search requests. It is thus not surprising that much previous work on Asian language retrieval has focussed on the development of effective indexing techniques [17] [4] [13] [15]. Since the emphasis of this paper is on Japanese text, we briefly describe the Japanese character sets and outline the currently available indexing strategies.

2.1 Japanese Character Sets

Japanese text contains 3 classes of character: *kanji*, *katakana*, and *hiragana*, each of which fulfill different linguistic functions. In addition, Japanese text often contains words in western characters (or *romaji*). kanji characters are based on Chinese ideograms, kanji may be used individually to form words with simple meanings or grouped together to form words with more complex meanings. There are around 2000 kanji in common usage, but many more are available when necessary. Hiragana characters are used for particles, auxiliary verbs, and conjugational parts. Katakana characters are mainly used to transliterate western words into Japanese. There are around 80 hiragana and katakana characters. Romaji characters are used for western words used in Japanese without transliteration. A typical Japanese natural language search request might thus take the form:

> 開始時間が午前１０時の日経ビジネススクール
> Nikkei business school that starts at 10 o'clock a.m.

2.2 Indexing Methodologies

Various methods for indexing Asian languages 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 for use as indexing units, without seeking to identify component words. There are various arguments in favour of each approach, a good review of these appears in [15].

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. However, such perfect segmentation is not possible, indeed it is sometimes not even clear what the definition of individual words should be. A significant source of this segmentation ambiguity is the free generation of new compound nouns in Japanese. For segmentation it is often not clear whether such 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 Morphological segmentation (often referred to as *dictionary-based* segmentation) divides continous 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 will tend to extract the component words (or *morphemes*) of compound words as separate indexing units. Unfortunately depending on the exact forms of the component words, this extraction may not appear to be applied entirely consistently. In fact morphological segmentation makes mistakes in segmentation which ultimately lead to degradation in 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. Another important factor is that morphological analysis is computationally expensive; however, since all analysis is done before retrieval, this is only an issue in indexing efficiency.

Statistical Segmentation Statistical segmentation is an alternative to dictionary-based word segmentation. This approach is based on the computation of the likelihood of word breaks between characters. While this segmentation method is less accurate than morphological segmentation it does not require a dictionary and hence can segment any character strings [16]. Improvements to this method, including use of overlapping segments described in [18], demonstrate that this method can be effective for Japanese text retrieval. The only prerequisite for this approach is that a suitable corpus is available to train the statistical segmentation parameters.

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 [2]. A slightly more complex variation is to ignore possible word boundaries and extract character n-grams, usually including overlapping ones, as the indexing units [17]. A more complex strategy for extracting n-grams after morphological segmentation is proposed in [13].

A hybrid character based analysis is applied in [6], kanji characters are indexed individually and katakana character strings are extracted as complete strings for use as individual indexing units. Similarly complete words in romaji characters can easily be extracted from documents and queries. To improve their contribution to retrieval performance romaji terms can be handled using standard English language retrieval techniques of stop word removal and suffix-stripping.

Comparison of Indexing Methodologies

Character-based n-gram indexing is simple and computationally cheap, it has also been shown to be as good as and perhaps better than word based indexing in various studies [6] [18]. The limitations of the test collections used in these studies mean that these results can generally only be taken as indicative. However, character-based indexing has two particular disadvantages:

- in Japanese there is often more than one way of writing a word, possibly using a different character set. Word-based indexing enables a thesaurus to be used to access different word forms, character-based indexing does not.
- interactive query expansion can only be used if the user is able to judge the usefulness of a possible new term in the expanded query. The user can offer judgement on complete words, but often not on arbitrary character strings.

The issues of synonymy in Japanese are quite complex and a good overview is contained in [6]. 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. However, such a dictionary is costly both to develop and to maintain.

In the experiments reported in this paper we restrict our investigation to retrieval using morphological segmentation and character-based indexing, both individually and in combination.

3 The NEAT Information Retrieval System

The NEAT Information Retrieval System is being developed for the retrieval of online Japanese text articles [12] [23]. 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 ranked by request-article matching score is returned.

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 field 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 each 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.

Specific projects at present are concentrated on the retrieval of online news articles using preset topic profiles, and personalised automated retrieval of web pages. The basic NEAT system is already commercially operational and is supplying news material to paying customers. Our current work, as described in this paper, is focussed on establishing and improving the retrieval effectiveness of the NEAT system. For reasons of space this paper describes only experiments using full-text retrieval without Boolean filtering and does not attempt to investigate the potential retrieval effectiveness of thesaurus expansion.

4 Test Collections

Ideally we would like to evaluate the NEAT system on large generally available Japanese test collections. Unfortunately, there are no such collections currently available. Thus our experiments use the BMIR-J1 and BMIR-J2 Japanese text retrieval collections which are only of small and moderate size respectively. In addition, we use our own TCIR-N1 collection as a training set for our routing experiments.

4.1 BMIR-J1

The BMIR-J1 collection consists of 600 articles from the *Nikkei* newspaper[1] and contains 60 natural language search requests with full document relevance assessment information for each one. The average number of relevant documents for each requests is 10.1.

BMIR-J1 was designed so that some search requests can be satisfied very easily, while for some others it is very difficult to retrieve the relevant documents using the request.

4.2 BMIR-J2

The BMIR-J2 collection consists of 5080 articles taken from the Mainichi Newspapers in the fields of economics and engineering, and a total of 50 main search requests[2]. Again, 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 each query is 33.6.

Like BMIR-J1, BMIR-J2 was designed so that some search requests can be satisfied very easily, while for some others it is very difficult to retrieve the relevant documents using the request.

4.3 TCIR-N1

The TCIR-N1 collection consists of 5048 documents taken from the *Nikkei* newspaper CD-ROM 1995 between July 1st and July 10th. It uses only 56 of the search requests from the BMIR-J1 collection, since no relevant documents were found for the other 4 requests. Relevance judgements were made by forming a very general Boolean expression from each request and manually assessing the relevance of each document retrieved by this expression. The average number of relevant documents per request for TCIR-N1 is 15.1. Note that this is proportionally lower than for BMIR-J1 and BMIR-J2 partially because the articles in TCIR-N1 are not focussed on the domain of the queries, unlike those in BMIR-J1 and BMIR-J2; and also probably since the expensive relevance assessment process was less exhaustive.

5 Information Retrieval Techniques

English language information retrieval systems typically make effective use of stop word removal and suffix stripping. As described already the nature of Japanese text means that we adopt a slightly different approach. The documents are morphologically segmented and separately indexed by extracting character strings. For retrieval the search request is first segmented using the morphological analyser. The segmented requests contain many one character hiragana particles, these are roughly equivalent to function words in English text, and hence are not useful for retrieval and are removed from all requests. Also many one character kanji are often present, many of these occur very frequently in different contexts and have very general meanings. It was not obvious whether all one character kanji should be ignored in retrieval or

[1] The data in the BMIR-J1 collection was provided to the Working Group for Benchmark Database for Evaluation of Information Retrieval Systems (in the SIG Database System of the Information Proccesing Society of Japan), courtesey of the Nihon Keizai Shimbun, Inc., and is based on articles that appeared in the Nikkei newspaper between Sept. 1, 1993 and Dec. 31, 1993.

[2] Data in BMIR-J2 is taken from the Mainichi Shimbun CD-ROM 1994 data collection. BMIR-J2 was constructed by the SIG Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

not. Preliminary experiments with the BMIR-J2 collection suggested that retaining these characters in search queries is beneficial to retrieval [11], and so they were retained in the requests used in this paper. Further investigation of Japanese request preprocessing should be carried out using larger and more demanding collections when they become available. Many verbs and adjectives are formed by appending a suitable ending, e.g. *suru* = "do" is appended to form a verb [6]. The base noun form can often be extracted in the morphological segmentation. In addition, the absence of plurals and other variations of noun forms means that matching between nouns in documents and queries is often straightforward and there is no significant requirement for any form of suffix stripping.

Returning to the earlier Japanese text example. When processed as a request using morphological segmentation and removing the resulting single character hiragana terms the following search *query* is produced.

開始, 時間, 午前, １０, 時の, 日経, ビジネス, スクール

The processed search query is then scored independently against the morphological and character-based document indexes. The query-document matching score for each indexing source is simply the usual sum of matching term weights.

5.1 Term Weighting

Effective term weighting is generally accepted to improve retrieval effectiveness. In these experiments we compare retrieval performance for *unweighted* (*uw*) terms with two term weighting schemes. These are standard *collection frequency weighting* (*cfw*) (also called *inverse document frequency* weighting) and the *combined weight* (*cw*), often known as BM25, originally developed in [21] and further elaborated in [22]. The *cw* model was chosen because it has been shown to be effective not only for English text retrieval, but also where documents have been imperfectly indexed, for example in Chinese text retrieval [2], and in retrieval of spoken documents [26].

The BM25 *cw* weight for a term is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i)$ is the standard collection frequency weight, $tf(i,j)$ is the document term frequency, and $ndl(j)$ is the normalized document length. $ndl(j)$ is calculated as,

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the length of j. $K1$ and b are empirically selected tuning constants for a particular collection[3]. $K1$ is designed to modify the degree of effect of $tf(i,j)$, while constant b modifies the effect of document length. High values of b imply that documents are long because they are verbose, while low values imply that they are long because they are multitopic.

The BM25 probabilistic model has previously been used for retrieval of Japanese text as reported in [18]. However, a particular contribution of our work in this respect is to examine the individual contributions of the weighting components for Japanese.

[3]The names of these constants are preserved from the original publications for consistency.

Document Length $dl(j)$

In retrieval of non-agglutinating languages the length of the document is usually taken as the number of terms contained in the document. However, when the document is indexed using a character-based approach an alternative measure must be used. Since $ndl(j)$ is the ratio between different lengths, the absolute length of the document is not important and thus alternative measures of $dl(j)$ can be used. This approach to document length measurement has been shown to be effective for spoken document retrieval [10]. In the experiments which follow we compare two measures of $dl(j)$:

- the number of morphologically segmented terms in the document.
- the number of individual characters in the document.

Experiments in term weighting for ad hoc retrieval are reported using both the BMIR-J1 and BMIR-J2 collections.

5.2 Relevance Feedback

Relevance feedback techniques seek to improve retrieval performance by taking advantage of document relevance information provided by the user. These techniques have been shown to be useful for a number of retrieval collections [25]. For the routing of online news stories we are specifically interested in using the user's relevance judgements about news articles from previous days to attempt to improve the initial retrieval ranking of today's articles.

Relevance information may be used either for query expansion or term reweighting. In our ongoing experiments we are investigating both of these techniques, however the experiments described here focus only on query expansion. For our current experiments the TCIR-N1 collection is used as a training collection with BMIR-J1 as the test collection. The TCIR-N1 and BMIR-J1 collections are used for these experiments rather than the BMIR-J2 collection since they are both taken from the *Nikkei* newspaper.

For each query, the terms appearing in the morphologically segmented version of the relevant documents in TCIR-N1 were ranked according to one of several investigated selection values. Various expansion term selection criteria have been developed for English language retrieval and here we investigated how well some of them perform for Japanese.

The selected terms were added to the original query which was then applied to the retrieval system as before to be scored against both morphologically segmented and character-based representations of the documents.

The expansion term ranking criteria investigated were: $r(i)$ the number of relevant documents in which term i appears, $rtf(i)$ the total number of times the term i occurs in relevant documents, and the Robertson selection value (rsv) [19]. The rsv is defined as,

$$rsv(i) = r(i) \times rw(i)$$

where $r(i)$ is again the number of relevant documents containing term i, and $rw(i)$ is the standard Robertson/Sparck Jones relevance weight [20]. $rw(i)$ is defined

as,

$$rw(i) = \log \frac{\frac{(r(i)+0.5)}{(R-r(i)+0.5)}}{\frac{(n(i)-r(i)+0.5)}{(N-n(i)-R+r(i)+0.5)}}$$

where $n(i)$ is the total number of documents containing term i, R is the total number of relevant documents for this query, and N is the total number of documents.

We have investigated several techniques for selecting appropriate terms from a ranked list.

- Add a fixed identical number of expansion terms not already present in the query from the top of the ranked list. Since expansion terms may not always be reliable they may be downweighted by a scalar constant factor relative to the original query terms.
- Add the optimal number of terms from the top of the ranked term list for each query as measured on the TCIR-N1 training set. Again the expansion terms may be downweighted by a scalar constant.
- Sequentially investigate the retrieval utility on the TCIR-N1 training set of adding each expansion term individually. Different scalar constant factors can be explored for each expansion term. Only terms which improve the retrieval performance for the query on TCIR-N1 are added to the query.

Methods of this type are of interest since it is important that selection of expansion terms can be completely automatic. It may of course be useful to offer possible expansion terms to the user for approval, but such options are beyond the scope of this current investigation.

5.3 Index Combination Methods

The combination of evidence from multiple information sources has been-shown to be useful for text retrieval in TREC [3]. Some form of index combination has already been shown to be effective for Japanese [5]. In our experiments we examine two forms of index combination defined in [3]: *data fusion* and *query combination*.

Data Fusion

For data fusion we combined the ranked document lists produced independently by morphologically segmented and character-based indexes in response to a query. The lists were combined by adding the corresponding query-document matching scores from the two lists and forming a new re-ranked list using the composite scores.

To explore optimization of the data fusion process we investigated scaling the scores of each list by a fixed constant before addition. Thus the overall combined score was given by,

$$cms(j) = (x.ms_{morph}(j)) + ((1-x).ms_{char}(j)), \quad 0 \le x \le 1$$

where $cms(j)$ is the combined query-document matching score of document j, $ms_{morph}(j)$ is the matching score for j using morphological segmentation, and $ms_{char}(j)$ is the matching score for j using character-based indexing.

Query Combination

In query combination different representations of a query are combined into a single representation to score against a document archive. Earlier we noted some of the problems associated with compound nouns for Japanese. In morphological analysis compounds may be separated into their component words depending on their context. In order to effectively index the possible different compound forms we investigated query combination.

Our requests are ordinarily segmented using morphological analysis. For our query combination experiments we investigated the use of a simple rule-based segmentation scheme. The input request is segmented using some simple heuristic rules based on where character type changes, e.g. from kanji to hiragana [24]. Requests segmented using this method have many more compound nouns preserved than are found using morphological analysis.

Retrieval was investigated using these new queries both in isolation, and also combined with our original queries generated using morphological analysis. A new set of combined queries was formed by taking the unique items from the corresponding query in both the existing sets.

6 Retrieval Experiments

In this section we present the results of our retrieval experiments in a series of comparison as follows. First, we examine the effectiveness of term weighting components for Japanese text retrieval. Second, we present results for retrieval using multiple evidence combination techniques for documents and queries. Finally, we examine query expansion for Japanese text routing using relevance information.

All results show precision at ranked cutoff of 5, 10, 15 and 20 documents, and standard TREC average precision.

Clearly for such small collections the specific figures are neither reliable nor significant. Therefore 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. The following sections report retrieval results for the BMIR-J1 and BMIR-J2 collections using morphological segmentation and character-based indexing.

Morphological Segmentation Indexing Table 1 shows retrieval performance for BMIR-J1 with text indexing using morphological analysis; Table 2 shows corresponding figures for BMIR-J2. In each case for *cw* weighting the values of $K1$ and b have been optimized individually for each collection.

Character-Based Indexing Table 3 shows retrieval performance for BMIR-J1 with character-based indexing; Table 4 shows corresponding figures for BMIR-J2. Again, for *cw* weighting the values of $K1$ and b are optimized for the individual collections.

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0.423	0.433	0.480	0.480
	10 docs	0.322	0.348	0.363	0.367
	15 docs	0.251	0.274	0.290	0.288
	20 docs	0.210	0.228	0.238	0.237
Av Precision		0.405	0.450	0.491	0.496

cw: $K1 = 1.0$, $b = 0.2$

Table 1: Retrieval precision values for BMIR-J1 using Morphological Indexing.

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0427	0.443	0.503	0.503
	10 docs	0.320	0.352	0.382	0.382
	15 docs	0.251	0.287	0.296	0.296
	20 docs	0.210	0.239	0.253	0.253
Av Precision		0.405	0.474	0.514	0.514

cw: $K1 = 1.0$, $b = 0.0$

Table 3: Retrieval precision values for BMIR-J1 using Character-Based Indexing.

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0.464	0.496	0.580	0.580
	10 docs	0.420	0.460	0.506	0.510
	15 docs	0.388	0.431	0.464	0.464
	20 docs	0.353	0.393	0.423	0.423
Av Precision		0.351	0.403	0.442	0.441

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

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

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0.460	0.516	0.588	0.584
	10 docs	0.418	0.468	0.508	0.508
	15 docs	0.389	0.424	0.463	0.464
	20 docs	0.355	0.380	0.420	0.417
Av Precision		0.351	0.406	0.443	0.442

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

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

Observations From the results in Tables 1 – 4 it can be seen that *cfw* is generally effective for Japanese text retrieval using either form of indexing. In all cases there is 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.

BM25 parameter values are of the same order as those reported by other researchers [18], although there is some difference between the two collections. Also from Tables 1 – 4 it can be seen that there is little difference in retrieval performance for document length measurement in Morphs or Chars.

For BMIR-J1 we observe the character-based indexing to be slightly more effective for retrieval than morphological segmentation. However, for BMIR-J2 the results are almost identical for the alternative forms of indexing. Other experiments using additional requests with BMIR-J2 show that the comparison between indexing methods can be very sensitive to behaviour for individual queries [11]. It has been observed in previous work [6] [18] that character-based appears to be better. Our results generally support this conclusion, but clearly more research is needed on larger collections.

6.2 Data Fusion

Tables 5 and 6 show data fusion retrieval performance for BMIR-J1 and BMIR-J2 respectively. All figures are shown for optimal weighting of retrieved document lists from morphological segmentation and character-based indexing. In general these suggest that a Morphs:Chars ratio of 1:2 gives best retrieval performance for BMIR-J1, but a ratio nearer 1:1 is better for BMIR-J2. These ratios could be anticipated from the observed retrieval behaviour of the indexing methods in isolation.

The small performance gain here indicates combination of retrieval using morphological segmentation and character-based indexing may be useful, particularly where the contributing systems have similar individual performance levels. However, in [5] it is reported that data combination yielded a 10% improvement in Japanese text retrieval performance. This is significantly higher than the improvement found so far in our work. The reasons for this apparent difference are not clear, although it suggests that we should explore alternative means of data combination.

6.3 Query Combination

Tables 7 and 8 show retrieval performance with optimized data fusion for queries segmented using the simple rule based technique. These results are substantially worse than those obtained previously using queries segmented using morphological analysis. Analysis of retrieval results for individual queries showed that often no documents at all are retrieved, leading to a much reduced average. However the new queries contain many search items not present in the original queries and combined queries may be more useful overall.

Tables 9 and 10 show retrieval performance for the combined queries. The figures here are similar to those observed for the original queries segmented using morphological analysis. However, closer analysis of individual queries shows that some are substantially improved by the use of query combination while others are much worse.

There are several possible reasons for the failure of query combination to improve retrieval performance. First the documents are indexed without using the rule-based segmentation method, thus many of the compound words introduced in the combined queries may be rare in the collections. This rarity may often lead to them having high term weights and make them likely to exert undue influence on matching scores. In further query combination experiments we explored downweighting of terms derived using rule-based segmentation. The results of these experiments showed downweighting to give a marginal improvement in performance, but not sufficient to alter our overall conclusions.

For the remainder of this paper we use only the original morphologically segmented queries.

Weight Scheme		uw	cfw	cw	
				Morphs	Chars
Prec.	5 docs	0.427	0.447	0.487	0.493
	10 docs	0.320	0.357	0.380	0.385
	15 docs	0.252	0.286	0.307	0.307
	20 docs	0.210	0.238	0.249	0.249
Av Precision		0.406	0.474	0.511	0.515

cw: $K1 = 1.0$, $b = 0.2$

Table 5: Retrieval precision values for BMIR-J1 using Data Fusion.

Weight Scheme		uw	cfw	cw	
				Morphs	Chars
Prec.	5 docs	0.460	0.504	0.580	0.576
	10 docs	0.418	0.466	0.508	0.506
	15 docs	0.389	0.424	0.463	0.461
	20 docs	0.355	0.384	0.424	0.427
Av Precision		0.352	0.410	0.449	0.448

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

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

Weight Scheme		uw	cfw	cw	
				Morphs	Chars
Prec.	5 docs	0.340	0.347	0.350	0.350
	10 docs	0.238	0.257	0.267	0.267
	15 docs	0.183	0.194	0.204	0.206
	20 docs	0.144	0.152	0.164	0.164
Av Precision		0.304	0.317	0.342	0.343

cw: $K1 = 1.0$, $b = 0.2$

Table 7: Retrieval precision values for BMIR-J1 using Data Fusion and rule segmented queries.

Weight Scheme		uw	cfw	cw	
				Morphs	Chars
Prec.	5 docs	0.364	0.384	0.424	0.424
	10 docs	0.314	0.338	0.366	0.366
	15 docs	0.292	0.313	0.333	0.333
	20 docs	0.274	0.292	0.314	0.315
Av Precision		0.264	0.293	0.319	0.319

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

Table 8: Retrieval precision values for BMIR-J2 using Data Fusion and rule segmented queries.

6.4 Routing Experiments

The morphologically segmented terms appearing in TCIR-N1 were ranked separately using $r(i)$, $rtf(i)$ and $rsv(i)$. In our first experiments retrieval performance for expansion terms chosen using each selection method was compared where up to the top 30 ranked expansion terms were added to each query. Expansion term weights were varied by a scalar constant downweighting factor q between 1.0 and 0.1. Overall expansion term ranking using $rsv(i)$ was shown to be consistently better than either of the other methods.

Tables 11 and 12 show retrieval performance for BMIR-J1 using the original queries and queries expanded by a fixed number of terms using $rsv(i)$ term selection ranking. All results use data fusion with a simple ratio of 1:1 between indexing sources for all query types. All retrieval results use *cw* indexing with document length measured in Chars. Table 11 shows retrieval performance using $K1 = 1.0$ and $b = 0.2$, the optimal values found for BMIR-J1. Table 12 shows retrieval performance for the more realistic routing situation where the optimal values for TCIR-N1 are used $K = 0.5$ and $b = 0.8$. The optimal expansion term query weight q was found to be 0.25. Using $q = 1.0$ actually led to a small decrease in retrieval performance.

From these results it can be seen that query expansion is quite effective for this routing task. Average precision is shown to be improved by more than 10% for retrieval using the optimal BM25 parameters for TCIR-N1. Rather surprisingly retrieval performance using $K1 = 0.5$ and $b = 0.8$ with query expansion appears to be slightly better than using the optimal parameters found for BMIR-J1 with the original queries. Obviously individual figures must be treated with caution, but the general trend seems to indicate that collection dependent setting of these parameters is less critical when more detailed queries are available. This is good news for routing applications, particularly if this is shown to be the case for larger collections.

The optimal average number of expansion terms appears to be around 20, but examination of the retrieval results for each query shows a wide variation in the optimal number for individual queries.

For a slightly more complex approach using selection of the optimal number of expansion terms for each query up to a maximum of 20 terms, a small improvement in BMIR-J1 retrieval performance was observed for $q = 1.0$, suggesting that this approach had potential to do better than the simpler method. However, while there was some further improvement in retrieval performance when the value of q was reduced, the best performance figures obtained were lower than those already observed in Tables 11 and 12 for the simpler expansion method.

Finally, expansion term selection using a more complex, and significantly more expensive, method was investigated. The effect of adding up to each of the top 20 $rsv(i)$ expansion terms for the query was explored sequentially, with the value of q optimised individually for each term during sequential iteration. If the addition of the term improved retrieval performance on the TCIR-N1 training set it was retained otherwise it was discarded. If more than 5 terms in a row were discarded training for this query was stopped. This training procedure is similar to that used for TREC routing experiments in [22]. Once again, the retrieval results obtained using these optimised queries were not as good as those achieved using the much simpler fixed number of terms method shown in Tables 11 and 12.

Observations These results indicate that query expansion with $rsv(i)$ works well for Japanese text. However, it is a little surprising that the largest improvement in retrieval performance was found with simplest method of term selection and weighting for query expansion. The TCIR-N1 training set is very small, as of course is the BMIR-J1 test set, and despite the fact that the news articles in BMIR-J1 and TCIR-N1 both come from the *Nikkei* newspaper there may be some mismatch between them since they contain articles from different periods. Thus attempts to use more sophisticated query expansion training algorithms may lead to overfitting to the training set. In this case it is possible that the simplest query expansion procedure produces more general and useful expanded queries. Clearly more experiments with

		original	expanded (no. expansion terms)					
			5	10	15	20	25	30
Prec.	5 docs	0.493	0.543	0.540	0.537	0.537	0.547	0.525
	10 docs	0.382	0.403	0.408	0.412	0.423	0.425	0.408
	15 docs	0.303	0.330	0.323	0.332	0.336	0.331	0.333
	20 docs	0.248	0.268	0.273	0.273	0.279	0.277	0.280
Av Precision		0.512	0.550	0.557	0.552	0.560	0.558	0.544
% change		—	+7.4%	+8.8%	+7.8%	+9.4%	+9.0%	+6.3%

cw: $K1 = 1.0$, $b = 0.2$, $q = 0.25$

Table 11: Retrieval precision values for BMIR-J1 using query expansion with *cw* weighting and *rsv* term selection.

		original	expanded (no. expansion terms)					
			5	10	15	20	25	30
Prec.	5 docs	0.460	0.493	0.507	0.543	0.550	0.547	0.533
	10 docs	0.362	0.398	0.398	0.402	0.417	0.407	0.403
	15 docs	0.296	0.322	0.331	0.332	0.332	0.333	0.333
	20 docs	0.247	0.268	0.270	0.273	0.276	0.276	0.275
Av Precision		0.498	0.536	0.549	0.550	0.564	0.558	0.551
% change		—	+7.6%	+10.2%	+10.4%	+13.3%	+12.0%	+10.6%

cw: $K1 = 0.5$, $b = 0.8$, $q = 0.25$

Table 12: Retrieval precision values for BMIR-J1 using query expansion with *cw* weighting and *rsv* term selection.

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0.437	0.460	0.473	0.477
	10 docs	0.325	0.355	0.373	0.376
	15 docs	0.262	0.289	0.299	0.300
	20 docs	0.213	0.236	0.246	0.245
Av Precision		0.419	0.477	0.498	0.498

cw: $K1 = 1.0$, $b = 0.4$

Table 9: Retrieval precision values for BMIR-J1 using Data Fusion and Query Combination.

Weight Scheme		*uw*	*cfw*	*cw*	
				Morphs	Chars
Prec.	5 docs	0.456	0.480	0.532	0.528
	10 docs	0.408	0.438	0.460	0.460
	15 docs	0.379	0.400	0.428	0.425
	20 docs	0.353	0.375	0.406	0.404
Av Precision		0.354	0.405	0.437	0.436

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

Table 10: Retrieval precision values for BMIR-J2 using Data Fusion and Query Combination.

larger collections are needed, but our results so far are nevertheless encouraging.

Similar improvements in retrieval performance for Japanese were observed for query expansion using an association thesaurus in [7]. An advantage of out work is that we do not need a thesaurus and potentially we can be responsive to the relevance judgements of our users.

7 Conclusions and Further Work

This paper has reported an experimental investigation into the use of probabilistic information retrieval techniques for Japanese text retrieval and routing. The results of these experiments suggest that these techniques are highly applicable to the Japanese language, although further investigation with much larger collections is required to establish our findings conclusively.

The experiments indicate that there is little difference in retrieval performance between text indexing using word-level or character-level analysis, but suggest that combination of these techniques can lead to a small overall improvement in retrieval performance. For further investigations in indexing it would be interesting to explore statistical segmentation techniques, such as those described in [18]. In addition to a straightforward comparison of retrieval performance using indexing with statistical segmentation, "terms" derived using this technique might be used as a source of query expansion, or perhaps in a hybrid segmentation method in combination with morphological analysis and character-based indexing.

In our current experiments we are extending our exploration of relevance feedback in Japanese to term reweighting both in isolation and in combination with query expansion. In addition, we intend to build a large collection based on BMIR-J2 to enable us to investigate incremental user feedback over time, taking work such as [1] as a starting point.

Finally, if relevance feedback is to be useful in the operational NEAT system we must collect relevance information from our users. Since many users may not provide such information on a regular basis it may be useful to explore automated analysis of their browsing behaviour as described in [14].

References

[1] J. Allan. Incremental Relevance Feedback for Information Filtering. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 270–278, Zurich, 1996. ACM.

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

[3] 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.

[4] L.-F. Chien. Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–120, Seattle, 1995. ACM.

[5] W. B. Croft, J. Broglio, and H. Fujii. Applications of Multilingual Text Retrieval. In *Proceedings of the 29th Annual Hawaii International Conference on System Sciences (Digital Document Track)*, pages 98–106, 1995.

[6] H. Fujii and W. B. Croft. A Comparison of Indexing Techniques for Japanese Text Retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 237–246, Pittsburgh, 1993. ACM.

[7] C. Han, H. Fujii, and W. B. Croft. Automatic Query Based Expansion for Japanese Text Retrieval. Technical Report 95-45, University of Massachusetts, 1994.

[8] D. K. Harman, editor. *The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD, 1996. NIST.

[9] D. K. Harman and E. M. Voorhees, editors. *The Fifth Text REtrieval Conference (TREC-5)*, Gaithersburg, MD, 1997. NIST.

[10] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 30–38, Zurich, 1996. ACM.

[11] G. J. F. Jones, T. Sakai, M. Kajiura, and K. Sumita. First Experiments on the BMIR-J2 Collection using the NEAT System. In *Information Processing Society of Japan Joint SIG DBS and SIG FI Workshop*, Yokohama, 1998. IPSJ.

[12] 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. In Japanese.

[13] J. H. Lee and J. S. Ahn. Using n-Grams for Korean Text Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 216–224, Zurich, 1996. ACM.

[14] M. Morita and Y. Shinoda. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–281, Dublin, 1994. ACM.

[15] J.-Y. Nie, M. Brisebois, and X. Ren. On Chinese Text Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–233, Zurich, 1996. ACM.

[16] Y. Ogawa. Effective and efficient document ranking without using a large lexicon. In *Proceedings of the 22nd Very Large DataBase (VLDB) Conference*, Bombay, 1996.

[17] Y. Ogawa and M. Iwasaki. A New Character-based Indexing Method using Frequency Data for Japanese Documents. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 121–129, Seattle, 1995. ACM.

[18] Y. Ogawa and T. Matsuda. Overlapping statistical word indexing: A new indexing method for Japanese text. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 226–234, Philadelphia, 1997. ACM.

[19] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46:359–364, 1990.

[20] 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.

[21] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, 1994. ACM.

[22] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In D. K. Harman, editor, *Overview of the Fourth Text REtrieval Conference (TREC-4)*, pages 73–96. NIST, 1996.

[23] 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 Information Processing Society of Japan National Conference*, pages 3–(301–302), Tokyo, 1997. IPSJ.

[24] T. Sakai, M. Kajiura, and K. Sumita. Profile Generation from Query Sentences for the NEAT Information Filtering System. In *Information Processing Society of Japan National SIG FI Workshop Notes 97-FI-4*, pages 83–88, Tokyo, 1997. IPSJ. In Japanese.

[25] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[26] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones, and K. Sparck Jones. Okapi at TREC-6: automatic ad hoc, VLC, routing, filtering and QSDR. In D. K. Harman and E. M. Voorhees, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, MD, 1998. NIST. To appear.