BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems

Tetsuya Sakai (Toshiba) *
Tsuyoshi Kitani (NTT DATA)
Yasushi Ogawa (Ricoh)
Tetsuya Ishikawa (Univ. of Library and Information Science)
Haruo Kimoto (NTT)
Ikuo Keshi (SHARP)
Jun Toyoura (Mitsubishi Electric)
Toshikazu Fukushima (NEC)
Kunio Matsui (Fujitsu Laboratories)
Yoshihiro Ueda (Fuji Xerox)
Takenobu Tokunaga (Tokyo Institute of Technology)
Hiroshi Tsuruoka (ERI, Univ. of Tokyo)
Hidekazu Nakawatase (NTT)
Teru Agata (Keio Univ.)
Noriko Kando (NACSIS)

Abstract BMIR-J2 is the first complete test collection generally available for evaluating Japanese information retrieval systems. BMIR-J2 features include a novel division of search requests based on various functions required to perform successful retrieval. BMIR-J2 and its smaller predecessor BMIR-J1 were constructed by a volunteer-based working group under the Information Processing Society of Japan. We hope that BMIR-J2 will come into wide use and that it will foster the development of Japanese IR systems.

1 Introduction

In Europe and the US, there are a variety of standard test collections available for objective evaluation of information retrieval(IR) systems [2]. However, in spite of the growing interest in IR research in Japan, no test collection for Japanese existed until 1993. Unlike European languages, there are no explicit word boundaries in Japanese text. Moreover, Japanese has several character classes, each of which fulfill different linguistic functions. Because of these language specific problems, Japanese test collections are indispensable for fostering research and development of Japanese IR systems.

In April 1993, we formed a working group under the Special Interest Group on Database Systems of the Information Processing Society of Japan(IPSJ SIGDBS), to design and construct Japanese test collections. In March 1996, with kind understanding from Nihon Keizai Shimbun, Inc., we released BMIR-J1, the first ever Japanese test collection [8] [11] . Although BMIR-J1 was a small preliminary collection, we collected many useful comments from the monitor users. This positive response encouraged the members of the working group to construct a much larger collection BMIR-J2, with revised search requests and criteria for relevance assessments. With kind understanding from Mainichi Shimbun and in collaboration with the government-funded Real World Computing Partnership(RWCP) [5], we released BMIR-J2 in March 1998 [10].

This paper introduces BMIR-J2 to the general IR community. It describes the issues we addressed and our experiences in building and using test collections. We hope that this paper encourages other researchers to experiment with BMIR-J2. Section 2 describes the features of BMIR-J2. Section 3 summarizes what we have learnt from building our preliminary collection BMIR-J1. Section 4 describes the construction process of BMIR-J2. Section 5 provides a discussion on the unique feature of BMIR-J2, namely *request complexity groups*. Section 6 gives conclusions.

2 Features of BMIR-J2

2.1 Overview

BMIR-J2 consists of 5,080 newspaper article IDs and 60 search requests, each with a list of relevant article IDs. The article IDs are as defined by Mainichi Shimbun, Inc., one of the major newspaper publishing companies in Japan. For copyright reasons, BMIR-J2 contains only the article IDs, so users must buy a copy of the Mainichi Shimbun CD-ROM'94 in order to use the actual contents of the articles for retrieval experiments. More details, including contact information for BMIR-J2 and the Mainichi Shimbun CD-ROM'94, can be found at `http://www.ulis.ac.jp/~ishikawa/bmir-j2/eindex.html`.

*Human Interface Laboratory, Toshiba R&D Center: email: tetsuya.sakai@toshiba.co.jp

2.2 Search Requests

A BMIR-J2 search request is a noun phrase describing a user's needs, which is augmented by a *narrative*, in TREC parlance. A narrative typically consists of a few sentences clarifying the meaning of the noun phrase and criteria for relevance assessments. For example, the narrative for the request "衛星放送 (satellite broadcasting)", in rough translation, is provided as follows:

```
Satellite broadcasting includes broadcasting by
broadcast satellites and that by communication
satellites. Even if a document contains keywords
such as ''satellite'' or ''TV with a BS tuner,''
it is considered nonrelevant unless it mentions
satellite broadcasting itself.
```

Compared to existing standard test collections for English, BMIR-J2 is unique in that the search requests are explicitly categorized into six *request complexity groups* according to various functions which the evaluated system might use to retrieve the relevant documents. Table 1 provides, for each complexity group, a brief description and an example of a request in translation. It is supposed that as we move along from group A to F, the requests become harder to satisfy and thus the average retrieval performance is likely to degrade. This feature was inherited from BMIR-J1 [11].

2.3 Relevance Levels

We provide three levels of relevance, namely, A, B and C. In an A-relevant document, the content of the search request is discussed as the main topic. Whereas in a B-relevant document, the content is only mentioned briefly at some point within the document. In evaluation of retrieval performance, users can choose whether to use A-relevant documents only or to use both A-relevant and B-relevant documents as the relevant sets, depending on their tasks and needs. In addition, we provide documents with relevance level C, which typically contain some terms that are associated with the search request but was eventually judged nonrelevant by the relevance assessors.

2.4 Statistical Soundness

Search requests with few relevant documents tend to influence the overall retrieval performance greatly. Therefore, in average-based evaluation using BMIR-J2, we encourage the users to use the *standard set*, that contain 50 requests with at least 5 relevant documents. The *additional set*, that contain the remaining 10 requests with few relevant documents, is provided as a separate file, and may be useful for close per-request analysis.

Table 2 summarizes and supplements the aforementioned information about BMIR-J2, together with information on the preliminary BMIR-J1 and TREC-4 collections for comparison [2]. It can be observed that BMIR-J2 is a small collection by current TREC standards. However, we believe that the relevance assessment data it contains are consistent and reliable. Although it is difficult to derive statistically sound results from such a small test collection, we argue that looking at general trends in the results can still be useful for exploring different retrieval techniques.

3 Lessons from BMIR-J1

The main purpose of releasing BMIR-J1 was to collect opinions from monitor users and set directions for constructing BMIR-J2. The 600 articles (See Table 2) were randomly chosen from the economics sections of Nihon Keizai Shimbun newspapers from 1993, and the search requests were devised by the working group members by browsing through the small document collection. The limited size of the collection enabled us to perform exhaustive relevance assessments for each search request without adopting a pooling method. Two relevance assessors were assigned to each request, who consulted the whole working group whenever there was disagreement in their assessments.

BMIR-J1 was released to 50 sites, 32 of which filled out and returned a questionnaire by e-mail to the working group. This included 11 sites to which each of the working group members belonged. An analysis of the questionnaire showed that the users found BMIR-J1 to be a useful test collection with reliable relevance assessments. However, as expected, many users said that we should enlarge the size of the collection for the sake of statistical soundness. Some also said that many of the search requests of BMIR-J1 are too hard for existing retrieval systems to deal with. There were also some other interesting opinions, some of which are summarized in Table 3.

4 Construction of BMIR-J2

The overall success of BMIR-J1 encouraged us to start working on the complete collection. Although other kinds of text sources such as patents and technical papers were also considered to be of importance by many monitor users, lack of funding and limited availability of text resources made us finally settle on newspaper articles again. With kind understanding from Mainichi Shimbun we decided to use a subset of Mainichi Shimbun newspaper articles from 1994. This enabled us to reuse many of the search requests developed for BMIR-J1. Table 4 presents information on the search requests inherited from BMIR-J1 and those newly developed for BMIR-J2. For some of the reused requests, the noun phrases and narratives were modified in order to disambiguate them and also to adjust the number of relevant documents.

Most of the requests newly developed for BMIR-J2 are relatively easy from the viewpoint of request complexity. This was intended to be a feedback from the questionnaire which suggested that BMIR-J1 contained too many difficult requests by current standards. Table 5 provides information on the distribution of requests over the complexity groups for BMIR-J1 and J2. It can be observed that, compared to BMIR-J1, BMIR-J2 contains more requests in the easier groups(A through C) and consequently fewer requests in the harder groups(D through F).

The relevance assessment process of BMIR-J2 was quite different from that of BMIR-J1 because of the

Table 1: Request Complexity Groups with Examples

Group	Description of Functions Required	Example Search Request
A	keyword search query expansion using thesauri	"liquid crystal"
B	numerical value/range comparison	"companies that plan to lay off more than 1,000 employees"
C	syntactic analysis	"construction of highways"
D	semantic/context analysis using linguistic knowledge	"price reduction due to the rise of yen"
E	knowledge processing using common sense/world knowledge	"export from countries in Southeast Asia to Japan"
F	using both D and E	"employment problems of women"

Table 2: Comparison of BMIR-J1, J2 and TREC-4

Collection	#Texts	Average #Terms per Text	#Requests	Average #Terms per per Request	Average #Relevant per Request
TREC-4	567,529	842.0	50	10	130

Collection	#Texts	Average #Characters per Text	#Requests	Average #Characters per Request (Noun Phrase)	Average #Characters per Request (including Narrative)	Average #A-relevant per Request	Average #A-relevant plus #B-relevant per Request
BMIR-J1	600	733.8	60	10.9	94.5	5.5	10.1
BMIR-J2	5,080	621.8	50+10(*1)	9.7(*2)	102.2	10.6	28.4

(*1) standard set + additional set.
(*2) The shortest noun phrase consists of 2 characters, and the longest consists of 26 characters.

enlarged size of the document collection. We had to give up the idea of exhaustive assessments and adopt a less resource-consuming method. Although the working group itself had no funding, RWCP agreed to assign the preliminary relevance assessment task to a subcontract company and cover the expenses. BMIR-J2 was completed as follows:

1. The full texts of 5,080 articles were selected from the economics, engineering, and industrial technologies sections of Mainichi Shimbun newspapers from 1994 as the document collection. Note that BMIR-J2 has a broader domain than J1.

2. Some working group members had their own information retrieval systems that allowed simple searches such as those using Boolean operators. For each candidate request, some recall-oriented search queries were devised by a relevance assessor and a working group member, and a set of candidate documents were obtained using one of the systems.

3. Preliminary relevance assessments were performed by the assessor on the candidate documents only.

4. Each request was assigned to a working group member, who checked the relevance assessments returned by the assessors and corrected them whenever necessary. For some requests, the working group asked the assessors to redo the relevance assessments after revising the recall-oriented query or clarifying the criteria for relevance assessments.

5. Finally, for some of the requests, a new working group member was assigned to cross-check the relevance assessments. As in the case of BMIR-J1, if there was disagreement between members, they consulted the whole working group.

5 Use of Request Complexity Groups

Because the search requests of BMIR-J2 are explicitly categorized into six complexity groups, the user can focus his attention on a complexity group with a particular function or compare results across different complexity groups. For example, if a system that performs fact retrieval using numerical comparisons is to be developed, group B can be used on its own for evaluation. Moreover, if the retrieval performance averaged over group D is considerably lower than that over group C, for example, this may imply that the basic retrieval methodology of the system is inadequate for advanced retrieval that go beyond syntactic analysis.

Using the preliminary BMIR-J1 collection, [9] and [14] showed that requests in groups D through F are actually significantly harder than those in groups A through C in terms of the traditional recall and precision measures. Recent work such as [15] showed that this is also the case for BMIR-J2. If BMIR-J2 comes into wide use, cases may arise in which, for instance, system X outperforms system Y for a particular complexity group but not for others.

Table 3: Excerpt from the BMIR-J1 Questionnaire Summary(*1)

Question/Answer	#Users	#Users excluding WG Members
Q2:Are the relevance levels (A,B and C) useful?		
yes	21	15
yes but level C is unnecessary	5	3
single relevance level is sufficient	2	1
there should be as many levels as possible	1	0
others	2	1
TOTAL	31	20
Q3:Is the number of search requests appropriate?		
there should be more	16	8
60 is sufficient	10	7
there are too many	1	1
others	1	1
TOTAL	28	17
Q8-1:What kinds of text should we add to our document collection?		
patents	14	11
technical papers/ abstracts	15	5
magazine articles	3	2
encyclopedia	3	2
manuals	3	3
web pages	3	1
system diagnosis/ logs	2	2
law reports	2	2
texts in foreign languages/ multilingual texts	2	2
TOTAL	47	30
Q8-2:Which domains should we add to our document collection?		
engineering/ computer science	16	9
politics/ international	4	3
legal matters	2	2
society/culture	2	2
hobbies/home affairs	2	1
advertisement/ press releases	2	1
editorials/ survey articles	1	1
miscellaneous domains	5	2
restricted domains	1	0
TOTAL	35	21

(*1)For Q8-1 and Q8-2, the users were allowed to choose more than one answer.

Table 4: Reused/New Requests in BMIR-J2

Query Origin	Standard	Additional	Total
inherited from J1 (not modified)	26	6	32
inherited from J1 (modified)	11	4	15
new	13	0	13
total	50	10	60

Table 5: Number of Requests in Each Complexity Group

Group	BMIR-J1	BMIR-J2(*1)
A	10	14+0=14
B	5	3+1=4
C	6	10+1=11
D	12	9+2=11
E	10	4+3=7
F	17	10+3=13
total	60	50+10=60

(*1)standard set + additional set

6 Conclusions

Research in English-based information retrieval has a history of some half a century. Compared to this, Japanese information retrieval is still in its infancy. Already, about 40 sites have completed the user registration of BMIR-J2, and we are beginning to see published research work based on BMIR-J2 [1] [6] [7] [12] [13] [15] [16] [17] [18] . We hope that BMIR-J2 will be widely used and that it will foster the development of advanced IR systems for Japanese. We also hope that our experiences will help in the development of other new test collections. In fact, inspired by our activities, the first TREC-like workshops for retrieval of Japanese text, namely IREX [3] and NTCIR [4] will start in 1999. The test collections associated with these workshops will be much larger than BMIR-J2 with around two to three hundred thousand documents. Some of the BMIR working group members are involved in planning these new projects.

The BMIR working group itself stopped its activities just after the release of BMIR-J2. Currently, the Third Subcommittee (Hypertext and Information Retrieval) under IPSJ SIGDBS is in charge of matters concerning BMIR-J2. The contact address is `bmir@rd.nacsis.ac.jp`.

Acknowledgements In the construction of BMIR-J1, Yoshifumi Masunaga(ULIS), Tomohiro Tanaka(NTT), Tadanobu Miyauchi(Fuji Xerox) and Seiji Miike(Toshiba) also took part as members of the working group. The authors would like to thank Professor Katsumi Tanaka, chair of IPSJ SIGDBS, for his continuous support of the Benchmark Database Working Group.

References

[1] Chun, Y. et al.: A Utility-based Information Retrieval System for User Information Usage - UBIR System, *Proc. of The 3rd International Workshop on Information Retrieval with Asian Languages*, pp. 22–27 (1998).

[2] Harman, D.(Moderator): Panel:Building and Using Test Collections, *Proc. of ACM SIGIR '96*, pp. 335–337 (1996).

[3] `http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html`

[4] `http://www.rd.nacsis.ac.jp/~ntcadm/index-en.html`

[5] `http://www.rwcp.or.jp/home-E.html`

[6] Jones, G. et al.: Experiments in Japanese Text Retrieval and Routing using the NEAT system, *Proc. of ACM SIGIR'98*, pp. 197–205 (1998).

[7] Jones, G. et al.: A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval, *Proc. of ACM SIGIR '99*, to appear (1999).

[8] Kimoto, H. et al.: Construction of a test collection for the evaluation of Japanese information retrieval systems(in Japanese), *IPSJ Transactions*, Vol. 40, No. 9, to appear (1999).

[9] Kitani, T. et al.: Information Retrieval Using a Full-text and Extracted Keywords(in Japanese), *IPSJ SIGFI Notes*, 96–FI–43(96–NL–115), pp. 129–134 (1996).

[10] Kitani, T. et al.: Lessons from BMIR-J2: A Test Collection for Japanese IR Systems, *Proc. of ACM SIGIR '98*, pp. 345–346 (1998).

[11] Matsui, K. et al.: Test Collection for Information Retrieval Systems from the Viewpoint of Evaluating System Functions, *Proc. of International Workshop on Information Retrieval with Oriental Languages*, pp. 42–47 (1996).

[12] Mochizuki, H. et al.: Passage-Level Document Retrieval Using Lexical Chains, *IPSJ SIGFI Notes*, 98–FI–51, pp. 39–46 (1998).

[13] Ogawa, Y. et al.: Overlapping Statistical Segmentation for Effective Indexing of Japanese Text, *Information Processing & Management*, to appear.

[14] Sakai, T. et al.: Generation and Evaluation of Search Queries Using Boolean Expressions and Document Structure for Information Filtering(in Japanese), *IPSJ Transactions*, Vol. 39, No. 11, pp. 3076–3083 (1998).

[15] Sakai, T. et al.: Application of Query Expansion Techniques in Probabilistic Japanese News Filtering, *Proc. of The 3rd International Workshop on Information Retrieval with Asian Languages*, pp. 46–55 (1998).

[16] Sakai, T. et al. : A Study on English-to-Japanese / Japanese-to-English Cross-Language Information Retrieval using Machine Translation (in Japanese), *IPSJ Transactions*, submitted (1999).

[17] Sugai, T. et al.: The Hierarchical Information Filtering Method and its Evaluation (in Japanese), *JSAI 12th National Conference Proceedings*, pp. 390–393 (1998).

[18] Yamada, H. et al.: A Proposal and Evaluation of a Text Retrieval Method Based on Numerical Value Extraction from Text(in Japanese), *IPSJ SIGFI Notes*, 99–FI–53, pp. 17–22 (1999).