Overview of Patent Retrieval Task at NTCIR-4

Atsushi Fujii*, Makoto Iwayama†, Noriko Kando‡

*Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

†Hitachi, Ltd.
1-280 Higashi-Kougakubo, Kokubunji, 185-8601, Japan
iwayama@crl.hitachi.co.jp

‡ National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430, Japan
kando@nii.ac.jp

Abstract

This paper describes the Patent Retrieval Task in the Fourth NTCIR Workshop, and the test collections produced in this task. We perform the invalidity search task, in which each participant group searches a patent collection for the patents that can invalidate the demand in an existing claim. We also perform the automatic patent map generation task, in which the patents associated with a specific topic are organized in a multi-dimensional matrix.

1 Introduction

In the Third NTCIR Workshop (NTCIR-3), which is a TREC-style evaluation forum for research and development on information retrieval and natural language processing, the authors of this paper organized the Patent Retrieval Task [1, 2]. This was the first serious effort to produce a test collection for evaluating patent retrieval systems.

The process of patent retrieval differs significantly depending on the purpose of retrieval. In NTCIR-3 Workshop, the "technology survey" task was performed, in which patents are regarded as technical publications rather than legal documents. In practice, given a query, which is a clipping of a newspaper articles related to a specific technology, two years of patent publications were searched for the documents relevant to the query. Search topics were in five languages. The same contents in Japanese, English, Korean, traditional/simplified Chinese were used to perform cross-language retrieval.

Given a success in NTCIR-3 Workshop, the authors are also performing the Patent Retrieval Task in NTCIR-4 Workshop, which is held from January 2003 to June 2004. However, unlike NTCIR-3 Workshop, we are focusing on the "invalidity search" and "patent map generation" tasks. This paper describes the test collections for both tasks.

Because NTCIR-4 Workshop is performed in one and half years, it is difficult to explore long-term research topics, such as the patent map generation task. Thus, while we perform the invalidity search task, which resembles the traditional ad-hoc IR task, as the main task, we perform the patent map generation task as a feasibility study, for which no quantitative evaluation is conducted.

2 Invalidity Search Task

2.1 Overview

The purpose of invalidity search is to find the patents that can invalidate the demand in an existing claim. This is an associative patent (patent-to-patent) retrieval task. In real world, invalidity search is usually performed by examiners in a government patent office and searchers of the intellectual property division in private companies.

The task was performed as follows. First, the task organizers (i.e., the authors of this paper) provided each participant group with the document collection

and search topics.

Second, each group submitted retrieval the results queried by the topics. In a single retrieval result, the top 1000 retrieved documents must be sorted by the relevance score. However, because patent documents are long, it is effective to indicate the important passages (i.e., fragments) in a relevant document. Thus, for each retrieved document, all passages in the document must be sorted as to which a passage provides grounds to judge if the document is relevant.

Third, human experts performed relevance judgment for the submitted results and produced a list of relevant documents and passages, on a topic-by-topic basis. Finally, the list was used to evaluate each submitted result.

In the dry run, which was performed from June to September in 2003, seven topics were produced and used for a preliminary evaluation. In the formal run, 103 search topics were produced and the evaluation results for each group will be released at the workshop final meeting in June 2004. The analysis of the formal run results has not been completed and is beyond the scope of this paper.

After the workshop final meeting, we complete a test collection consisting of the search topics, the document collection, and the relevance judgments for each topic.

2.2 Document Sets

The document set used as a target collection consists of five years of unexamined Japanese patent applications published in 1993-1997. The file size and number of documents are approximately 40GB and 1.7M, respectively.

For the sake of passage-based evaluation, the passages in each document were standardized. In Japanese patent applications, paragraphs are identified and annotated with the specific tags by applicants. We used these paragraphs as passages, and therefore the passage identification process was fully automated.

The English patent abstracts, which are human translations of the Japanese Patent Abstracts published in 1993-1997, were also provided for training English-to-Japanese cross-language IR systems.

2.3 Search Topics

A search topic is a Japanese patent application rejected by the Japanese Patent Office. For each topic patent, one or more citations were identified by examiners to invalidate the demand in the topic patent. If these citations are included in our document collection, they can be used as relevant documents for the topic.

We asked 12 members of the Intellectual Property Information Search Committee in the Japan Intellectual Property Association (JIPA) to produce seven topics for the dry run and 34 topics for the formal run. Each JIPA member belongs to the intellectual property division in the company he or she works for, and they are all experts in patent searching. The JIPA member also performed relevance judgment to enhance the relevant documents.

A search topic file includes a number of additional SGML-style tags. The claim as a target of invalidation is specified by <CLAIM>.

A claim usually consists of multiple components (e.g., parts of a machine and substances of a chemical compound) and relevance judgment is performed on a component-by-component basis in real world case. To simulate this scenario, human experts annotate each component with <COMP>.

To invalidate an invention in a topic patent, relevant documents must be the "prior art", which had been open to the public before the topic patent was filed. Thus, the date of filing is specified by <FDATE> and only the patents published before the topic was filed can potentially be relevant.

To perform cross-language retrieval, the claims translated into English and simplified Chinese are also used. Thus, the topic language is specified by <LANG>. However, the translated claims do not maintain the order of phrases and sentences in Japanese claims and thus do not include the <COMP> tags. Figure 1 shows an example topic claim translated into English.

```
<TOPIC>
<NUM>008</NUM>
<CLAIM>(Claim 1) A sensor device, characterized in that an open recessed part is formed on a box-shaped forming base, a conductive film of a designated pattern is formed on the surface of the forming base including the inner surface of the recessed part, an element for a sensor is bonded to the recessed part, and the forming base is closed with a cover.</CLAIM>
</TOPIC>
```

Figure 1. The claim in an English search topic (008).

Through a preliminary study in collaboration with JIPA, we found that for invalidity search the number of relevant documents for a single topic is small, compared with existing IR test collections. Consequently, the evaluation results obtained with our collection can potentially be unstable.

The same problem is identified in the question answering task, and thus the hundreds of questions are usually used to resolve this problem [4].

To increase the number of topics with a limited cost,

we produced additional 69 topics for which only the citations provided by the Japanese Patent Office were used as the relevant documents. However, the validity of rejection was verified manually, the process of producing additional topics was not fully automated.

2.4 Baseline Document Retrieval System

To participate in the main task, each participant group was required to develop an entire retrieval system and perform a number of processes, such as query processing, document indexing, and passage indexing. Amongst of these, document indexing for the 1.7M patent applications was a prohibitive process.

Thus, to facilitate a partial participation, the organizers developed a baseline document retrieval system and provided the participants with an API to use the system on the Web. In other words, a group that developed only query processing and passage retrieval modules was able to participate in the main task. In addition, by sharing the document retrieval system, we can facilitate a glass-box comparative evaluation.

The baseline retrieval system is based on an existing probabilistic method [3] and uses word-based index terms. In addition, non-textual constraints, such as the IPC code and date, can be used to reduce the number of retrieved documents. In the formal run, two groups used the baseline system.

2.5 Submissions

Each group was allowed to submit one or more retrieval results, in which at least one result must be obtained using only the <CLAIM> and <FDATE> fields. For the remaining results, any information in a topic file, such as the International Patent Classification (IPC) codes, can be used.

The results of the dry run showed that for specific topics, an IPC-base system successfully retrieved relevant patents that could not be retrieved by the text-based systems.

2.6 Relevance Judgments

The relevance degree of a document with respect to a topic is determined on the basis of the relevance degrees of the document with respect to each component in the topic. Relevance judgment for patents is performed based on the following two ranks:

- patent that can invalidate a topic claim (A)
- patent that can invalidate a topic claim, when used with other patents (B)

The documents that can invalidate the demands of all essential components in a target claim were judged as "A". The documents that can invalidate demands of most of the essential components in a target claim (but not all essential components) were judged as "B".

For the main 34 topics, to identify relevant documents exhaustively, the pooling method and manual search were used. The human experts who produced the topics performed manual searches to collect as many relevant patents as possible. The experts were allowed to use any systems and resources, so that we were able to obtain a patent document set retrieved under the circumstances of their daily patent searching. The citations provided by the Japanese Patent Office were also used as the relevant documents.

For the 34 topics, the resultant number of A and B documents were 159 and 185, respectively. We analyzed details of the number of relevant documents obtained by the different sources. In Figure 2, "C", "J", and "S" denote the sets the relevant documents (A and B) obtained by the citations, the manual searches by the JIPA members, and the 30 systems participated in the pooling, respectively.

In the formal run, 111 run files including one CLIR result with the English topics were submitted, from which we used 30 files for the purpose of pooling (we selected up to 4 run files for each group, according to the preference).

It should be noted that because the JIPA members collected the citations before the manual search, $|C \cap J|$ is always zero. Looking at this figure, each source was independently effective to collect the relevant documents. While $|C|$ and $|J|$ were almost equivalent, $|S|$ was comparable with $|C \cup J|$.

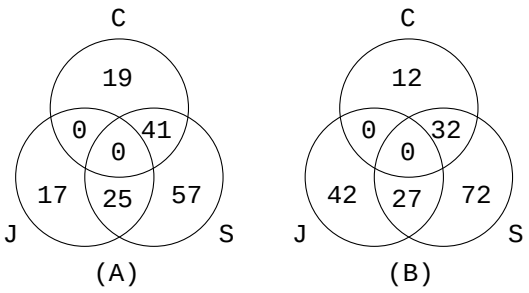

Figure 2. Details of the number of relevant documents.

The evaluation score is fundamentally determined by the conventional mean average precision. At the same time, each group is encouraged to propose new evaluation measures effective for patent IR systems.

In addition to the conventional document-based evaluation, we also explore the passage-based evaluation. Relevant passages were determined based on the following criteria:

- If a single passage can be grounds to judge the

document in question as relevant (either A or B), this passage is judged as relevant.

- If a "group" of passages can be grounds to judge the document in question as relevant, this passage group is judged as relevant.

The experts exhaustively identified all relevant passages and passage groups.

It should be noted that a relevant passage group is equally informative as a single relevant passage. In other words, we newly introduce the concept of "combinational relevance".

This feature provides a salient contrast to the conventional IR evaluation method, in which all relevant passages or documents are independently important and thus combinations of partially relevant documents are not considered.

The evaluation score for each system is determined by the number of passages which would have to be searched until a user obtains a sufficient grounds to judge the document as relevant.

3 Patent Map Generation Task

In principle, the purpose of the patent map generation task is to generate a patent map driven by a specific theme, such as automobiles, by (semi-)automatic method. This can be seen as a text mining task.

In practice, the organizers provided participants with the patent documents retrieved by a specific topic, and participants are requested to organize those documents in a two-dimensional matrix. The x and y axes can vary depending on the topic, but they are usually "problems to be solved" and "solutions", respectively.

To produce the topics and documents, we used the test collection produced for the NTCIR-3 Patent Retrieval Task. We selected six search topics for which more than 100 relevant documents were identified. The NTCIR-3 collection includes the following three document sets:

- two years worth of unexamined Japanese patent applications published in 1998 and 1999,
- Japanese abstracts, the JAPIO Patent Abstracts, which are human-edited abstracts for the above applications,
- English abstracts, the Patent Abstracts of Japan (PAJ), which are human translations of the JAPIO Patent Abstracts.

Any document set can be used for patent map generation purposes. Because the search topics are in the five languages independently (see Section 1), cross-language patent map generation can also be performed.

However, the patent map generation task is as a feasibility study, and thus human experts evaluated the submitted maps subjectively. We had two participant groups for this task.

4 Overview of Formal Run for Main Task

Out of the 69 additional search topics, two topics were discarded (#058 was the same as #21 and the citations for #093 were not included in the document collection). Thus, we used the remaining 101 topics to evaluate the submitted run files.

The number of groups submitted at least one result was eight and the total number of run files was 111, including one CLIR result obtained with the 34 English topics.

We had alternative evaluation measures. Mean average precision (MAP), which has commonly been used in past IR literature, was a feasible choice. However, MAP can potentially be problematic for our purpose, because the evaluation result obtained by MAP can be unreliable if the number of correct answers (relevant documents) is small.

Additionally, for the invalidity search task, users usually investigate more than 100 documents and thus the order of top 100 documents is not always important. However, MAP values can significantly be changed with a small number of top documents (for example, the top 10 documents).

The organizers and participants occasionally had round-table meetings to discuss this problem, but unfortunately there has not been a consensus about the evaluation measure. In addition, the passage-based evaluation has not been conducted mainly due to the delay of data processing.

In summary, we used MAP as the measure for the document-based evaluation purpose. However, the validity of the evaluation result needs to be further investigated.

Table 1 shows the MAP values of the top two results for each group. The column of "Rigid" denotes the case in which the documents judged A were regarded as the correct answers and the column of "Relaxed" denotes the case in which the documents judged B were also regarded as the correct answers.

Although JAPIO achieved the best MAP value for the main topics combined with the Rigid evaluation, RDNDC achieved the best MAP values in most cases. Figures 3-8 depict the recall-precision curves for each case in Table 1.

5 Conclusion

We built test collections for the patent-to-patent invalidity search and automatic patent generation tasks

Table 1. MAP values for different runs.

All topics				Main topics				Add topics			
Rigid		Relaxed		Rigid		Relaxed		Rigid		Relaxed	
RDNDC9	.1693	RDNDC9	.1755	JAPIO10	.2714	RDNDC9	.2666	RDNDC13	.1404	RDNDC13	.1444
RDNDC13	.1636	RDNDC1	.1622	JAPIO14	.2705	RDNDC2	.2465	RDNDC14	.1391	RDNDC14	.1432
JAPIO6	.1630	LAPIN2	.1571	RDNDC2	.2476	JAPIO20	.2465	LAPIN2	.1284	LAPIN2	.1265
JAPIO14	.1597	JAPIO6	.1570	RDNDC9	.2475	JAPIO2	.2441	JAPIO13	.1188	JAPIO13	.1165
LAPIN2	.1570	JAPIO14	.1526	PLLS6	.2408	LAPIN3	.2180	JAPIO15	.1180	JAPIO15	.1159
IFLAB6	.1464	LAPIN3	.1426	fj002-19	.2384	LAPIN2	.2174	IFLAB6	.1082	TRL7	.1071
PLLS6	.1445	IFLAB6	.1343	IFLAB8	.2354	IFLAB11	.1983	TRL7	.1066	IFLAB6	.1057
IFLAB1	.1383	IFLAB14	.1317	fj002-22	.2252	IFLAB12	.1974	LAPIN3	.1054	LAPIN3	.1044
LAPIN3	.1365	PLLS6	.1223	IFLAB6	.2239	fj002-10	.1920	IFLAB14	.1032	IFLAB14	.1015
fj002-13	.1273	fj002-10	.1166	LAPIN2	.2152	fj002-01	.1887	TRL8	.0985	PLLS6	.0988
fj002-04	.1268	fj002-01	.1153	LAPIN3	.1996	PLLS6	.1685	PLLS6	.0971	TRL8	.0975
TRL8	.1024	TRL7	.1107	PLLS1	.1734	PLLS1	.1625	fj002-13	.0838	fj002-13	.0829
TRL7	.0997	TRL8	.1088	TRL8	.1104	TRL8	.1310	fj002-04	.0836	fj002-04	.0827
PLLS1	.0907	PLLS3	.0908	TRL12	.1089	TRL12	.1300	PLLS3	.0557	PLLS3	.0574
NUT1	.0235	NUT1	.0300	NUT1	.0626	NUT1	.0800	NUT1	.0039	NUT1	.0042

in the NTCIR-4 Workshop. After the NTCIR-4 final meeting, the test collection will be available to the public for research purposes[1].

The test collections can directly be used for the following research purposes:

- retrieval of very long semi-structured documents,
- associative document retrieval,
- passage retrieval,
- evaluation of retrieval systems on the basis of combinational relevance,
- classification and text mining.

Future work would include detailed analysis for the formal run results and the passage-based evaluation.

6 Acknowledgments

The authors would like to thank the Japan Intellectual Property Association for their support in the NTCIR-4 Patent Retrieval Task.

References

[1] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: Patents and newspaper articles. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 251–258, 2003.

[2] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at NTCIR-3. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[3] S. 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, 1994.

[4] E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, 2000.

[1] http://www.slis.tsukuba.ac.jp/~fujii/ntcir4/cfp-en.html

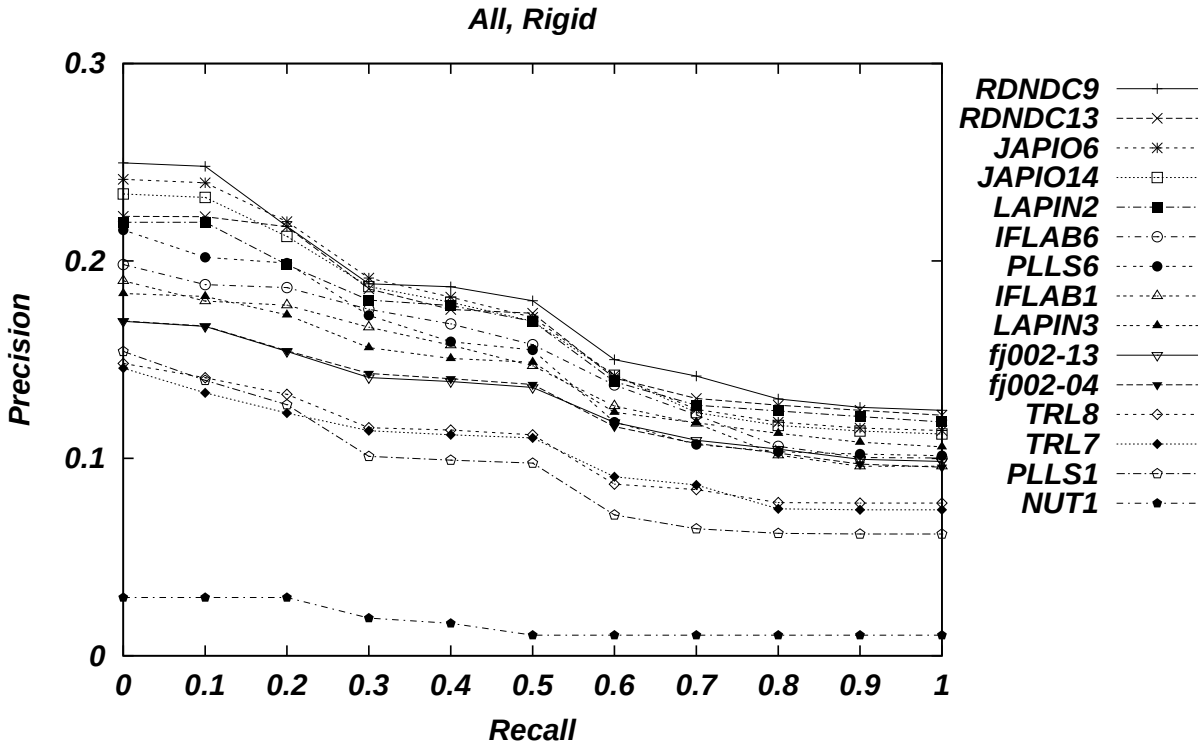

Figure 3. Recall-precision curves for all topics (Rigid).

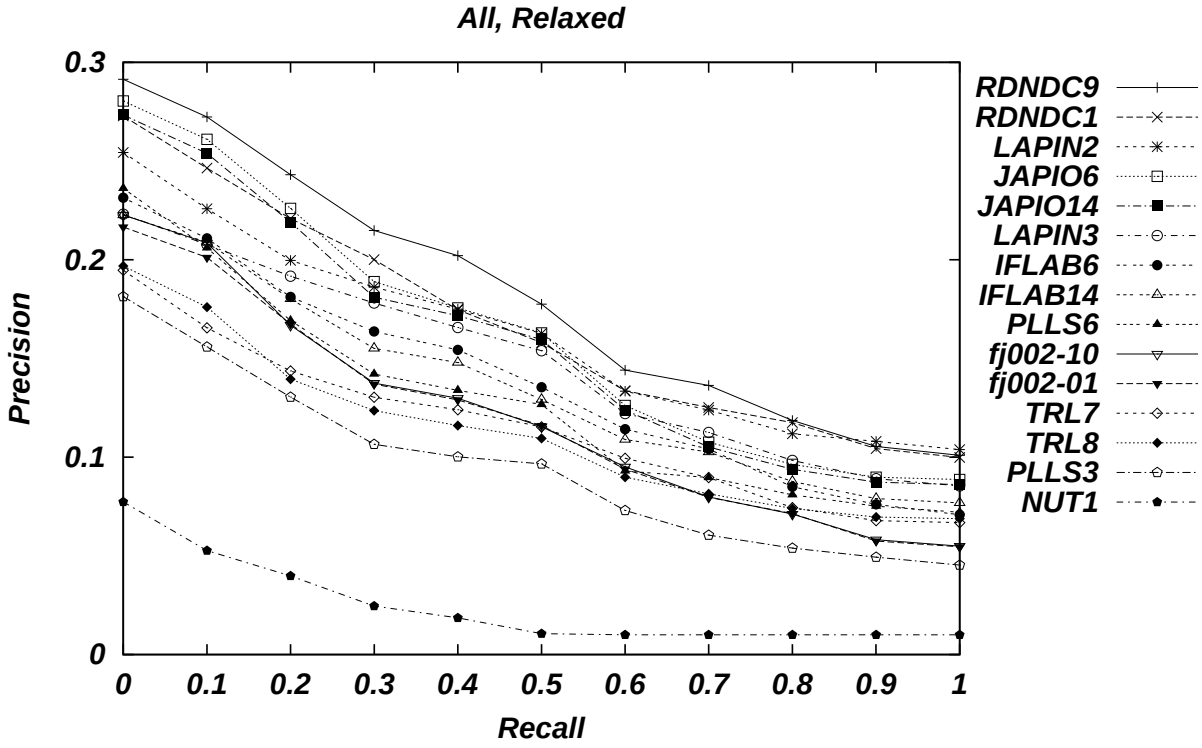

Figure 4. Recall-precision curves for all topics (Relaxed).

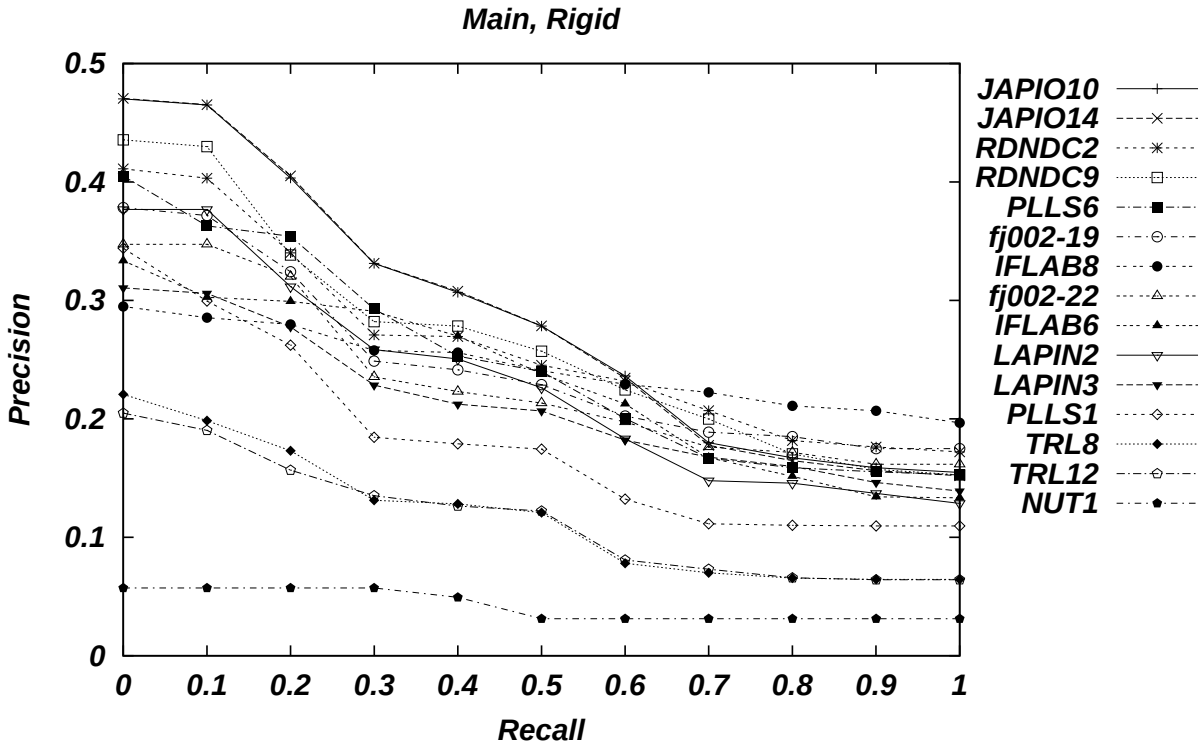

Figure 5. Recall-precision curves for main topics (Rigid).

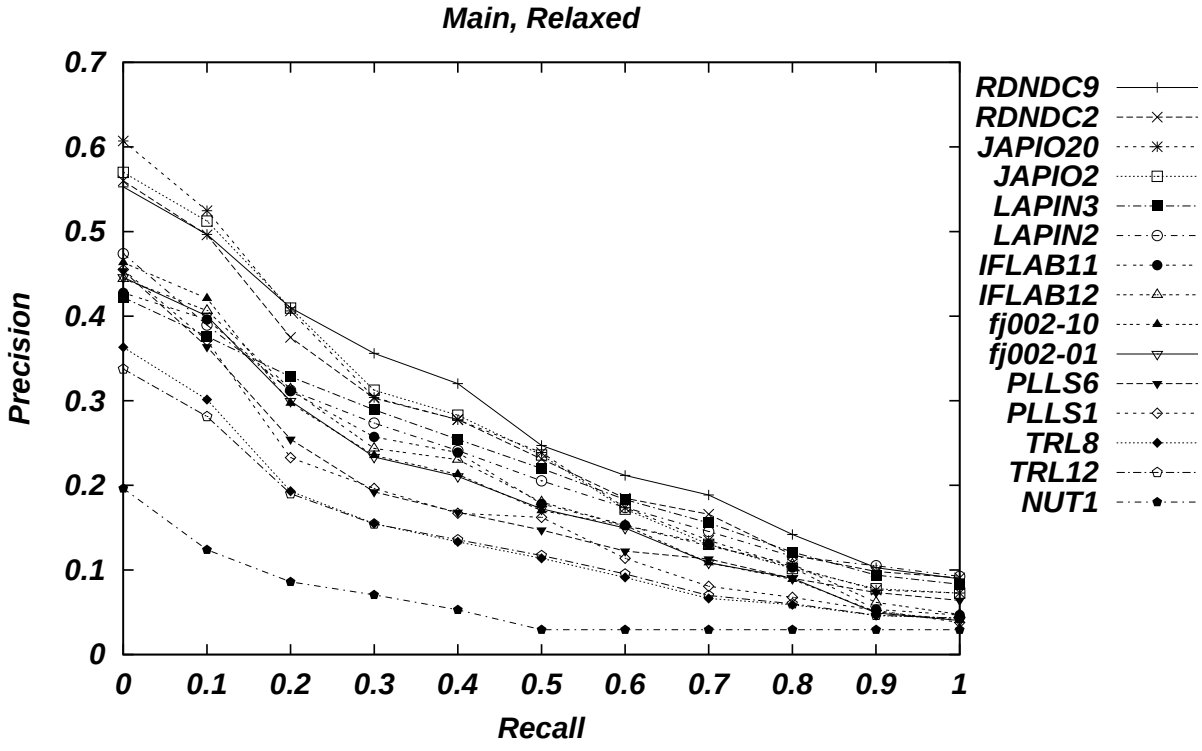

Figure 6. Recall-precision curves for main topics (Relaxed).

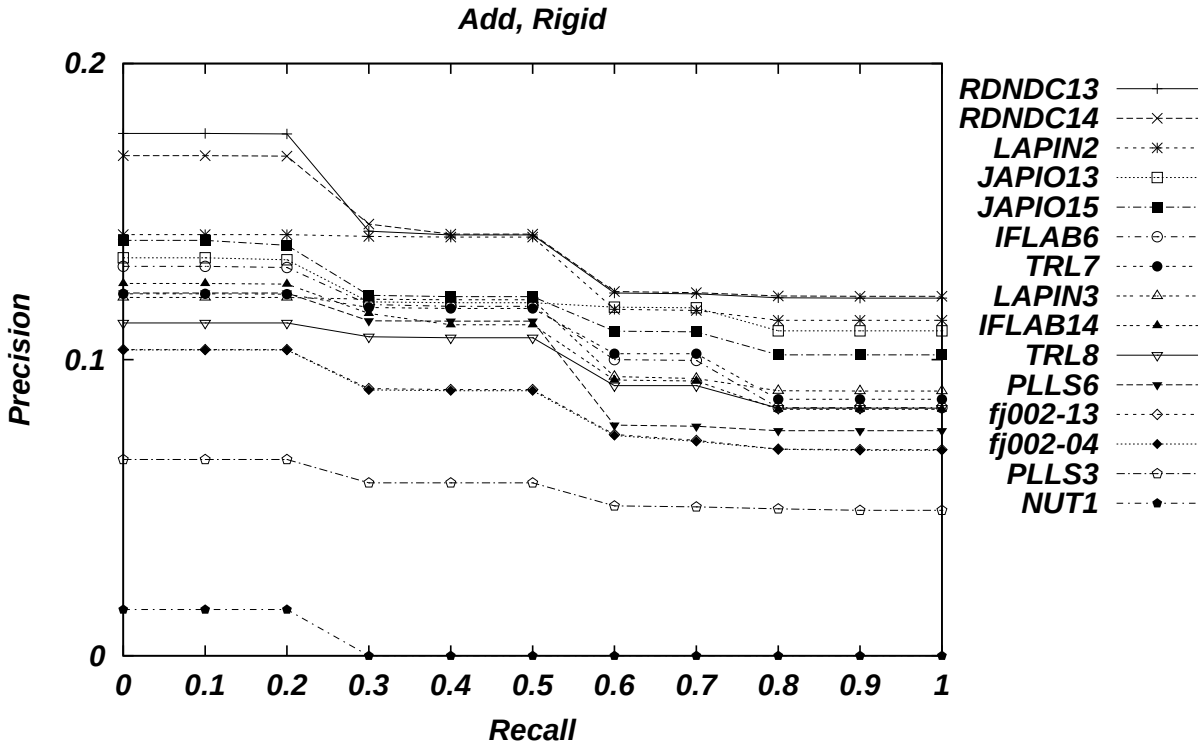

Figure 7. Recall-precision curves for additional topics (Rigid).

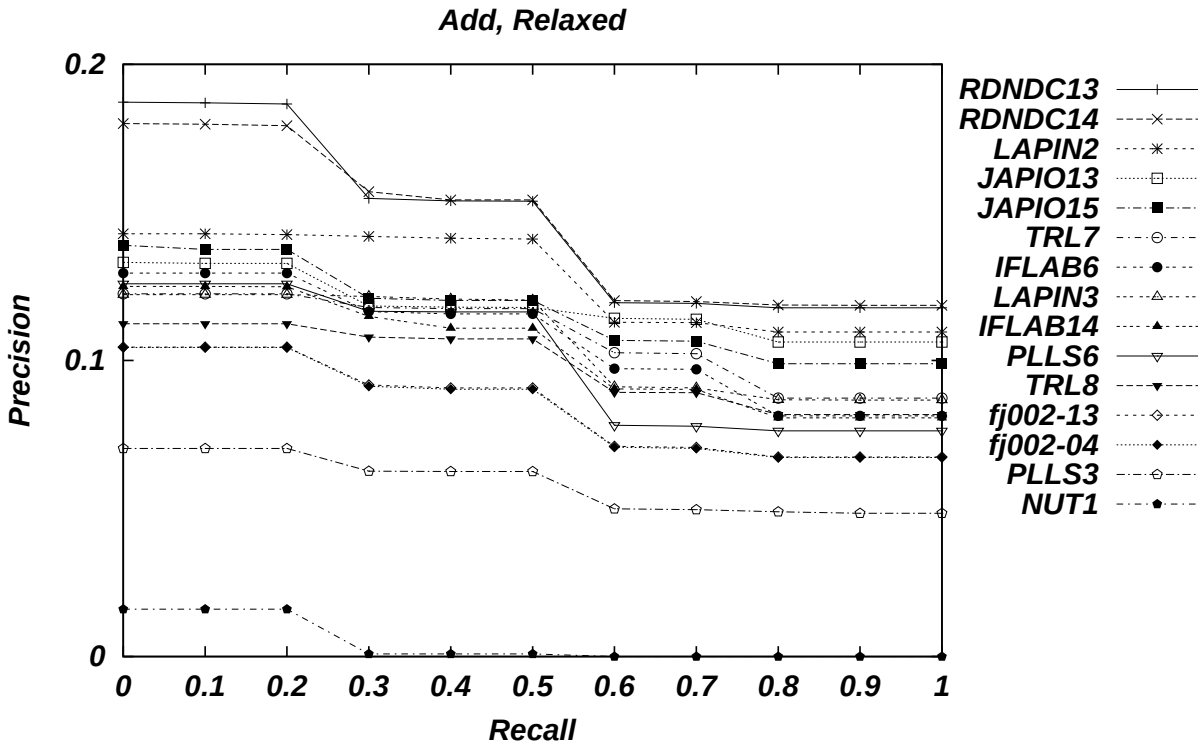

Figure 8. Recall-precision curves for additional topics (Relaxed).