Overview of Patent Retrieval Task at NTCIR-3

Makoto Iwayama
Tokyo Institute of Technology/Hitachi Ltd.
iwayama@crl.hitachi.co.jp

Atsushi Fujii
University of Tsukuba/Japan Science and Technology Corp.
fujii@slis.tsukuba.ac.jp

Noriko Kando
National Institute of Informatics
kando@nii.ac.jp

Akihiko Takano
National Institute of Informatics
aki@acm.org

Abstract

We describe the overview of patent retrieval task at NTCIR-3. The main task was the technical survey task, where participants tried to retrieve relevant patents to news articles. In this paper, we introduce the task design, the patent collections, the characteristics of the submitted systems, and the results overview. We also arranged the free-styled task, where participants could try anything they want as far as the patent collections were used. We describe the brief summaries of the proposals submitted to the free-styled task.

1 Introduction

In the field of information retrieval, there have been held successive evaluation workshops, such as TREC [8], CREF [1], and NTCIR [5], to build and utilize various kinds of test collections. In the Third NTCIR Workshop (NTCIR-3), which was held from June 2001 to December 2003, a serious effort was first made in the "Patent Retrieval Task" to explore information retrieval targeting patent documents.

The goal of Patent Retrieval Task is to provide test collections for enhancing research on patent information processing, from patent retrieval to patent mining. Although there exist many commercial patent retrieval systems and services, patent retrieval has not been paid much attention in the research field of information retrieval. One of the reasons is the lack of test collection on patent. TREC used patent documents as a part of the document collections, but there was no treatment specially applied to the patent collection.

In SIGIR2000, the first workshop on patent retrieval was held [4] and there were many fruitful discussions on the current status and future directions of patent retrieval. The workshop convinced us that there was the need of test collections specifically for patents.

We then asked for PATOLIS Co. [7] to provide patent collections for the patent retrieval task. Consequently, we could release three kinds of patent collections; those were two years' Japanese full texts, five years' Japanese abstracts, and five years' English abstracts. At the same time, we could fortunately have cooperation with JIPA (Japan Intellectual Property Association) [3] in creating search topics and assessing the relevance. Since each member of JIPA belongs to the intellectual property division in her/his company, they are all experts in patent searching. All the above contributions enabled us to kick off the first evaluation workshop designed for patent information processing.

There are various phases and aspects in patent information processing. For example, various kinds of users (researchers, patent searchers, business managers, and so on) search patents for various purposes (technical survey, finding conflicting applications, buying/selling patents, and so on). Corresponding to each situation, an appropriate search model should be developed. The standard of the relevance judgments may also depend on each situation. In some cases, retrieving relevant patents is not enough but further analysis on the retrieved patents might be necessary. For example, creating a patent map of a product would clarify the patent relations between the techniques used to make the product. Cross-lingual patent retrieval is also important when applying patents to foreign countries. All of these are within scope of our project and this task was the first step toward our goal.

2 Task Design

In this workshop, we focused on a simple task of technical survey. End-users we assumed in the task were novice users, for example, business managers. The major reason of adopting such general task was that we could only use the two years' full texts that were not enough for trying more patent-oriented task like finding conflicting applications from patents.

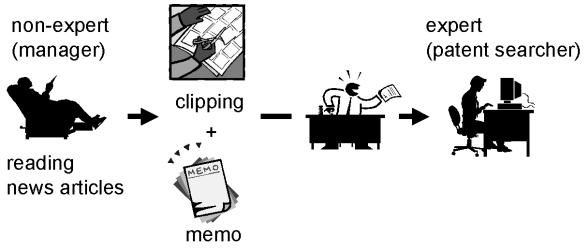


Figure 1: Scenario of technology survey

To fit the task to a real situation, we used Japanese news articles as the original sources of search topics, so the task was conducting cross-database retrieval, searching patents by news articles. The task assumed the following situation that is depicted in Figure 1. When a business manager looks through news articles and is interested in one of them, she/he clips it out and asks a searcher to find related patents to the clipping. The manager passes the clipping to the searcher along with her/his memorandum, and this clipping with memorandum became the search topic in this task. The memorandum helps the searcher to have the exact information need the manager has, when the clipping contains non-relevant topics or the clipping has little description on the information need. Task participants played the role of the searcher and tried to retrieve relevant patents to the clipping. Since the purpose of the searching was technical survey, the claim part in patent was not treated specifically in assessing the relevance. Patent documents were treated as if those were technical papers.

Cross-database retrieval itself is so general that techniques investigated in the task can be applied to various combinations of databases. This is another purpose of the task.

We prepared search topics in four languages, Japanese, English, Korean, and Chinese (both traditional and simplified). Participants could try cross-lingual patent retrieval by using one of the non-Japanese topics. Unfortunately, only two groups submitted cross-lingual results and both of them used English topics.

In addition to the technical survey task explained so far, we arranged the optional task, where participants could try anything they want as far as they used the patent collections provided. One of the purposes of this free-styled task is to explore next official tasks.

3 Characteristics of Patent Applications

In this section, we briefly review the characteristics of patent applications (patent documents).

- There are structures, for example, claims, purposes, effects, and embodiments of the invention.
- Although the claim part is the most important in patent, it is written in an unusual style especially for Japanese patent; all the sub-topics are written in single sentence.
- To enlarge the scope of invention, vague or general terms are often used in claims.
- Patents include much technical terminology. Applicants may define and use their original terms not used in other patents.
- There are large variations in length. The longest patent in our collections contains about 30,000 Japanese words!
- The search models would be significantly different between industries, for example, between chemical / pharmaceutical industries and computers / machinery / electric industries.
- Classification exists. IPC (International Patent Classification) is the most popular one.
- The criterion of evaluation depends on the purpose of searching. For example, high recall is required for finding conflicting applications.
- In some industries, images are important to judge the relevance.

Our task focused on few of the above characteristics. We treated patent documents as technical documents rather than legal statements, so we did not distinguish between the claim part and the oth-

ers in assessing the relevance. High recall was not necessary, so we used the standard averaged precision to evaluate the results. Few groups used structures and classifications. Images were not included in the patent collections provided.

4 Patent Collections

PATOLIS Co. provided and we released the following patent collections.

- **kkh**: Publication of unexamined patent applications (1998, 1999) (in Japanese)
- **jsh**: JAPIO Patent Abstracts (1995–1999) (in Japanese)
- **paj**: Patent Abstracts Japan (1995– 1999) (in English)

"Kkh" contains full texts of unexamined patent applications in Japanese. Images were eliminated. "Jsh" contains human edited abstracts in Japanese. Although all the texts in "kkh" have the abstracts written by the applicants, experts in JAPIO (Japan Patent Information Organization) [2] shortened/lengthened about half of them to fit the length within about 400 Japanese characters. They also normalized technical terms if necessary. "Paj" is English translation of "jsh".

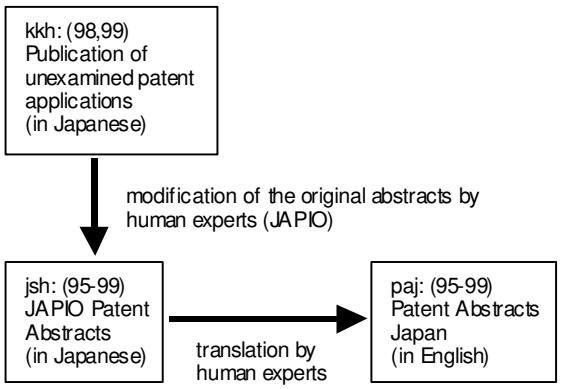


Figure 2: Relationships between the patent collections

Figure 2 shows the relationships between these three collections. Here, we see parallel relations, for example, full texts vs. abstracts, original abstracts vs. edited abstracts, and Japanese abstracts vs. English abstracts. Researchers can use these parallel collections for various purposes, for example, finding rules of abstracting, creating a term normalization dictionary, acquiring translation knowledge, and so on.

Table 1 summarizes the characteristics of the three collections.

	kkh	jsh	paj
Type	Full text	Abstract	Abstract
Language	Japanese	Japanese	English
Years	98,99	95-99	95-99
Number of documents	697,262	1,706,154	1,701,339
Bytes	18139M	1883M	2711M

Table 1: Characteristics of the patent collections

5 Topics

JIPA members created topics, six for the dry run and 25 for the formal run. Since the topics for the dry run were substantially revised after the dry run, we decided to re-use those in the formal run. In consequence, we had the total 31 topics for the formal run.

Figure 3 is an example of the topics in English and Table 2 shows the explanations of the fields in the topics. In our task, <ARTICLE> and <SUPPLEMENT> correspond to the news clipping and the memorandum respectively.

The topics also contain <DESCRIPTION> and <NARRATIVE> fields we are familiar with. Since many NTCIR tasks already have the results for using <DESCRIPTION> and <NARRATIVE> fields, we can compare our results of using these fields with the results of other tasks.

Along with the grade of relevance (i.e., "A", "B", "C", or "D"), each judged patent has a mark ("S", "J", or "U") representing the origin from which the patent was retrieved. Table 3 explains about the marks. For example, a document with "BJ" means that the document was judged as "partially relevant" (i.e. "B-") and only found by experts in their preliminary search (i.e., "-J").

Here, note that all the submitted runs contributed to collecting the "S" patents, but only the top 30 patents for each run were used. Note also that we can restore the patent set retrieved by the manual search (i.e., "PJ" set) by collecting "J" and "U" patents.

```
<TOPIC><NUM>P004</NUM><LANG>ENK</LANG>
<PURPOSE>technology survey</PURPOSE>
<TITLE>Device to judge relative merits by comparing
codes such as barcodes with each other</TITLE>
<ARTICLE>
<A-DOC>
<A-DOCNO>JA-981031179</A-DOCNO>
<A-LANG>JA</A-LANG>
<A-SECTION>Society</A-SECTION>
<A-AE>No</A-AE>
<A-WORDS>189</A-WORDS>
<A-HEADLINE>BANDAI lost a lawsuit for piracy filed by
EPOCH at Tokyo District Court</A-HEADLINE>
<A-DATE>1998-10-31</A-DATE>
<A-TEXT>In settlement of the lawsuit filed by EPOCH
INC., the toy manufacturer, against BANDAI CO., LTD. As
compensation of 264 million for damages for infringement
of a card game patent, the Tokyo District Court ordered
BANDAI to pay about 114 million on the 30th. The presid-
ing judge, Mr. Yoshiyuki Mori, indicated that some func-
tions including key operation for the "Super Barcode
Wars" mini game machine manufactured and sold by BANDAI
CO., LTD. in July, 1992 to March, 1993 fell under the
"technical range of a patent licensed to EPOCH
INC.".</A-TEXT>
</A-DOC>
</ARTICLE>
<SUPPLEMENT>Determination of victory or defeat by com-
paring each other's values based on codes from barcode
readings does not conflict with the patent.</SUPPLEMENT>
<DESCRIPTION>What kind of devices determines leaders or
victors by reading several codes such as barcodes and
comparing the values corresponding to these
codes?</DESCRIPTION>
<NARRATIVE>"Super Barcode Wars" is a type of mini game
machine where recorded barcodes are read in cards fea-
turing characters and the game proceeds in semi-real
time by operating offence and defense keys. Sample codes
include barcodes and magnetic codes, but shall not be
defined as limited only to these.</NARRATIVE>
<CONCEPT>Sign, barcode, code, superiority or inferior-
ity, victory or defeat, comparison, judgment</CONCEPT>
<PI>PATENT-KKH-G-H01-333373</PI>
</TOPIC>
```

Figure 3: Example of the topics

Field	Explanation
<LANG>	Language code
<PURPOSE>	Purpose of search
<TITLE>	Concise representation of search topic
<ARTICLE>	MAINICHI news article in NTCIR format
<SUPPLEMENT>	Supplemental information of news article
<DESCRIPTION>	Short description of search topic
<NARRATIVE>	Long description of search topic
<CONCEPT>	List of keywords
<PI>	Original patents of news article

Table 2: Explanations of the fields in topics

6 Results Overview

6.1 Participants

Eight groups submitted the 36 runs. One group submitted runs only for pooling. We briefly describe the characteristics of each group. Refer to the proceedings of Patent Retrieval Task [6] for each detail.

LAPIN: This group focused on the "term distillation" in cross-database retrieval, where the difference between the term frequency in source database and that in target database was integrated into the overall term weighting.

SRGDU: This group tried several pseudo relevance feedback methods in the context of patent retrieval. The proposed method using Taylor formula was compared with the traditional Rocchio method.

daikyo: This group made long gram-based index from the patent collections. Compared with the traditional gram-based indexing, proposed method produce more compact index.

DTEC: This group searched various kinds of abstracts rather than full texts, and compared the effectiveness of those. The abstracts were JAPIO patent abstracts and the combinations of "title", "applicant's abstract", and "claims". Manual and automatic runs were compared.

DOVE: This group also submitted manual and automatic runs. In the manual runs, non-relevant passages in <ARTICLE> were eliminated manually.

IFLAB: This group evaluated their cross-lingual IR system PRIME through several monolingual runs. They also evaluated their translation extraction method by using Japanese-US patent families, which were not provided in this task.

brkly: This group submitted both monolingual and cross-lingual runs. In the cross-lingual runs, words in English topics were translated into Japanese words by using English-Japanese dictionary automatically created by the aligned bilingual corpus (i.e., "paj" and "jsh"). Their method of creating the dictionary is based on word co-occurrence with the association measure.

sics: This group also submitted cross-lingual runs, where they automatically created a cross-lingual thesaurus form the aligned bilingual corpus, "paj" and "jsh", and used the thesaurus for word-based query translation. The Random Indexing

vector-space technique was used to extract the cross-lingual thesaurus. Note that, in both the "sics" and the "brkly" groups, there was no member who understands Japanese.

6.2 Recall/Precision

The recall/precision graphs of the mandatory runs are shown in Figure 4, and those of the optional runs in Figure 5. In each figure, there are both results for the strict relevance ("A") and the relaxed relevance ("A" + "B"). For each run in the figures, brief system description is specified; the description includes the searching mode (automatic or manual), the topic fields used in query construction, and the topic language.

6.3 Topic-by-topic Results

Figure 6 shows the median of the average precisions for each topic. Figure 7 shows the breakdown of the relevance judgments. Detailed analysis on each topic will be given by JIPA, where it will be discussed about the reasons why systems could not find some patents human experts found and vise versa.

6.4 Recall of the relevant patents retrieved in the preliminary human search

Figure 8 shows the recall of the relevant patents retrieved in the preliminary human search. In the process of making pool, we used only the top 30 documents for each run. Here, we extracted more documents from each run and investigated how many human retrieving relevant patents could be covered by the systems.

7 Optional (Free-styled) Task

The following two groups applied to the optional task. Refer to the proceedings of Patent Retrieval Task [6] for each detail.

CRL: This group investigated the method of extracting various rules from the existing alignments in patents. The "diff" command of UNIX was used to find the alignments between JAPIO patent abstracts and the original abstracts by applicants, between claims and embodiments, and between different claims in an application.

TIT: This group focused on the unusual style of Japanese claims, and tried to automatically structure the claims to raise the readability of claims. Rhetorical structure analysis was applied for this purpose.

8 Summary and Future Directions

In this paper, we described the overview of patent retrieval task at NTCIR-3. We are planning to continue our effort for the next patent retrieval task along with the following directions.

- Longer range of years will be covered.
- Purpose of search would shift to more real one, for example, searching conflicting applications.

Acknowledgements

We are grateful to PATOLIS Co. for providing the patent collections of this task. We also thank all the members of JIPA who created the topics and assessed the relevance. Without their expertise in patent, this task would not be realized. Lastly, we thank all the participants for their contributions to this task.

References

[1] CLEF (Cross Language Evaluation Forum) (http://clef.iei.pi.cnr.it/)

[2] JAPIO (Japan Patent Information Organization) (http://www.japio.or.jp/)

[3] JIPA (Japan Intellectual Property Association) (http://www.jipa.or.jp/)

[4] ACM-SIGIR Workshop on Patent Retrieval, organized by Mun-Kew Leong and Noriko Kando, 2000. (http://research.nii.ac.jp/ntcir/sigir2000ws/)

[5] NTCIR (NII-NACSIS Test Collection for IR Systems) (http://research.nii.ac.jp/ntcir/index-en.html)

[6] Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, 2003.

[7] PATOLIS Co. (http://www.patolis.co.jp/e-index.html)

[8] TREC (Text Retrieval Conference) (http://trec.nist.gov/)

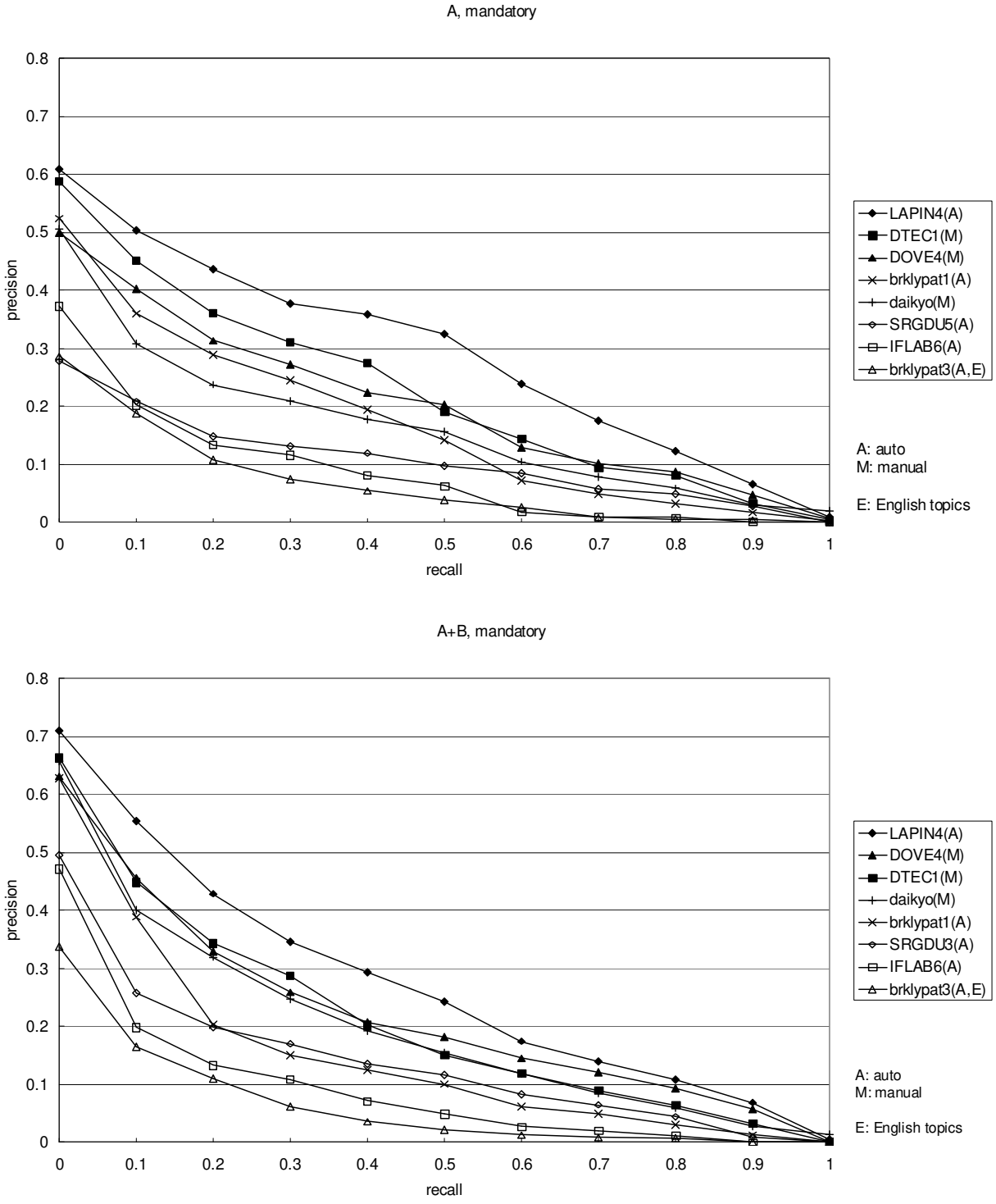
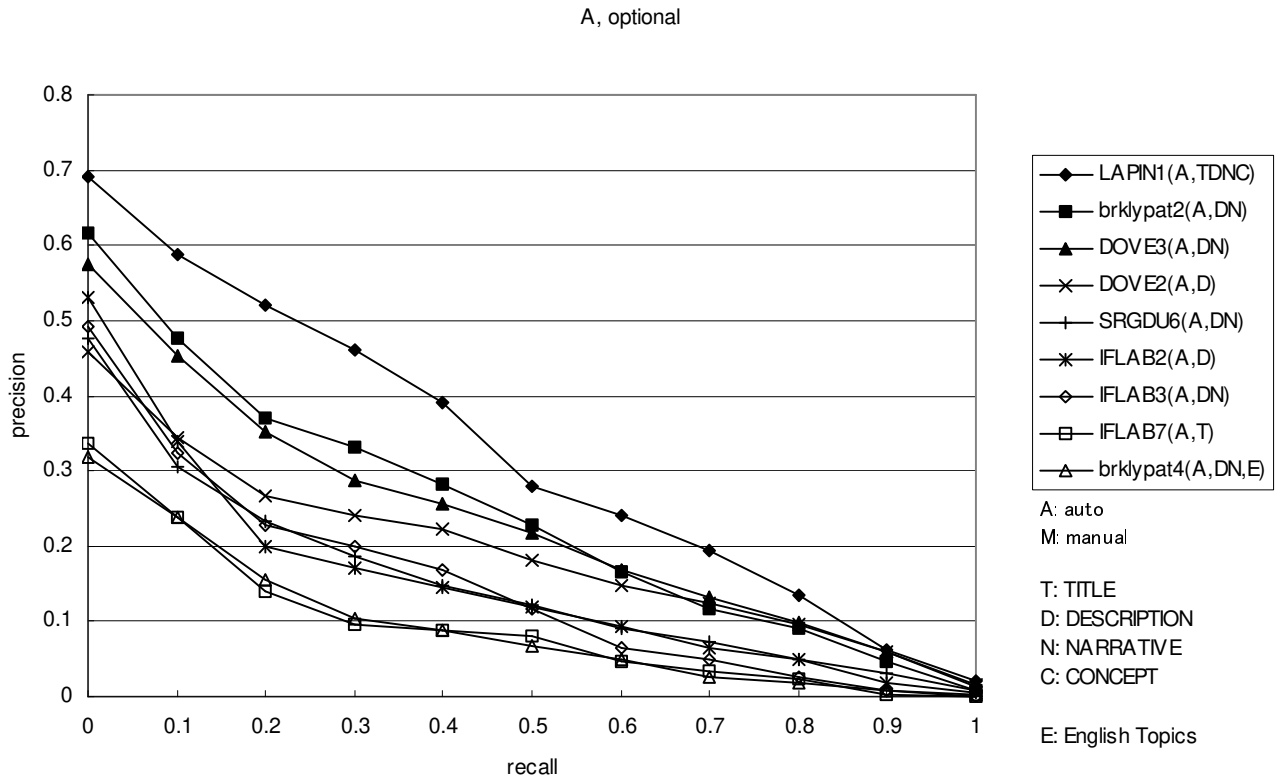


Figure 4: Recall/Precision of mandatory runs



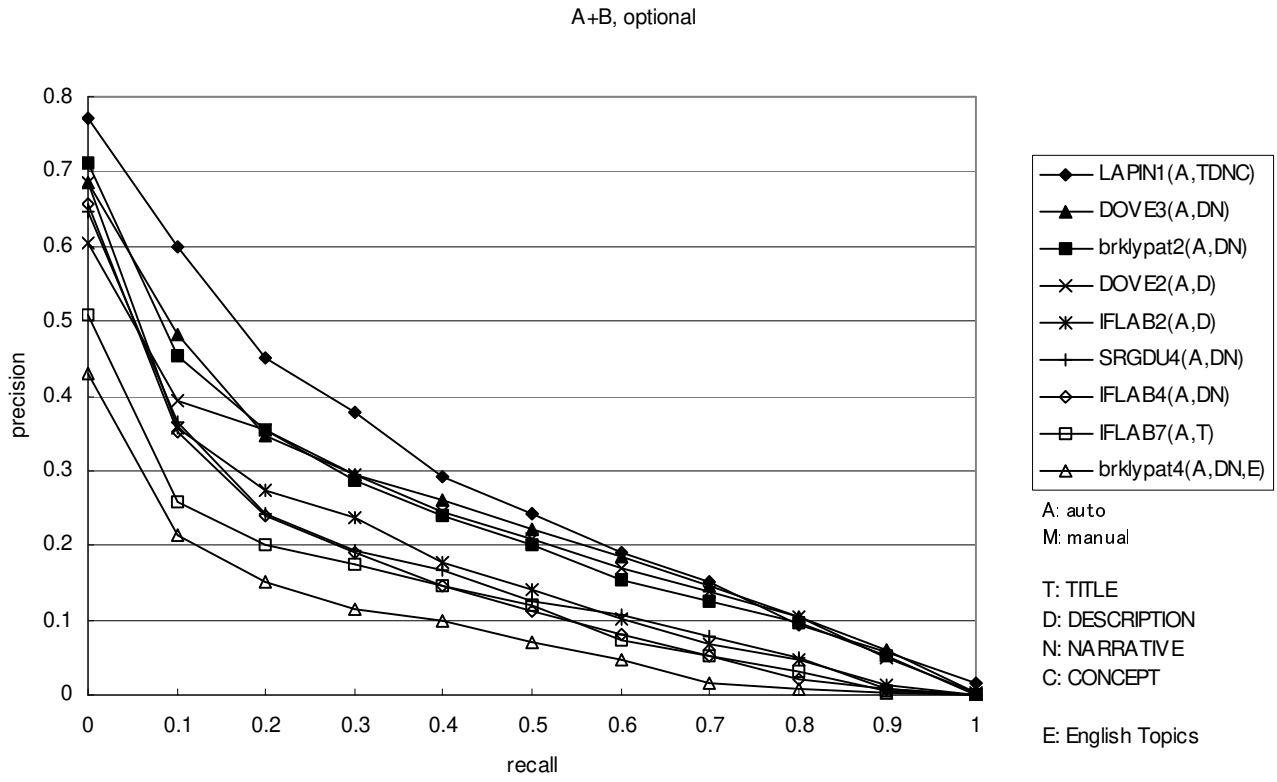


Figure 5: Recall/Precision of optional runs

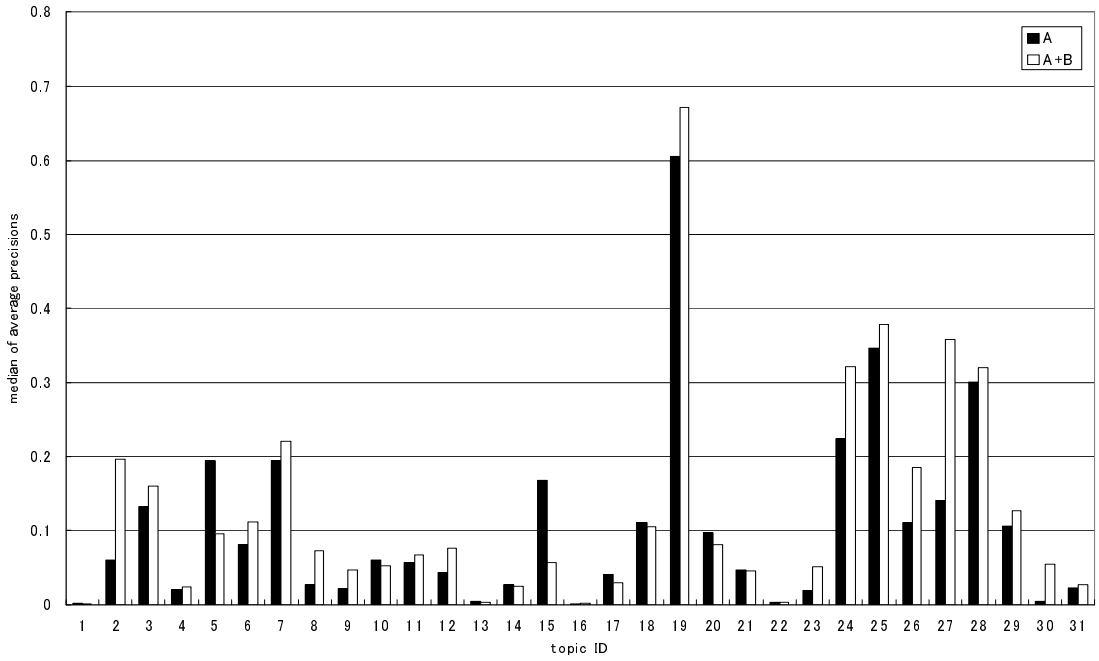
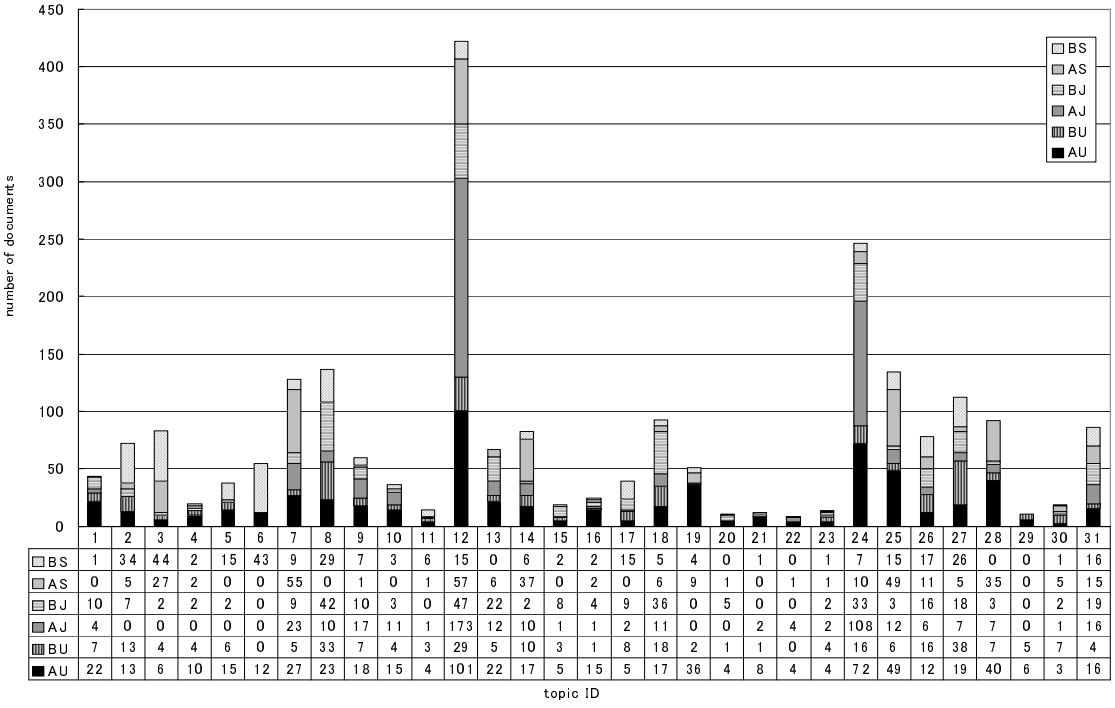
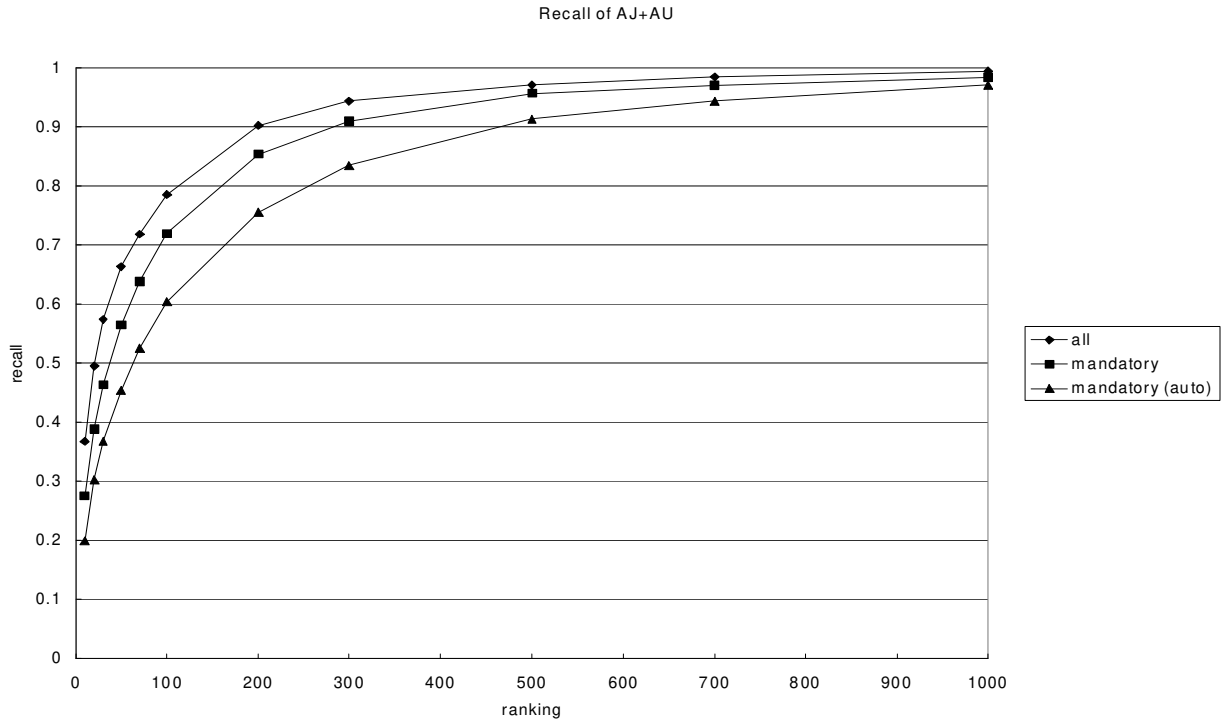


Figure 6: Median of average precisions (all runs)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
BS	1	34	44	2	15	43	9	29	7	3	6	15	0	6	2	2	15	5	4	0	1	0	1	7	15	17	26	0	0	1	16
AS	0	5	27	2	0	0	55	0	1	0	1	57	6	37	0	2	0	6	9	1	0	1	1	10	49	11	5	35	0	5	15
BJ	10	7	2	2	2	0	9	42	10	3	0	47	22	2	8	4	9	36	0	5	0	0	2	33	3	16	18	3	0	2	19
AJ	4	0	0	0	0	0	23	10	17	11	1	173	12	10	1	1	2	11	0	0	2	4	2	108	12	6	7	7	0	1	16
BU	7	13	4	4	6	0	5	33	7	4	3	29	5	10	3	1	8	18	2	1	1	0	4	16	6	16	38	7	5	7	4
AU	22	13	6	10	15	12	27	23	18	15	4	101	22	17	5	15	5	17	36	4	8	4	4	72	49	12	19	40	6	3	16

Figure 7: Breakdown of relevance judgments



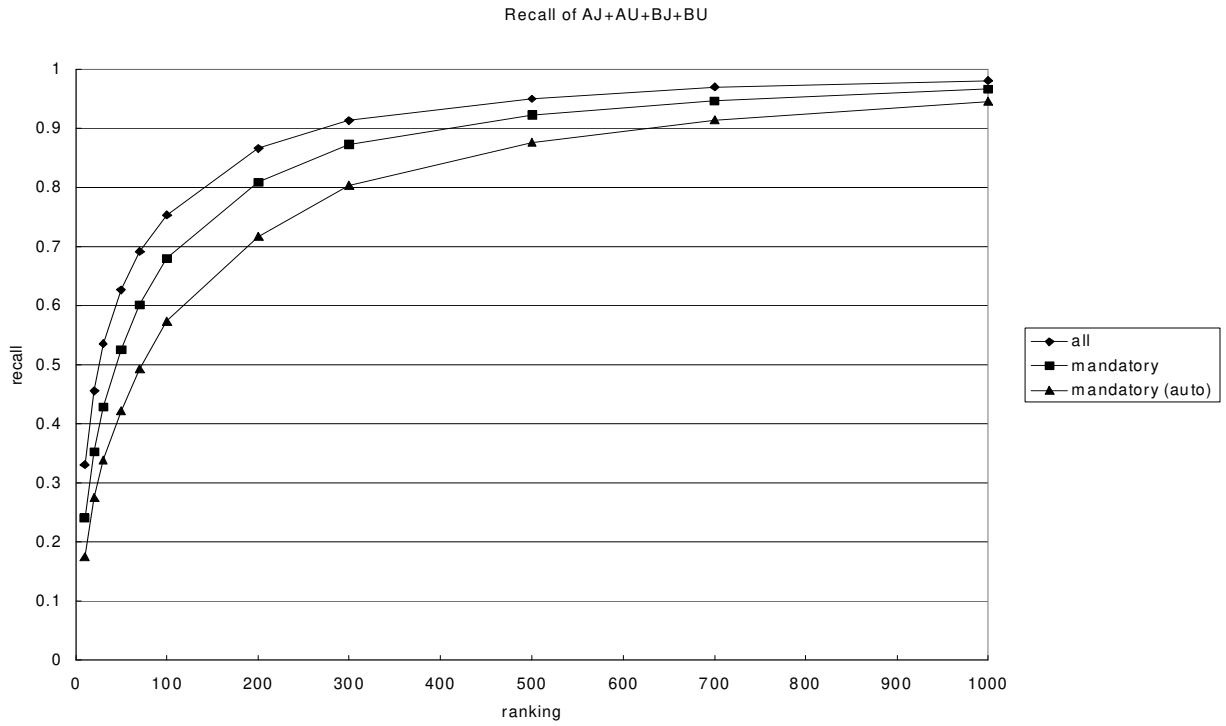


Figure 8: Recall of the relevant patents retrieved in the preliminary human search