Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task

Makoto Iwayama[*], Atsushi Fujii[†], Noriko Kando[‡]
[*] Hitachi, Ltd., 1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan
makoto.iwayama.nw@hitachi.com
/ Tokyo Institute of Technology
[†] University of Tsukuba, 1-2 Kasuga, Tsukuba 305-8550, Japan
fujii@slis.tsukuba.ac.jp
[‡] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8 30, Japan
kando@nii.ac.jp

Abstract

This paper describes the Classification Subtask of the NTCIR-5 Patent Retrieval Task. The purpose of this subtask is to evaluate the methods of classifying patents into multi-dimensional classification structures called F-term (File Forming Term) classification systems. We report on how this subtask was designed, the test collection released, and the results of the evaluation.

Keywords: *Patent categorization, F-term, Patent map*

1. Introduction

Organizations attempting to utilize their patents have to survey the existing patents in the targeting domain and clarify the advantages and disadvantages of their patents as compared with their competitors' patents. Patent maps help this kind of analysis by providing distributional information of patents from various perspectives.

In the NTCIR- Patent Retrieval Task [2], we started a subtask (named the Feasibility Study Subtask) for automatically creating a patent map that offers a bird's eye view of patents in a specific technological field. The patent map we targeted was a two-dimensional matrix that summarizes patents from two viewpoints, namely problems to be solved and solutions. Figure 1 is an example. In the map, columns ("crystalline", "reliability", "long life", etc.) are possible problems to be solved by the patents and rows ("structure of active layer", "electrode composition", etc) are possible solutions claimed in the patents. Patents in each cell solve the corresponding problem with the corresponding solution. For example, the patent 1998-107318 solves the problem of reliability of blue light-emitting diodes with an approach to electrode composition.

		problems to be solved				
		crystalline	reliability	long life	emission stability	emission intensity
solutions	structure of active layer			1998-145000 1998-233554		
	electrode composition		1998-107318		1998-190063 1998-209498	1998-209495
	electrode arrangement		1998-215034 1998-223930	1998-242518	1998-173230 1998-209499 1998-256602	1998-242515 1998-270757
	structure of light emitting element	1998-135516 1998-242586 1998-247761		1998-135514 1998-256668		1998-012923 1998-247745 1998-256597

Figure 1. Patent map of blue light-emitting diodes.

Although this subtask revealed a couple of promising approaches to automatic patent map creation, the initial evaluation was insufficiently thorough. We subjectively evaluated only six topics. In addition, the dimensions of patent maps were fixed to problems to be solved and solutions for every topic. However, appropriate viewpoints in a patent map are different from one technological field to another.

In the last Classification Subtask of the NTCIR-5 Patent Retrieval Task [3], we focused on the evaluation of patent categorization by using multi-dimensional classification structures called the F-term (File Forming Term) classification system[1][5], which is used in the Japan Patent Office. The F-term classification system has over 2,500 themes covering all technological patent fields. Patents in any theme can be classified from several viewpoints, such as purpose, problem, solution, effect, and so on. The set of possible viewpoints varies from theme to theme. Each viewpoint defines a set of possible elements, and a pair consisting of a viewpoint and an element is called an F-term. F-terms are a powerful tool for specifying relevant patents in patent searches. F-terms also help create patent maps such as the one shown in Figure 1 by selecting an appropriate pair of dimensions from the possible viewpoints and by classifying patents based on the selected viewpoints.

Experts assign F-terms to a patent in two steps. They first determine themes of the patent, and then

[1] http://www.ipdl.ncipi.go.jp/HELP/pmgs_en/database/format_summary.html#fterm

5B001		Detection and correction of errors					
AA	AA00	AA01	AA02	AA03	AA04	AA05	…
	CODES	. Parity	.. Multiple parity	. Error-correction codes (ECC)	.. Cyclic-redundancy check (CRC)	.. Single-bit error correction and double-bit error detection (SECDEC)	…
AB	AB00	AB01	AB02	AB03	AB04	AB05	…
	PURPOSE	. Error detection	. Error correction	. Code generation	.. Prediciton	. Decoding	…
AC	AC00	AC01	AC02	AC03	AC04	AC05	…
	MEANS	. Code operations	.. Tables	.. Counting	. Comparison	. Interleaving	…
AD	AD00	AD01	AD02	AD03	AD04	AD05	…
	ERROR LOCATION	. Arithmetic circuits	.. Decoders	. Memories	.. Magnetic tapes	. Interfaces	…
AE	AE00	AE01	AE02	AE03	AE04	AE05	…
	TYPES OF ERRORS	. Program instructions	. Data	.. Multiple errors	...Burst errors	. Addresses	…

Figure 2. Example of an F-term classification system.

for each theme they assign F-terms to the patent. Based on this procedure, we divided the last Classification Subtask (at NTCIR-5) into two parts, the Theme Categorization Subtask and the F-term Categorization Subtask. In the Theme Categorization Subtask, participants determined one or more themes for each patent. This can be seen as a simplified version of classifying patents into the world standard taxonomy of the IPC (International Patent Classification). Refer to [1] for approaches to automatic patent categorization based on the IPC[2]. In the F-term Categorization Subtask, participants determined one or more F-terms for each patent whose theme had been given. The F-term Categorization Subtask was a new attempt in that the targeted categories were multi-dimensional (in other words, multifaceted) categories of F-terms.

In this NTCIR-6, focusing on the F-term Categorization Subtask, we evaluated the multi-dimensional patent categorization extensively. We increased the number of targeted themes from five (at NTCIR-5) to 108. Those themes were randomly selected. There was a total of 21,606 test documents. In this subtask, we did not conduct the Theme Categorization Subtask because this type of conventional text categorization has already been extensively evaluated.

The rest of this paper is organized as follows. Section 2 reviews the F-term classification system. Section 3 describes the task overview, and Section introduces the datasets we released. Section 5 describes the evaluation method, and Section 6 shows the evaluation results. Section 7 concludes the paper.

2. F-term Classification System

The most common patent classification taxonomy is the IPC, which is internationally uniform. The IPC is structured based on the single viewpoint of technological contents of inventions. However, patent searchers often have to explore patents from various viewpoints such as the purpose of the invention, the problem to be solved, the solution, the effect of the invention, and so on. To this end, the Japan Patent Office provides multi-dimensional classification structures called F-term classification systems based on which most Japanese patents are classified. Figure 2 shows an example.

In the F-term classification system, each technological field is defined as a theme corresponding to a set of FI (a Japanese extension of the IPC) codes. For example, the theme denoted by 5B001 is the technological field of detection and correction of errors (in computers) and corresponds to the FI codes of G06F11/08-11/10,330@Z. A theme is expressed by a sequence of a digit, a letter, and three digits. There are over 2,500 themes covering all the technological fields in patents.

Each theme has a set of viewpoints for specifying possible aspects of the inventions under this theme[3]. For example, 5B001 has PURPOSE, MEANS, ERROR LOCATION, and other viewpoints. The set of possible viewpoints varies from theme to theme. In the example, ERROR LOCATION is a unique viewpoint for this theme. A viewpoint is denoted by two letters. For example, AC represents the viewpoint MEAN. Note that the viewpoint naming policy is not uniform across themes, meaning that AC may not represent MEAN in other themes.

Each viewpoint has a list of possible elements. For example, MEANS of 5B001 can be Code operations, Comparison, Interleaving, and so on. The set of possible elements varies from viewpoint to viewpoint. An element is represented as two digits. For example, Interleaving under MEAN corresponds to 05. As an exception, 00 sometimes represents others, i.e., the elements not enumerated in the list. The 00 element may also be used to designate its belonging viewpoint,

[2] Only class-level or subclass-level IPC categories (the numbers are 11 and 51 respectively) are considered in [1].

[3] Some themes do not have viewpoints mainly because their FI codes are sufficient to classify the patents.

as seen in Figure 2.

A pair consisting of a viewpoint and its element is briefly called F-term. For example, AC05 is an F-term representing mean (of error collection and correction) (AC) is interleaving (05). Although F-terms can have an additional letter for expressing more detailed information, we ignored the additional codes in this subtask.

There are general/specific relations between F-terms. This relationship is defined by dot (.) characters written in the description of each F-term. Figure 3 shows examples of such descriptions.

3E003 (Container packaging and wrapping operations)

AA00 CONTAINERS
AA01 . Rigid containers
AA02 .. Rigid containers with integrated internal dividers
AA03 .. Rigid containers with separate internal dividers
AA04 .. Rigid containers with cushioning materials
AA05 . Soft containers

Figure 3. Examples of F-term descriptions.

The number of dots signifies the level of the hierarchy. Absence of a dot signifies the highest level, which is followed by single dot (.), double dots (..), and triple dots (…) in descending hierarchical order. The F-terms in Figure 3 correspond to the hierarchy in Figure .

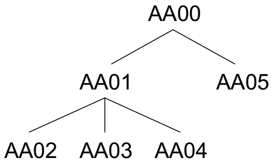

Figure 4. Hierarchy of F-terms.

3. Task Overview

Figure 5 is an overview of the Classification Subtask at this NTCIR-6. In this subtask, participants had to submit a ranked list of 200 possible F-terms for each test document (patent) whose theme had been given at the release of the test document. In a submitted list of F-terms, higher ranked ones are more likely to be assigned to the test document than lower ranked ones. For each submitted F-term in the list, participants also had to decide whether the F-term is confidently assigned to the patent or not. Only the confident F-terms were used to calculate the F-measure. Submitted lists were evaluated based on recall/precision. The evaluation will be described in detail in Section 5.

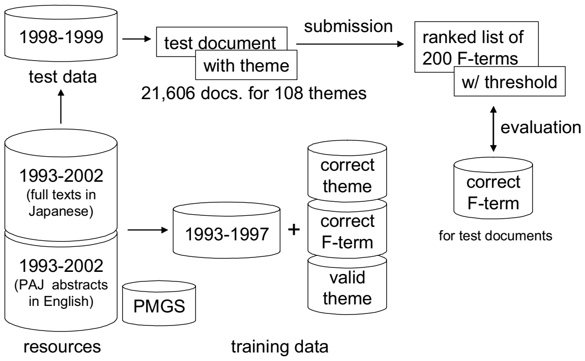

Figure 5. Overview of Classification Subtask.

4. Datasets

4.1 Document Resources

Unexamined Japanese patent applications published from 1993 to 2002 were released for this subtask. These applications are full texts of Japanese patents (written in Japanese). The same years' English abstracts were also released. That is, every full text in Japanese has a corresponding abstract in English. This collection of English abstracts is called PAJ (Patent Abstract Japan).

The PMGS (Patent Map Guidance System) contains descriptions of the themes and the F-terms. The PMGS is provided in both Japanese and English.

4.2 Training data

We released the list of correct themes and correct F-terms for every patent published from 1993 to 1997 as the training data. Those themes and F-terms were extracted from the Seirihyoujunka (Standardized) Data which contains the latest bibliographic information of patents. The Seirihyoujukna Data are the dumped copies of the master databases in the Japan Patent Office. Note that the themes and F-terms in the full text of a patent are not the latest ones. There may be revisions of the assigned themes and F-terms after publication, which would appear only in the master databases.

Participants could use the PMGS for training purposes.

4.3 Test data

To make the test data, we first excluded invalid themes, i.e., discontinued themes, themes under revision, and themes called partial F-term themes where F-terms are defined for only a portion of the theme. After the filtering, we had a list of 1,200 valid themes, which we released to participants. In the dry run, we found that some valid themes had F-terms not listed in the PMGS. In the formal run, we excluded these themes, obtaining 1,119 valid themes.

Next, we randomly selected 108 themes from the

http://www5.ipdl.ncipi.go.jp/pmgs1/pmgs1/pmgs_E.

1,119 valid themes, and, for each selected theme, we randomly sampled about 200 test documents from 1998 and 1999. Table 1 lists statistics for the test documents. From the table, it can be seen that the number of correct F-terms per document is largely between one and 13 .

Table 1. Statistics of test documents.

number of themes	108
number of documents	21606
average number of correct F−terms per doc.	9.32
maximum number of correct F−terms per doc.	134
minimum number of correct F−terms per doc.	1

5. Evaluation Method

We did document-driven evaluation in which a contingency table is made for each document rather than each category. Here the conventional evaluation compares the correct categories of a document with submitted categories posted by a system. For example, if a document has a set of categories $\{b, f\}$, and a system submits a set of categories $\{c, f\}$ for the document, the recall of the correct categories becomes 1/2, and the precision of the submitted categories becomes 1/2.

In this subtask, we evaluated the effectiveness of each categorization method through text retrieval where documents for searching are assigned to categories by the method in evaluation, and queries that can retrieve the documents are evaluated. Given a test document and its correct categories, we first identify the queries that can retrieve this test document (ranking order is ignored here), and these queries are compared with the queries that can retrieve the test document given that the test document has the categories submitted by a system. To simplify the query comparison, we considered only the queries in the form of "retrieve documents with category x" and that specify only a single category. We call the category specified in the query (i.e., x) query category and denote it with the corresponding capital letter (i.e., X). In particular, we call query categories that can retrieve a document with correct categories correct query categories, and those for submitted categories candidate query categories. Recall and precision are calculated by comparing these two sets of query categories.

In the above example, since the test document with $\{b, f\}$ can be retrieved by two queries (retrieve documents with b and retrieve documents with f), the correct query categories are $\{B, F\}$. When a system assigned $\{c, f\}$ to the test document, the candidate query categories are $\{C, F\}$. By comparing the two sets, the recall becomes 1/2 and the precision becomes 1/2. These recall/precision values are the same as those obtained by the conventional evaluation. In general, if we restrict queries to the above-mentioned simple form, query categories are exactly the same as the categories of a test document, and the recall/precision of our retrieval-based evaluation is always the same as that of the conventional evaluation.

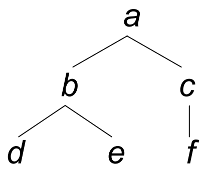

Figure 5. A category hierarchy.

If categories have a hierarchical structure like an F-term hierarchy, we may use queries that can retrieve documents with subcategories of the specified query category. A typical form is "retrieve documents with category x or one of the subcategories under x". To distinguish this type of query from the above-mentioned simple one, we note the query category in this type as X*. In the example, assuming a hierarchy in Figure 5, the document with $\{b, f\}$ can be retrieved by the queries with $\{B, F, A^*, B^*, C^*, F^*\}$ as query categories. Note that a document with a category x can be retrieved by two similar queries; one is the exact match query X of "retrieve documents with x" and another is the relaxed match query X* of "retrieve documents with x or one of the subcategories under x". Although the relaxed match query X* can retrieve all the documents which can be retrieved by the corresponding exact match query X, we distinguish the two queries as different ones in the recall/precision calculation. This is because the two queries retrieve different document sets depending on the different search purposes[5]. Coming back to the example, if a system assigned $\{c, f\}$ to the example document, the candidate query categories for this assignment would be $\{C, F, A^*, C^*, F^*\}$. By comparing this with the correct query categories $\{B, F, A^*, B^*, C^*, F^*\}$, we have a recall of /6 and a precision of /5. Remember that the recall and the precision of the conventional evaluation were both 1/2. Allowing partial matches on the category hierarchy makes the recall and the precision values larger than those by the conventional evaluation (or our evaluation using only exact match queries).

In summary, we evaluated the submitted results on the following two levels.

- Exact match (A): We used only exact match queries. The evaluation results by this method are the same as those by the conventional evaluation of directly comparing correct categories to submitted categories.
- Relaxed match (B): We used relaxed match queries in addition to exact match queries. This

[5] If a searcher knows that relevant documents should have x, the searcher uses the exact match query rather than the relaxed match query.

evaluation reflects partial matches on F-term hierarchies.

6. Evaluation Results

6.1 Approaches

We had six participating groups. Before showing the evaluation results, we briefly introduce the approaches of the participating groups. For more detailed information, refer to their original papers.

GATE

Their system was based on the SVM and a bag of words (BOW) representation was used. Using a morphological analyzer ChaSen, their system extracted words (nouns, verbs, adjectives, and unknown words) from full texts. They also used F-term descriptions in the PMGS to extract document features. In addition to the conventional SVM, they used an SVM variant (called H-SVM) for hierarchical classification. They evaluated the results using their original measure for the partial matching on category structure [ ].

JSPAT

Their system was based on the SVM. They used a different method from GATE's one for bundling binary (one-vs-rest) classifiers. Features were nouns and noun phrases extracted from full texts. They used F-term descriptions in the PMGS to tune the dictionary used by a morphological analyzer MeCab.

NCS

NCS constructed hybrid binary classifiers, each a naive Bayes model estimated from each patent component of the title, the bibliographic information (the applicants and the inventors), the abstract, the claims, and the description. MeCab extracted nouns, verbs, and adjectives as document features in a BOW representation. The classifiers for the five components were then combined based on the maximum entropy (ME) principle.

NICT

NICT used a K-NN approach where similarity was calculated based on the SMART measure or the BM25 measure. For each document, ChaSen analyzed the abstract and the claims and extracted nouns as a BOW representation.

NUT

NUT used a classifier where the chi-square statistics of words (nouns) and N-grams were estimated from the training data and were combined linearly into a single score. The targeting document fields were the abstract and the claims, and ChaSen was used as a morphological analyzer. They also used F-term descriptions in the PMGS to boost the weight of a term when the term appeared in the description of the targeting F-term.

RDNDC

RDNDC used a K-NN approach based on the BM25 similarity measure. The document representation was a set of nouns and compound nouns. These features were extracted from the abstract and the first claim using ChaSen.

6.2 Results and discussion

The six participating groups submitted a total of 3 runs. Figure 6 plots the 11pt interpolated precisions of the best runs from each group. The plots for both the exact match (A) and the relaxed match (B) are shown. The best run is the run with the highest MAP (Mean Average Precision) among runs from a group. Table 2 shows the MAP values and the F-measures of the best runs. The baseline performance was obtained by assigning F-terms to each test document in decreasing order of its frequency in the training documents.

Overall:

Among the best runs, the MAP values and the F-measures have almost the same order of ranking except that the top two runs are swapped between the MAP values and the F-measures.

The top two approaches are the one based on maximum entropy (NCS02) and the one based on SVM (GATE03). The difference between them is subtle. The recall/precision of the K-NN based approach (NICT01) is not as good as that of these two approaches, but the difference is not so large. This result is similar to those observed in the comparative studies of conventional text categorization. Unfortunately, none of our findings was particularly noticeable for multi-dimensional text categorization.

The effect of the partial match on the F-term hierarchy:

In Figure 6 and Table 2, we see that the results for the exact match (A) and the relaxed match (B) have exactly the same order of ranking. We will have to do theme-by-theme analysis to investigate the effect of the partial match. Some themes have deep F-term structures, but some have almost flat structures.

GATE did their own evaluation using their original measure for the partial match. Refer to their paper [ ] for more information.

The effect of using the patent structure:

GATE, JSPAT, and NCS used the full text, and NICT, NUT, and RDNDC used only the abstract and the claims. The results do not show a significant difference between these two approaches. Using the full text seems effective but NICT (NICT01) performed well only with the abstract and the claims. Note that NCS learned different classifiers from different parts of patent and combined them. NCS also was the only group that used bibliographic information (i.e., applicants and inventors) for classification.

The effect of using the PMGS:

GATE and NUT utilized the PMGS (F-term descriptions) in their classifiers. JSPAT used the PMGS to extend the dictionary of morphological analyzer. Although it is apparent that the PMGS contains very useful information for classifiers, no significant effect of the PMGS can be seen in Figure 6 and Table 2.

Document features:

All groups used the BOW representation, and the elements were basically words. JSPAT and RDNDC also used compound words but the effect can not be discerned in Figure 6 and Table 2. NUT proposed a hybrid approach using words and N-grams. All groups used ChaSen or MeCab as the morphological analyzer.

7. Conclusion

In the Classification Subtask of the NTCIR-6 Patent Retrieval Task, we released a test collection for patent categorization. The test collection was based on the F-term classification system, which has a multi-dimensional category structure. Using the test collection, we performed the task of assigning F-term categories to each document and evaluated the results.

References

[1] C.J. Fall, A. Torcsvari, K. Benzineb, and G. Karetka. Automated Categorization in the International Patent Classification. ACM SIGIR Forum, Vol.37, No.1, pp.10-25, 2003.

[2] A. Fujii, M. Iwayama, and N. Kando. Overview of Patent Retrieval Task at NTCIR- . In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization, 200 .

[3] M. Iwayama, A. Fujii, and N. Kando, Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task. In Proceedings of the Fifth NTCIR Workshop, 2005.

[ ] Y. Li, K. Bontcheva, and H. Cunningham, New Evaluation Measures for F-term Patent Classification. In Proceedings of the First International Workshop on Evaluating Information Access (EVIA 2007), 2007.

[5] I. Schellner, Japanese File Index classification and F-terms. World Patent Information, Vol.2 , pp.197-201, 2002.

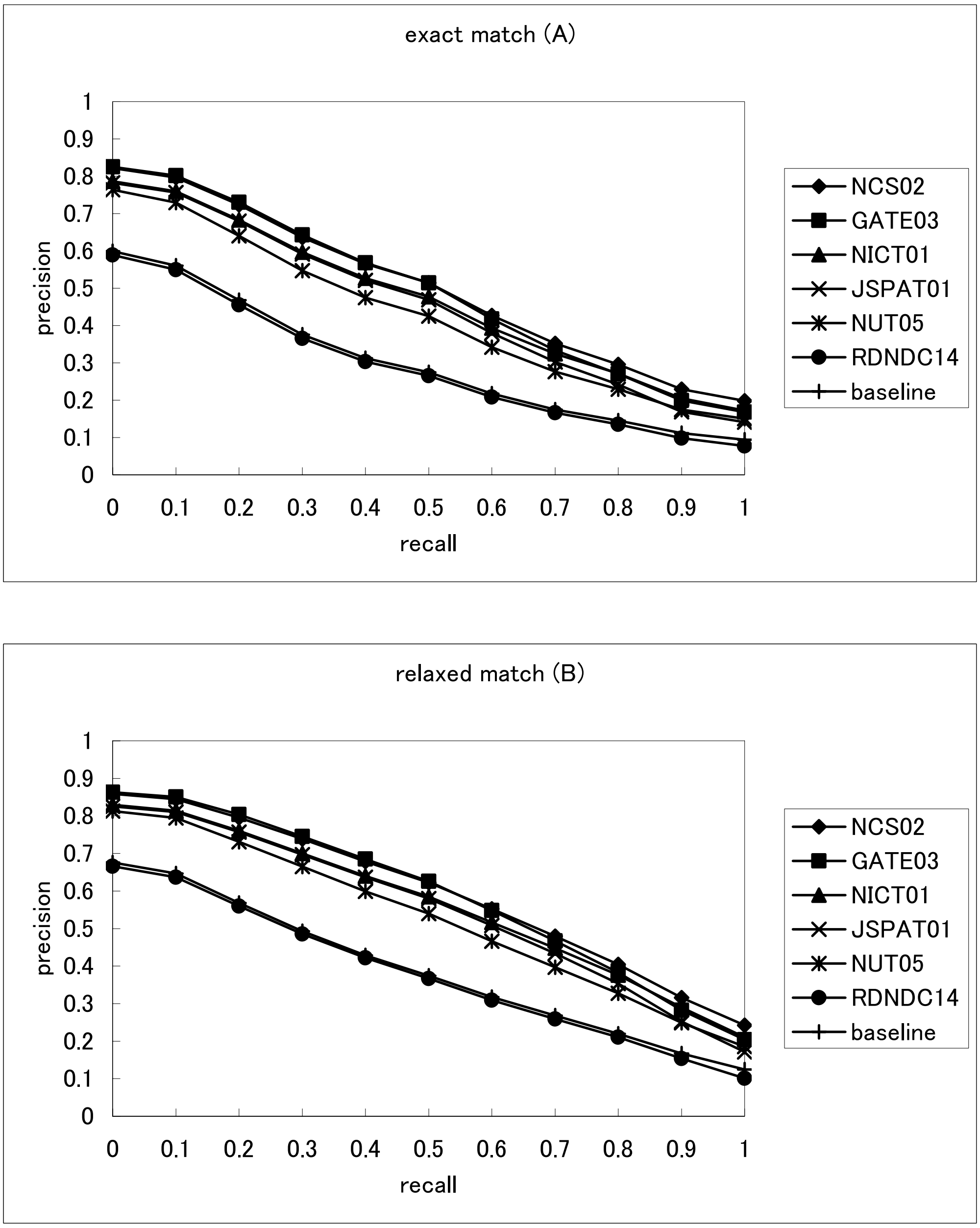

Figure 6. Recall/precision curves of best runs.

Table 2. MAPs and F-measures of best runs.

run	exact match (A)		relaxed match (B)	
	MAP	F-measure	MAP	F-measure
NCS02	0.4852	0.4037	0.5810	0.4970
GATE03	0.4779	0.4125	0.5755	0.5109
NICT01	0.4518	0.3840	0.5473	0.4767
JSPAT01	0.4381	0.3038	0.5355	0.3680
NUT05	0.4101	0.2432	0.5093	0.3838
RDNDC14	0.2717	0.2414	0.3622	0.3431
baseline	0.2821		0.3715	