Multi-lingual Multi-media Information Retrieval System

Shoji Mizobuchi, Sankon Lee, Fumihiko Kawano, Tsuyoshi Kobayashi, Takahiro Komatsu
Graduate School of Engineering, University of Tokushima
2-1 Minamijosanjima, Tokushima, 770-8506, Japan
+81-88-656-7486
{shoji, sam}@is.tokushima-u.ac.jp
Jun-ichi Aoe
Department of Information Science and Intelligent Systems, University of Tokushima
2-1 Minamijosanjima, Tokushima, 770-8506, Japan
+81-88-656-7486
aoe@is.tokushima-u.ac.jp

1. INTRODUCTION

In the first NACSIS Test Collection for Information Retrieval (NTCIR) Workshop [3][4][5], we participated in the ad hoc Information Retrieval (IR) task. In this participation, we focus on evaluating the Japanese IR engine of our prototype Multi-Lingual Multi-Media Information Retrieval (MLMMIR) system using NTCIR, which is the first large-scale test collection in Japanese.

A MLMMIR system is an integrating system of Multi-Lingual Information Retrieval (MLIR) techniques [2][9] and Multi-Media Information Retrieval (MMIR) techniques [1]. It is capable of mutually retrieving informative data that is represented by many languages and various kinds of media by crossing boundaries among them.

Aiming to realize such a system, we build the prototype MLMMIR system, called BOSS [7]. BOSS can retrieve a mixture of various kinds of useful data by first translating the contents that are included in lingual and media data into Japanese keywords, and then providing them to the Japanese IR engine that we developed. The performance of the system depends heavily on the Japanese IR engine, and the overall system is now under construction. Therefore, we limit the range of evaluation to not the overall system but the Japanese IR engine of the system.

We submitted one official ad hoc run, JALAB, to NACSIS. However, it was excessively bad because the collection of documents that was used in automatic term indexing is not the ones of NTCIR, but the corpus that we collected by ourselves. Therefore, we report new five ad-hoc runs, JALAB1, JALAB2, JALAB3, JALAB4, and JALAB5, in addition to JALAB. They are the runs that were performed using the index terms that were extracted from all documents of NTCIR.

The remainder of this paper is organized as follows. The next section describes the overview of the system. Section 3 describes keyword extraction methods for multi-lingual data and multi-media data. Section 4 describes the Japanese IR engine of the system. Section 5 describes the evaluation results for the ad hoc task. Finally, Section 6 describes conclusion.

2. OVERVIEW OF THE SYSTEM

BOSS can retrieve the information that is relevant to a query from a database and display the result in relevant order. The information that is inputted as a query is lingual data such as Japanese and English texts, or media data such as images and sounds. In addition, the information that is stored in the database and is returned in the result is a mixture of these data.

To break down the walls that stand among lingual and media data, BOSS retrieves relevant information through Japanese keywords. In the searching process, Japanese keywords are first extracted from any kinds of information that the system supports, and then they are provided for the conventional IR system. By doing in such a way, it is possible to search for various kinds of information mutually.

Figure 1 shows the overview of BOSS. It consists of three main modules: the user interface module, the keyword extraction module, and the IR engine. Each of them is easily explained below.

- User interface module: The user interface module takes various kinds of command inputted by a user and displays the result of executing it. If data objects are included in the result, they are outputted using a suitable method for each of their formats.
- Keyword extraction module: The keyword extraction translates a data object that is sent from the user interface module into Japanese keywords using the extraction method and the conversion dictionary that specialize in its format. Conversion dictionaries are used to translate the features that are extracted from a data object into Japanese keywords except that it is a Japanese text. The detail is described in Section 3.
- IR engine: The IR engine retrieves data objects that are relevant to Japanese keywords using the inverted index. The relevance between the keywords and each of data objects that are stored in the database are computed using a vector-space method [6]. The inverted index is used to compute these relevances fast. The detail is described in Section 4.

3. KEYWORD EXTRACTION

The input data is translated into some Japanese keywords in the keyword extraction module. This process is performed using two different methods for lingual data and media data.

3.1 Method for Multi-lingual Data

In the method for lingual multi-data, the keyword extraction that is specialized to the language used in a lingual data object is

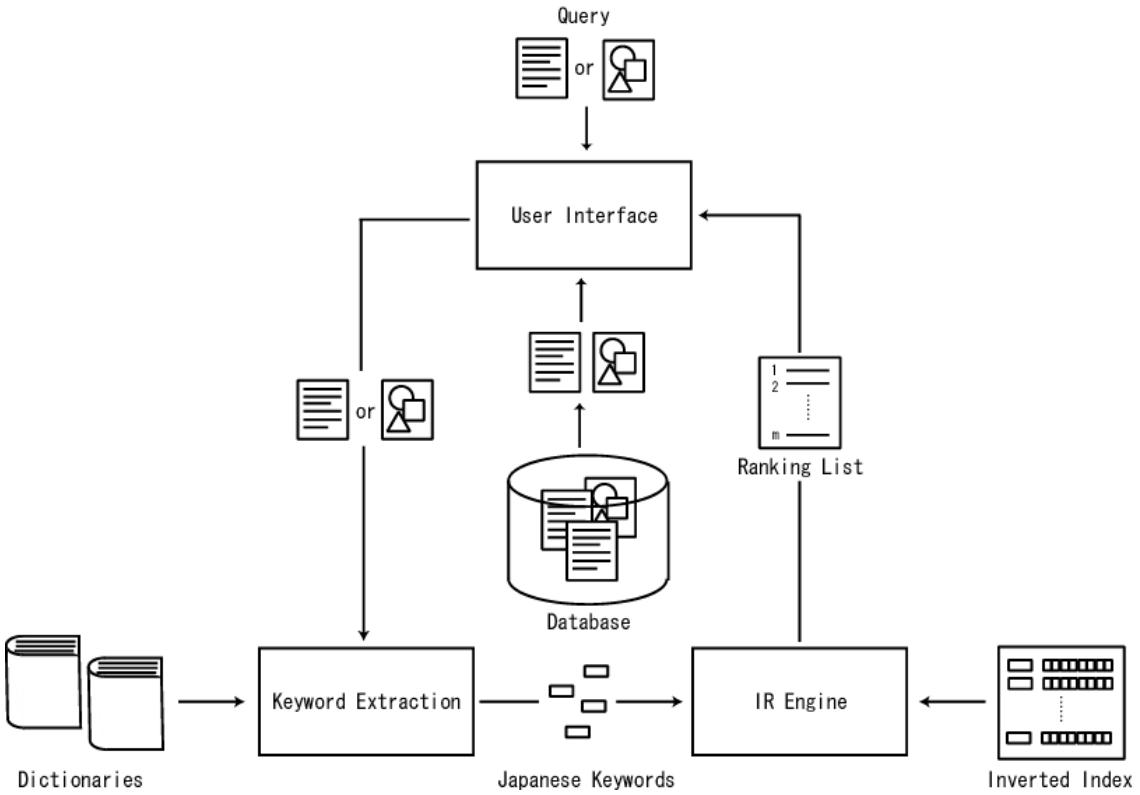

Figure 1. Overview of BOSS

performed. If Japanese is used in the data object, Japanese keywords are gotten after this process. On the other hands, if other language is used, the keywords that are extracted in this process are then translated into Japanese keywords using a bilingual dictionary.

In the keyword extraction specialized to a language, keyword candidates are recognized, modified, and reduced only using surface information such as character type and frequency. This method can not extract the words that occur in the data object as keywords, but it can extract keywords fast because it does not require complex linguistic analysis.

In the translation of keyword candidates, the bilingual dictionary including only nouns of each language is used. In the case that there are some Japanese words equivalent to a word in a certain language, only one Japanese word with a general meaning is selected.

3.2 Method for Multi-media Data

In the method for media data, the signature that identifies a certain data object is first made from the inputted data, and then the Japanese keywords that are associated with the signature in advance are retrieved. A signature is a list of bits of fixed length and consists of several attributes of files such as file names, file sizes, and bit information that is extracted from the data itself. The assignment of Japanese keywords to each media data is performed manually.

4. JAPANESE IR ENGINE

The Japanese IR engine searches for data related to the keywords that are extracted from a query, and outputs the result in descending order of its relevance to the keywords. The relevance between keywords and each stored data is computed using the vector-space model. In this model, both of them are represented by vectors on a vector space. In the IR engine, index terms that correspond to each element of a vector are not keywords, but strings are further extracted from the keywords. They are selected in advance through learning process on test documents.

4.1 Vector Representation

The contents of queries and stored data objects are represented by vectors. Each element of a vector is associated with a particular character bigram. If a character bigram appears in the extracted Japanese keywords, the corresponding element of a vector is set to 1. Otherwise, the element is set to 0.

Vectors that are generated for stored data objects are retained in an inverted index to speed up the computation of their relevance to a query. The inverted index is a table identifying data objects corresponding to index terms. Each entry of the index includes a particular index term and the corresponding list of data object numbers. The use of the inverted index enables us to search for relevant data objects quickly because it loses the need of computation of relevance for all vectors.

Index terms are selected in advance through automatic indexing process for data objects of a collection. In this process, keywords

are first extracted from the data objects. Then, character bigrams are further extracted from the keywords. The frequency of each unique bigram in the keywords are computed. Finally, the indicated number of bigrams with high frequencies is selected as index terms.

4.2 Searching in Relevance Order

Let $q = (q_1, q_2, \ldots, q_n)$ and $d = (d_1, d_2, \ldots, d_n)$ be the vector for a query and the vector for a stored data object, respectively. The relevance r between them is computed as follows.

$$r = \frac{q \bullet d}{N(q) + N(d) - q \bullet d}$$

where $q \bullet d$, $N(q)$, $N(d)$ are computed as follows.

$$q = \sum_{i=1}^{n} q_i \times d_i$$

$$N(q) = \sum_{i=1}^{n} q_i$$

$$N(d) = \sum_{i=1}^{n} d_i$$

5. EVALUATION

5.1 Demonstration of BOSS

BOSS works on Windows NT and handles texts in English, Japanese, and Korean, images of JPEG format, and sounds of MIDI format. Figure 2 shows the result of retrieving information that is relevant to the English text about a golf tournament. The upper left, bottom, and upper right windows in the figure are the content of the English text, the list of data objects that relevant to it, and the content of the most relevant Japanese text in the list, respectively. Figure 3 shows the result of retrieving information that is relevant to the jpeg image, the portrait of Mozart. The list at the bottom window in the figure includes several kinds of data objects like texts, images, and sounds. The first relevant data object is a sound data. When clicking a mouse on the play button at the upper right of the window, music gets started. Like this, BOSS can retrieve various kinds of data objects for a particular data object.

5.2 Results of the Ad Hoc Task

In NTCIR-1, we participate the ad hoc IR task to evaluate the Japanese IR engine of BOSS. In the task, documents for each of 53 topics are retrieved, and top 1,000 documents that are retrieved for each topic are judged whether they are relevant to the topic or not. This task is performed using the NTCIR.

The NTCIR is presented by NACSIS to enhance research in Japanese text retrieval. It consists of documents, search topics, and relevance assessments for each search topic. The documents are 339,483 abstracts taken form NACSIS Academic Conference Papers Database. The search topics are formatted descriptions of user's needs. They consist of 5 fields: Title, Description, Narrative, Concept and Field. The relevance assessments are lists of the documents that are judged relevant to each of search topics.

We submitted one official ad hoc run, JALAB, to NACSIS. After this submission, we perform new five ad hoc runs, JALAB1,

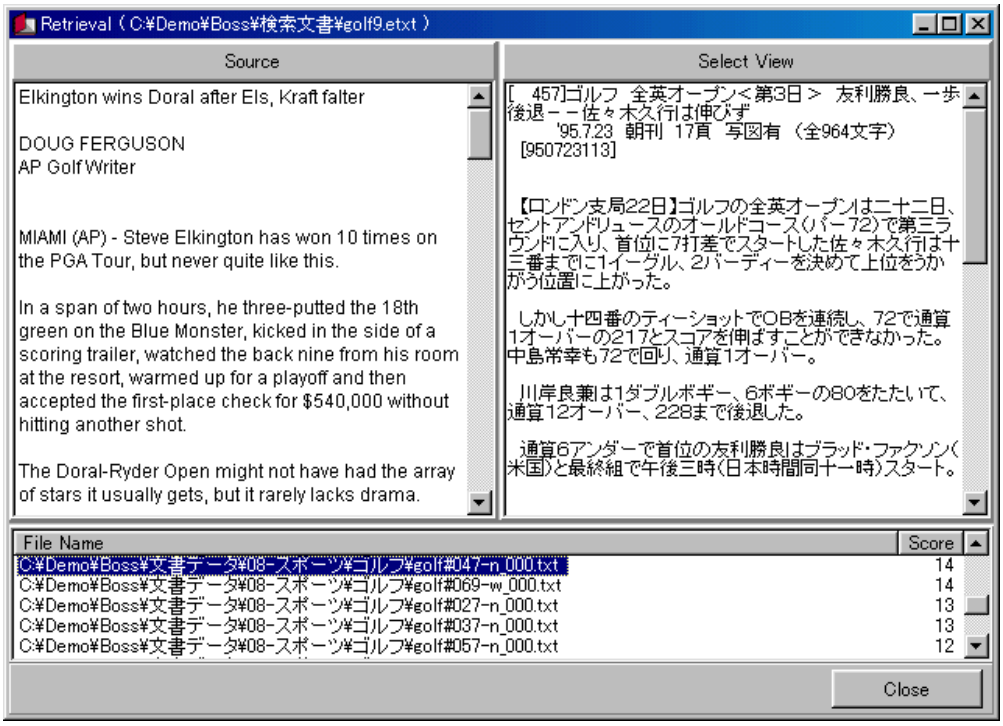

Figure 2. Result of retrieving information that is relevant to the English text about a golf tournament

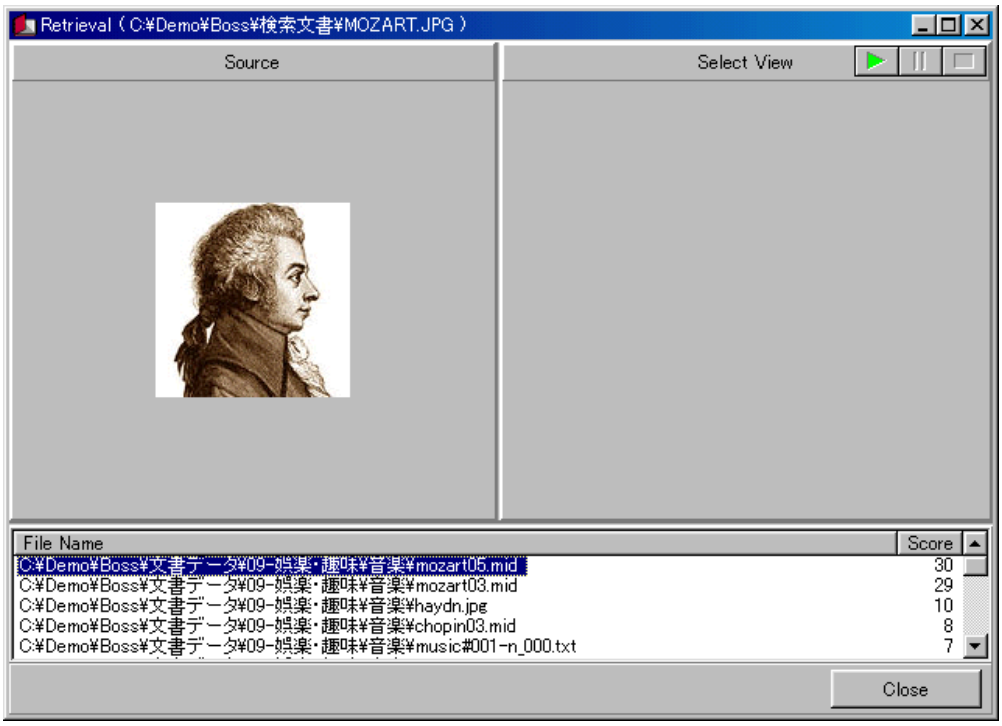

Figure 3. Result of retrieving information that is relevant to the jpeg image, the portrait of Mozart

JALAB2, JALAB3, JALAB4, and JALAB5. The contents of the six runs are as follows.

- JALAB (submitted): The Aoe laboratory (AL) corpus is used in the selection of index terms. The AL corpus is a collection of about 15,000 documents that we collected from various fields by itself. The total number of the index terms that are extracted from the documents in the corpus is 56,000. In this run, the description field is used for each topic.
- JALAB1 (new): The collection of documents in the NTCIR is used in the selection of index terms. The total number of index terms that are extracted from the documents in the collection is 366,886. In this run, the title field is used for each topic.
- JALAB2 (new): The used documents in the selection of index terms and the total number of index terms are the same as JALAB1. In this run, the description field is used for each topic.
- JALAB3 (new): The used documents in the selection of index terms and the total number of index terms are the same as JALAB1. In this run, the narrative field is used for each topic.
- JALAB4 (new): The used documents in the selection of index terms and the total number of index terms are the same as JALAB1. In this run, the title, description, and narrative fields are used for each topic.
- JALAB5 (new): The used documents in the selection of index terms and the total number of index terms are the same as JALAB1. In this run, the description and concept fields are used for each topic.

The performance of the IR engine is measured in precision and recall. Precision is the number of relevant data objects retrieved divided by the total number of data objects retrieved. Recall is the number of relevant data objects retrieved divided by the total number of relevant data objects.

The results of the ad hoc task are as follows. Table 1 and Table 2 show the summary of the results for the above 6 runs using the relevance assessments, Relevant qrels (JE-1) and Partial Relevant qrels (JE-2), respectively. Figure 4 and Figure 5 show the recall precision curves for the results using JE-1 and JE-2, respectively.

6. CONCLUSION

In this paper, we have described our developing MLMMIR system, BOSS, and reported the evaluation results of the Japanese IR engine of BOSS for the ad hoc task in NTCIR-1.

With respect to BOSS, we have just explained the overview of the system because BOSS is not so complete as the overall evaluation can be made for the system. Completing the system, many existing and new techniques must be introduced into it.

With respect to the evaluation, we have actually recognized that it is difficult to retrieve relevant documents in the way of weighting index terms only by using the information of whether they occur in documents or not. We need to introduce term weighting methods like TF-IDF[8].

Table 1. Result of the ad-hoc task for JE-1

Run ID	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Run desctiption	Automatic	Automatic	Automatic	Automatic	Automatic	Automatic
Number of topics	53	53	53	53	53	53
Used fields of topics	Description	Title	Description	Narrative	Title Description Narrative	Description Concept
Collection used in term indexing	AL Corpus	NTCIR1	NTCIR1	NTCIR1	NTCIR1	NTCIR1
Number of index terms	56,000	366,886	366,886	366,886	366,886	366,886

Total number of documents over all topics						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Retreived:	52,051	53,000	53,000	53,000	53,000	53,000
Relevant:	1,910	1,910	1,910	1,910	1,910	1,910
Rel-ret:	84	745	758	613	699	825

Recall Level Precision Averages						
Recall	Precision					
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
0.00	0.0075	0.3945	0.4461	0.3664	0.4101	0.4505
0.10	0.0064	0.3025	0.3130	0.2368	0.2929	0.3477
0.20	0.0049	0.2213	0.2263	0.1691	0.1792	0.2597
0.30	0.0025	0.1630	0.1762	0.1161	0.1305	0.2100
0.40	0.0014	0.1278	0.1366	0.0942	0.0902	0.1564
0.50	0.0006	0.1007	0.1124	0.0766	0.0793	0.1331
0.60	0.0002	0.0789	0.0817	0.0506	0.0587	0.1061
0.70	0.0000	0.0602	0.0567	0.0281	0.0416	0.0827
0.80	0.0000	0.0422	0.0448	0.0176	0.0309	0.0522
0.90	0.0000	0.0177	0.0188	0.0118	0.0032	0.0161
1.00	0.0000	0.0119	0.0154	0.0104	0.0023	0.0130

Average precision over all relevant docs						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Non-interpolated	0.0017	0.1227	0.1345	0.0937	0.1032	0.1535

Document Level Averages						
Document Level	Precision					
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
At 5 docs	0.0000	0.2113	0.2264	0.1736	0.1849	0.2491
At 10 docs	0.0000	0.1849	0.1830	0.1264	0.1547	0.2094
At 15 docs	0.0000	0.1560	0.1673	0.1270	0.1346	0.1736
At 20 docs	0.0090	0.1349	0.1472	0.1142	0.1255	0.1585
At 30 docs	0.0025	0.1195	0.1283	0.0975	0.1094	0.1346
At 100 docs	0.0034	0.0685	0.0702	0.0491	0.0562	0.0696
At 200 docs	0.0029	0.0459	0.0458	0.0320	0.0368	0.0460
At 500 docs	0.0022	0.0239	0.0242	0.0182	0.0209	0.0246
At 1,000 docs	0.0016	0.0141	0.0143	0.0116	0.0132	0.0156

R-Precision (precision after R docs retrieved (where R is the number of relevant documents))						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Exact	0.0003	0.1367	0.1497	0.1084	0.1327	0.1676

Table 2. Result of the ad-hoc task for JE-2

Run ID	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Run desctiption	Automatic	Automatic	Automatic	Automatic	Automatic	Automatic
Number of topics	53	53	53	53	53	53
Used fields of topics	Description	Title	Description	Narrative	Title Description Narrative	Description Concept
Collection used in term indexing	AL Corpus	NTCIR1	NTCIR1	NTCIR1	NTCIR1	NTCIR1
Number of index terms	56,000	366,886	366,886	366,886	366,886	366,886

Total number of documents over all topics						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Retreived:	52,051	53,000	53,000	53,000	53,000	53,000
Relevant:	2,345	2,345	1,910	1,910	1,910	1,910
Rel-ret:	164	998	978	820	902	1,052

Recall Level Precision Averages						
Recall	Precision					
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
0.00	0.0194	0.4347	0.4892	0.4325	0.4685	0.4896
0.10	0.0108	0.3288	0.3284	0.2598	0.3125	0.3493
0.20	0.0053	0.2321	0.2436	0.1937	0.2099	0.2749
0.30	0.0034	0.1928	0.1900	0.1421	0.1395	0.2187
0.40	0.0025	0.1601	0.1534	0.1133	0.1068	0.1686
0.50	0.0006	0.1135	0.1232	0.0829	0.0851	0.1383
0.60	0.0002	0.0842	0.0933	0.0601	0.0523	0.1030
0.70	0.0000	0.0662	0.0653	0.0375	0.0355	0.0736
0.80	0.0000	0.0479	0.0473	0.0288	0.0243	0.0558
0.90	0.0000	0.0203	0.0241	0.0162	0.0031	0.0220
1.00	0.0000	0.0111	0.0178	0.0149	0.0027	0.0170

Average precision over all relevant docs						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Non-interpolated	0.0028	0.1369	0.1447	0.1103	0.1115	0.1571

Document Level Averages						
Document Level	Precision					
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
At 5 docs	0.0075	0.2566	0.2792	0.2377	0.2377	0.2943
At 10 docs	0.0075	0.2264	0.2340	0.1774	0.2075	0.2528
At 15 docs	0.0050	0.1925	0.2113	0.1748	0.1799	0.2113
At 20 docs	0.0057	0.1642	0.1887	0.1557	0.1660	0.1934
At 30 docs	0.0069	0.1484	0.1597	0.1333	0.1434	0.1635
At 100 docs	0.0072	0.0877	0.0906	0.0670	0.0740	0.0840
At 200 docs	0.0062	0.0607	0.0602	0.0423	0.0475	0.0575
At 500 docs	0.0042	0.0319	0.0316	0.0245	0.0269	0.0312
At 1,000 docs	0.0031	0.0188	0.0185	0.0155	0.0170	0.0198

R-Precision (precision after R docs retrieved (where R is the number of relevant documents))						
	JALAB	JALAB1	JALAB2	JALAB3	JALAB4	JALAB5
Exact	0.0066	0.1587	0.1712	0.1332	0.1544	0.1914

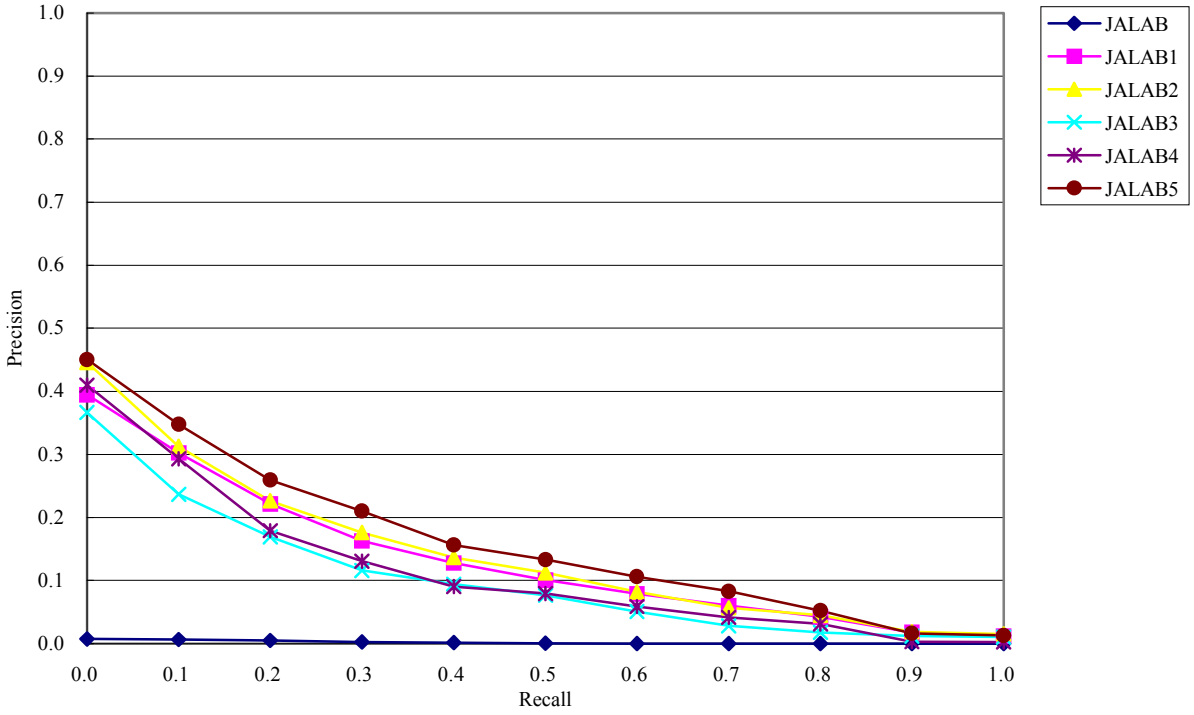

Figure 4 Recall-precision curves for JE-1

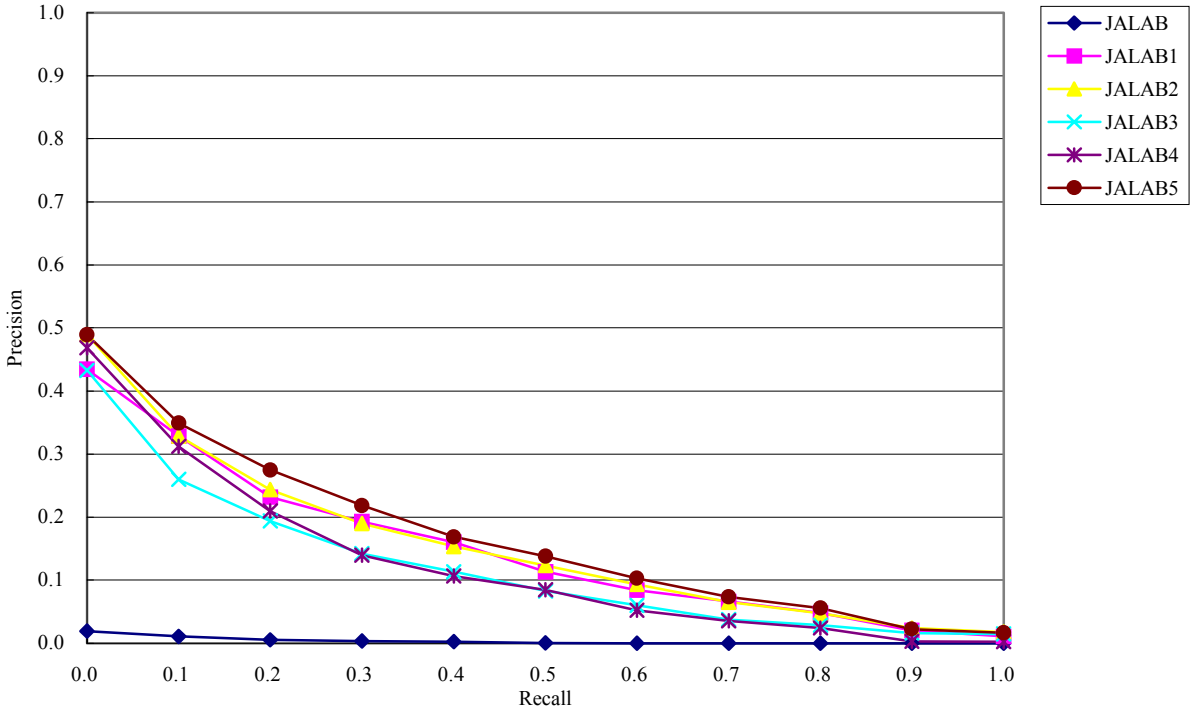

Figure 5 Recall-precision curves for JE-2

REFERENCES

[1] Gupta, A. and Jain, R. Visual Information Retrieval. Communications of the ACM 40, 5, (May 1997), 71-79.

[2] Hull, D.A. and Grefenstette, G. Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (August 1996), 49-57, Zurich, Switzerland.

[3] Kageura, K. et al. NACSIS Corpus Project for IR and Terminological Research. Natural Language Processing Pacific Rim Symposium '97, (December 1997), 493-496, Phuket, Thailand.

[4] Kando, N. et al. NACSIS Test Collection Workshop (NTCIR-1). Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (August 1999), Berkeley, CA, USA.

[5] Kando, N. et al. NTCIR: NACSIS Test Collection Project. Proceedings of the 20th Annual Colloquium of BCS-IRSG, (March 1998), Autrans, France.

[6] Lee, D. L. et al. Document Ranking and the Vector-Space Model. IEEE Software 14, 2, (March 1997), 67-75.

[7] Lee, S. et al. Cross-language multi-media information retrieval system: Boss. Proceedings of the 18th International Conference on Computer Processing of Oriental Languages 1, (March 1999), 493-496, Tokushima, Japan.

[8] Salton, G. and McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill, London, 1983.

[9] Sheridan, P. and Ballerini, J.P. Experiments in multilingual information retrieval using the spider system. Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (August 1996), 58-65, Zurich, Switzerland.