Lessons from BMIR-J2: A Test Collection for Japanese IR Systems

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

Abstract BMIR-J2 is the first complete Japanese test collection available for use in evaluating information retrieval systems. It contains sixty queries and the IDs of 5080 newspaper articles in the fields of economics and engineering. The queries are classified into five categories, based on the functions the system is likely to use to interpret them correctly and retrieve relevant texts. This collection has two levels of relevance, topically relevant and partially relevant. Also discussed are design issues such as collection types and size. This collection and the principles derived in designing it should be helpful in the future development of new test collections.

1 Introduction

A variety of standard test collections are available for objective evaluation of information retrieval systems[1]. In spite of increasing interest in IR research in Japan, however, until now no Japanese test collection existed. In 1993, we formed a working group (WG) under the Special Interest Group of Database Systems in the Information Processing Society of Japan to develop a Japanese test collection. In 1996, a preliminary version called BMIR-J1 was distributed to fifty sites. We enlarged the collection size and revised the queries and relevance assessments based on comments from BMIR-J1 users. Distribution of BMIR-J2, the first complete Japanese test collection, started in March 1998.

In designing a test collection, it is important to consider to which types of IR systems it is applicable, what kinds of texts should be included and how relevant texts should be selected. Based on our experience, we also discuss these design issues.

2 Overview of BMIR-J2

2.1 Text selection

Initially, we considered text sources such as patent descriptions and technical papers for use in the collection. However, we settled on newspaper articles, since they are generally available on CD-ROMs and are widely used. Thus, BMIR-J2 uses only articles from the Mainichi Newspapers. In response to requests from the BMIR-J1 users, we focused on the fields of economics and engineering.

2.2 Query development

A total of sixty queries were developed in BMIR-J2. Each query consists of a natural language phrase describing a user's needs and additional comments augmenting it. In the following English translation, the first line, showing a query phrase, is followed by two narrative lines:

```
Q:F=oxoxo:"Utilizing solar energy"
Q:N-1:Retrieve texts mentioning uses of
      solar energy.
Q:N-2:Include texts concerning generating
      electricity and drying things with
      solar heat.
```

2.3 Query categorization

Information retrieval often requires a deep understanding of the user's needs. Sometimes, however, merely matching words appearing in the query with those in the texts is enough to retrieve relevant texts. Since a system's architecture ranges from simple word matching to rich natural language processing, it is worth categorizing queries according to the functions that the system uses to process them. This variety of categories allows BMIR-J2 users to select queries that their systems will be able to deal with. BMIR-J2 provides five categories, shown as "F=oxoxo" in the previous example. Each digit, denoted as "o" (necessary) or "x" (unnecessary), represents the following functions, starting from the left[3]:

- *The basic function*: The relevant texts can be retrieved simply using words extracted from the query or their thesaurus expansion.
- *The numeric range function*: The system will need to handle a numeric range description such as "reduction of more than one thousand employees."
- *The syntactic function*: Analyzing a syntactic relationship among query words will help understand the query.
- *The semantic function*: A semantic analysis will be required to understand the query. In a query, "a trend in the facsimile market," a system must determine how to identify a description as a "trend," since the word "trend" may not appear in the text.

*Laboratory for Information Technology, NTT Data Corporation. Email: tkitani@lit.rd.nttdata.co.jp.

Table 1: Comparisons with TREC-4

Collection	Number of Texts	Average Terms /Text(*1)	Number of Queries	Average Terms /Query(*1)	Average Relevant /Query
TREC-4	567529	842.0	50	10	130
BMIR-J1	600	733.8	60	10.9/94.5(*2)	5.5/10.1 (*3)
BMIR-J2	5080	621.8	60	9.7/102.2(*2)	10.6/28.4 (*3)

(*1) Average number of characters in BMIR. (*2) Noun phrase / noun phrase and narratives. (*3) Topically relevant / topically and partially relevant.

- *The world knowledge function*: World knowledge will be required to process the query. In a query, "a joint business operation between companies in different types of industries," the system must know that the companies belong to different types of industries. Such information is often missing in the texts or the system's lexicon.

2.4 Relevance assessments

For each query, relevancy of articles was assessed in the following four steps in general[2]:

1. Pre-screening of possibly relevant texts: A Boolean query was made manually in such a way that most relevant texts were likely to be retrieved. Results from one or two IR systems were merged.
2. Relevance assessments by database searchers: Several database searchers first assessed the relevancy from the results obtained in step 1. They also cross-checked their work.
3. Relevance assessments by one of the WG members: Relevant texts of each query from step 2 was checked and corrected by one of the WG members.
4. Relevance assessments by another WG member: Relevant texts from step 3 were cross-checked and corrected by another WG member.

3 Design issues and discussion

3.1 Types of test collection

BMIR-J2 allows batchwise evaluation of IR systems, which corresponds to the ad hoc task in TREC. Need is increasing for other types of test collection, such as by text categorization and interactive testing with users. These issues should be examined further.

3.2 Collection size

How many texts and queries, and how many relevant texts for each query, are necessary to make IR system evaluation statistically sound? Queries with few relevant texts tend to have a great influence on the overall system performance. Of the sixty queries in BMIR-J2, ten contained fewer than five relevant articles. They are provided as an additional set separated from a standard set containing fifty queries. BMIR-J2 users may include the additional set for system evaluation.

3.3 Text sources and domains

Texts from limited domains and specific sources allow controlled system evaluation. Texts from various domains and sources, on the other hand, allow an evaluation that is close to real-world situations. Choice of text selection is valuable, but other issues such as copyright of text sources and getting enough funding for the license, are often dominant. To get around the copyright issue, we included a list of article IDs, rather than full texts, in the distribution set. Users must get a copy of Mainichi Newspaper CD-ROM'94 themselves.

3.4 Relevance assessment process

As recent test collections go, BMIR-J2 is not particularly large[1]. Table 1 shows some comparisons with TREC-4. Even with 5080 articles and sixty queries, checking all possibilities manually is already overwhelming work. Thus, pre-screening relevant texts using IR systems was necessary in developing our collection. In order not to miss relevant texts in pre-screening, the pooling method using IR systems with different architectures is desirable. Our pooling method was limited as to the number and variety of architectures of IR systems used. Relevance results should be cross-checked by at least two people. From our experience, cross-checking gave us a chance to adjust relevance assessment criteria and keep them consistent among WG members.

4 Conclusion

BMIR-J2, the first complete Japanese test collection, is now available with a handling cost of 2,000 yen. More information can be found at `http://www.ulis.ac.jp/~ishikawa/bmir-j2`. It contains 5080 newspaper article IDs in the fields of economics and engineering, sixty queries, and their relevance assessments. We hope that BMIR-J2 will be widely used and that our experience will help in the development of other new test collections.

Acknowledgements This work is done in collaboration with the Real World Computing Project. We thank Professor Katsumi Tanaka for his continuous support on this project. Thanks are also due to Noriko Kando who joined the discussion for the development.

References

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

[2] T. Kitani et al. BMIR-J2 - A Test Collection for Evaluation of Japanese Information Retrieval Systems (in Japanese). In *Proc. of IPSJ SIG Notes, DBS-114-3*, pages 15–22, 1998.

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