Overview of the Patent Mining Task at the NTCIR-7 Workshop

Hidetsugu Nanba
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194, Japan

Atsushi Fujii
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba 305-8550, Japan

Makoto Iwayama
Hitachi, Ltd. / Tokyo Institute of Technology
1-280 Higashi-Koigakubu, Kokubunji 185-8601, Japan

Taiichi Hashimoto
Integrated Research Institute, Tokyo Institute of Technology
4259 Nagatsuta, Yokohama 226-8503, Japan

Abstract

This paper introduces the Patent Mining Task of the Seventh NTCIR Workshop and the test collections produced in this task. The task's goal was the classification of research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system, which is a global standard. For this task, 12 participant groups submitted 49 runs. In this paper, we also report the evaluation results of the task.

Keywords: *test collection, classification of research papers, patent, International Patent Classification (IPC)*

1 Introduction

The Patent Mining Task in the Seventh NTCIR Workshop (NTCIR-7) investigated the effective retrieval of necessary information from research papers and patent databases. In this paper, we introduce the task and report the evaluation results.

For a researcher in a field with high industrial relevance, retrieving research papers and patents has become an important aspect of assessing the scope of the field. Examples of these fields are bioscience, medical science, computer science, and materials science. In fact, the development of an information retrieval system of research papers and patents for academic researchers is central to the Intellectual Property Strategic Programs for 2006[1] and 2007[2] of the Intellectual Property Strategy Headquarters in the Cabinet Office, Japan.

In addition, research paper searches and patent searches are required by examiners in government Patent Offices, and by the intellectual property divisions of private companies. An example is the execution of an invalidity search among existing patents or research papers, which could invalidate a rival company's patents or patents under application in a Patent Office.

However, the terms used in patents are often more abstract or creative than those used in research papers, to try to widen the scope of the claims. Therefore, the Patent Mining Task aims to develop fundamental techniques for retrieving and classifying both research papers and patents.

In previous NTCIR Workshops, Patent Classification Subtasks were conducted [4][5]. In these subtasks, participants were asked to classify Japanese patent applications in terms of the File Forming Term (F-term) system, which is a classification system for Japanese patent documents. Here, we are focusing on the classification of research papers in addition to patents. The aim of the Patent Mining Task in NTCIR-7 was the classification of research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system, an alternative patent classification system.

The remainder of this paper is organized as follows. In Section 2, we describe some related works. In Section 3, we explain the task description. In Section 4, we describe the participants of the task. In Section 5, we report the evaluation results. Finally, we conclude in Section 6.

2 Related Works

The task of research paper classification into the IPC system is considered a cross-genre text classification study. Although patent classification tasks were conducted in the Fifth and Sixth NTCIR Workshops [4][5], and in Falls' study [1], they did not focus on cross-genre text classification. In other evaluation workshops, such as the Text REtrieval Conference (TREC)[3] and the Cross-Language Evaluation Forum (CLEF)[4], no tasks that focused on cross-genre information access have been conducted.

[1] http://www.kantei.go.jp/jp/singi/titeki2/keikaku2006_e.pdf
[2] http://www.kantei.go.jp/jp/singi/titeki2/keikaku2007_e.pdf
[3] http://trec.nist.gov/
[4] http://www.clef-campaign.org/

Although the goal of the task was not text classification, a related subtask was conducted in the Patent Retrieval Task in the Third NTCIR workshop [6]. This subtask aimed to retrieve patents relevant to a given newspaper article. In this task, Itoh et al. focused on "Term Distillation" [3]. The distribution of the frequency of the occurrence of words was considered to differ between heterogeneous databases. For example, the word "president" often appears in newspaper articles but seldom appears in patents. Therefore, unimportant words were assigned high scores when using the Term Frequency Inverse Document Frequency (TFIDF) method to weight words. Term Distillation is a technique that can prevent such cases by filtering out words that could be assigned incorrect weights.

There is another approach for cross-genre information retrieval. Nanba et al. proposed a method to integrate a research paper database and a patent database by analysing citation relations between research papers and patents [11]. For the integration, they extracted bibliographic information of cited literatures in "prior art" fields in Japanese patent applications. Using this integrated database, users can retrieve patents that relate to a particular research paper by tracing citation relations between research papers and patents. However, the number of cited papers among patent applications is not enough to retrieve related papers or patents, even though the number of opportunities for citing papers in patents or for citing patents in papers has been increasing recently.

Kamaya et al. proposed a method to paraphrase scholarly terms into patent terms (e.g., paraphrase "floppy disc" into "magnetic recording medium") [7]. They used hypernym-hyponymy relations between terms for the paraphrasing [10]. Some patent terms (e.g., "magnetic recording medium") are the hyponyms of scholarly terms (e.g. "floppy disc"). Therefore, a paraphrase of some scholarly terms can be realized by finding their hypernyms. However, some scholarly terms do not have such a relationship with their corresponding patent terms. For example, a hypernym of the scholarly term "machine translation" is "natural language processing," but the terms "automatic translation" or "language translation" are used in patents instead of "natural language processing". Therefore, they also used citation relationships between research papers and patents for paraphrasing. Generally, a research paper and a patent that have citation relationships with each other, tend to be in the same research field. Using this idea, a paraphrase of a scholarly term can be realised by using the following procedure:

1. Retrieve research papers that contain a given scholarly term in their titles.
2. Collect patents that have citation relationships with the papers retrieved in Step 1.
3. Extract patent terms from patents collected in Step 2.
4. Output patent terms extracted in Step3.

They combined the hypernym-hyponym-based and the citation-based methods, and finally obtained the patent terms as the paraphrase of the given scholarly term.

These related works are a part of solutions for cross-genre information access. The organizers of the Patent Mining Task expected participants to propose other fundamental techniques for cross-genre information access.

3 The Patent Mining Task

3.1 Task Overview

As we described in Section 1, the goal of the Patent Mining Task was the classification of research papers into the IPC system, which is a global standard hierarchical patent classification system. One or more IPC codes are manually assigned to each patent, aiming for effective patent retrieval.

The sixth edition of the IPC system contains more than 50,000 classes at the most detailed level. The goal of this task was to assign one or more of these 50,000 classes to a given topic, as expressed in terms of the title and abstract of a research paper. An example of a topic is shown in Figure 1. Here, <TOPIC-ID> specifies the topic identification number, and <TITLE> and <ABSTRACT> specify the title and abstract of the research paper to be classified.

<TOPIC>
<TOPIC-ID> 100 </TOPIC-ID>
<TITLE> DTMF (Dual Tone Multi-Frequency) transmission method for a mobile communication system </TITLE>
<ABSTRACT> A highly efficient speech-encoding scheme called VSELP is adopted for Japanese digital mobile communication systems. However, DTMP (Dual Tone Multi-Frequency) signals are distorted by using this encoding scheme. This paper presents a DTMF signal transmission scheme. DTMF signals are transmitted in the form of call control messages from mobile stations (MS) to the mobile control centre (MCC). In addition, necessary control capabilities in MS and MCC are described. </ABSTRACT>
</TOPIC>

Figure 1. An example of a topic in the English subtask

Within the overall task, the following subtasks were conducted.

- Japanese subtask: classification of Japanese research papers using patent data written in Japanese.
- English subtask: classification of English research papers using patent data written in English.

In addition to these subtasks, we also conducted the following more challenging subtasks, which require both cross-genre and cross-lingual information access techniques.

- Cross-lingual subtask (J2E): classification of Japanese research papers using patent data written in English.
- Cross-lingual subtask (E2J): classification of English research papers using patent data written in Japanese.

These four subtasks are summarized in Figure 2.

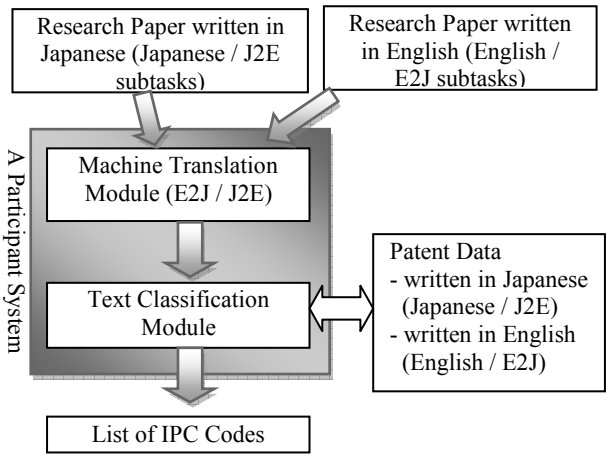

Figure 2. Summary of subtasks

In the next subsection, we describe in detail the patent data used in these subtasks.

3.2 Patent Data

An overview of the patent data used in each subtask is shown in Table 1. In the following, we describe details of the data.

Table 1. Document Sets

Data	Year	Size	Number	Language
(1) Unexamined Japanese patent applications	1993–2002	100 GB	3.50M	Japanese
(2) USPTO patent data	1993–2000	33 GB	0.99M	English
(3) Patent Abstracts of Japan (translated into English)	1993–2002	4.2 GB	3.50M	English
(4) NTCIR-1 and NTCIR-2 CLIR Task test collection (Abstracts of research papers)	1988–1999	1.4 GB	0.26M	Japanese/ English

(1) Unexamined Japanese Patent Applications

These data were distributed to the teams participating in the Japanese subtask and the Cross-lingual subtask (E2J). To standardize the format of the documents, the organizers provided an official tool, which inserts SGML-style tags into each document. Table 2 shows the tags inserted by that tool. Although passages were extracted from the specific fields, such as claims and detailed descriptions of the invention, any fields can be used for categorization purposes.

Table 2. Tags for Japanese Patent Applications

Tags	Description
<DOC>	document
<DOCNO>	document identifier
<TEXT>	text body
<PASSAGE>	passage
<PNUM>	passage identifier

(2) USPTO Patent Data

These data were distributed to the groups participating in the English subtask and the Cross-lingual subtask (J2E). To standardize the format of the documents, the organizers provided an official tool, which inserts SGML-style tags into each document. Table 3 shows the tags inserted by that tool. Because the format of the source data was more complicated than that for the Japanese patent applications, we inserted a large number of tags to enhance the readability of the USPTO patent data. The participant groups were allowed to use <DOC>, <DOCNO>, <TITLE>, <ABST>, <SPEC>, and <CLAIM> for the classification purpose.

Table 3. Tags for USPTO Patent Data

Tags	Description
<DOC>	document
<DOCNO>	document identifier
<APP-NO>	application number
<APP-DATE>	application date
<PUB-NO>	publication number
<PUB-TYPE>	publication type
<PUB-DATE>	publication date
<PRI-IPC>	primary IPC
<IPC-VER>	IPC version
<PRI-USPC>	primary USPC
<PRIORITY>	priority information
<CITATION>	citation(s)
<INVENTOR>	inventor(s)
<ASSIGNEE>	assignee(s)
<TITLE>	title
<ABST>	abstract
<SPEC>	specification
<CLAIM>	claim(s)

(3) Patent Abstracts of Japan (PAJs)

These data were distributed to the groups participating in the English and J2E subtasks. The tags shown in Table 4 were assigned to each document in PAJs. Participant groups were allowed to use all tags.

Table 4. Tags for Patent Abstracts of Japan

Tags	Description
<B110>	number of the patent document
<B121>	plain language designation of the kind of document
<B130>	kind of document code according to WIPO Standard ST.16
<B190>	WIPO Standard ST.3 code, or other identification, of the office or organization publishing the document
<B210>	number(s) assigned to the application(s)
<B220>	date(s) of filing the application(s)
<B310>	number(s) assigned to priority application(s)
<B320>	date(s) of filing of priority application(s)
<B511> <B512>	International Patent Classification
<B542>	title of the invention
<B711>	name(s) of applicant(s)
<B721>	name(s) of inventor(s) if known to be such

(4) NTCIR-1 and NTCIR-2 CLIR Task Test Collection

This database was distributed to all participant groups, and they were allowed to use it for any purposes. The database was originally used in the Cross-lingual Information Retrieval (CLIR) tasks in the first and second NTCIR Workshops (NTCIR-1 and NTCIR-2) [8][9]. It contains 255,960 records of Japanese-English paired documents, with each record comprising a title, the author(s), an abstract, keywords, a publication year, and a conference name.

3.3 Relevance Judgements

Sets of topics with manually assigned IPC codes are necessary for the evaluation. However, it is very costly and time consuming to create such data sets. Therefore, we have produced the data sets using the following idea.

Essentially, an invention is not patentable if it was already known before the date of filing. However, Article 30 in the Japanese patent law provides a six-month grace period for disclosures made via a publication or a presentation at a conference or exhibition. In this case, the applicants must mention the proceedings' title (or the conference name) and the date it was published in an "Indication of exceptions to lack of novelty" field (or *exception field*) in the patent. Figure 3 gives an example of an exception field.

(original)
【新規性喪失の例外の表示】特許法第３０条第１項適用申請有り２０００年３月１４日　社団法人情報処理学会発行の「第６０回（平成１２年前期）全国大会講演論文集（４）」に発表

(English translation)
[Indication of exceptions to lack of novelty] The provisions set forth in Article 30, Paragraph 1 in Japanese patent law. Proceedings (Volume 4) of the 60th Annual Meeting of the Information Processing Society of Japan, published in March 14, 2000.

Figure 3. An example of the "Indication of exceptions to lack of novelty" field

We can assume that most of the content of the paper mentioned in the exception field overlaps with the patent. Therefore, if we regard the IPC codes that were assigned to the patent as the codes that should be assigned to the research paper mentioned in the exception field, it becomes possible to create a large-scale data set at low cost. In fact, there are more than 9,000 applications with exception fields in the 3,496,253 Japanese patent applications published in the 10-year period 1993–2002.

The procedure used to create the data set was as follows. Firstly, we extracted publication years and proceedings titles from the exception fields in the 9,000 applications. Although the title and authors of a paper are not mentioned in the exception field, the authors are usually the same as the inventors of the patent. We therefore extracted and used the inventors of the patent instead of the authors' names.

Secondly, we compared these extracted data with records in a research paper database using a simple string matching method. From this automatic matching, we obtained, on average, six candidate records for each exception field.

Thirdly, we manually identified the correct match from among the candidate records. Here, we identified the match from the following two viewpoints.

- A paper in an exception field and a candidate paper are exactly the same (group A)
- Authors and research topics of the two papers are almost the same, but the publication years are different (group B)

We obtained 976 pairs (525 pairs of group A and 451 of group B) of matching patents and research papers.

From these pairs, we created English and Japanese topics (titles and abstracts) and their correct classifications (IPC codes extracted from patents). For each topic, an average of 2.3 IPC codes was assigned.

We then randomly assigned 97 topics to the "dry run" and the remaining 879 topics to the "formal run".

Table 5 shows the breakdown of the topic numbers used in the dry run and the formal run.

Table 5. Breakdown of Topics

	group A	group B
dry run	100-151	200-244
formal run	300-772	1000-1405

The dry run data were provided to the participant teams as training data for the formal run. A list of pairs of a patent ID and one or more IPC codes were also provided as additional training data[5]. These IPC codes were extracted from each patent in the data sets (1), (2), and (3).

Participant teams were asked to submit one or more ranked lists[6] of IPC codes for each topic, to be evaluated using Mean Average Precision (MAP), Recall, and Precision measurements. To calculate these measurements for each submitted run, the organizers produced a Perl program that was compatible with the trec_eval program[7]. The values for MAP, Recall, and Precision can potentially be different depending on the version of trec_eval used.

4 Participants

We had 24 participating systems for the Japanese subtask, 20 for the English subtask, and five for the Cross-lingual subtask. As far as the number of groups is concerned, we had 12 participating groups of universities and companies. Table 6 shows the breakdown of the groups.

Table 6. Breakdown of Participants (Please note that one group consists of universities in North America and Asian Countries)

	Japan	Other Asian Countries	Europe	North America
University	3	4	0	2
Company	2	0	1	0

The number of runs for each subtask was as follows.

- Japanese subtask: 24 runs from five groups
- English subtask: 21 runs from nine groups
- J2E: five runs from two groups

There were no runs submitted to E2J.

5 Results

5.1 Evaluation Results of the Patent Mining Task

We show the evaluation results for the Japanese, English, and Cross-lingual subtasks in Tables 7, 8, and 9, respectively. We also show the recall-precision curves for each subtask in Figures 4, 5, and 6. The systems "HTC13", "NEUN1_S1", and "xrce_j2e" obtained the best scores in Japanese, English, and Cross-lingual subtasks, respectively. These systems employed the k-Nearest Neighbours (k-NN) method in common, and exceeded 0.4 of MAP scores. In addition to these, many other systems, such as HCU1 and HCU2 in the Japanese subtask and "xrce_e2j2e", "xrce_en_lm", "xrce_en_filter", and "xrce_en_pp" in the English subtask, also employed the k-NN method.

On the other hand, some systems using machine learning approaches also obtained remarkable results. The "nttcs4" system in the Japanese subtask applied logistic regression to the dry run data for tuning of the model, and obtained 39.64 of the MAP score for the formal run data. In the English subtask, "nttcs2" obtained 34.79 of the MAP score using the hybrid model of logistic regression and Naïve Bayes.

5.2 Comparison with the Results of Patent Classification

To investigate the effectiveness of a cross-genre text classification system, we compared it with a patent classification system. For the cross-genre text classification system, we used "HCU1" for the Japanese subtask. The "HCU1" system comprises a patent retrieval engine [10] developed for the Patent Retrieval Task in NTCIR-6 [2]. The engine introduced the Vector Space Model as a retrieval model and SMART [12] for term weighting. The "HCU1" system obtained a list of IPC codes using the follow procedure.

1. Retrieve top 170 results using the patent retrieval engine for a given query (research paper).
2. Extract IPC codes with relevance scores for the query from each retrieved patent in step 1.
3. Rank IPC codes using the following equation.

$$\text{Score}(X) = \sum_{r=1}^{n} \text{Relevance score of each patent}$$

Here, X and n indicate an IPC code and the number of patents that X is assigned to within the top 170 retrieved patents, respectively.

We also used this system for patent classification. As we described in Section 3.3, each query has its counterpart in a patent database. We regarded these counterparts as queries and conducted patent classification using the "HCU1" system. We used the full

[5] From the results of the dry run, we found that detecting patents that are the counterparts of given research papers (topics) and using them for the classification of papers was more effective than using all the data in the list. We therefore removed the data for the 976 patents from the list and distributed the list to the participant teams before the formal run, because the approach of detecting the counterpart patents is not our intended purpose here.

[6] The maximum number of IPC codes for a single topic is 1,000.

[7] http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz

text of patents as queries. Instead of using the top 170 retrieved patents, we used the top 20 for the calculation of each IPC code in Step 3, which was determined using the dry run data.

From these experiments, we obtained 37.06 for the MAP score, which is almost the same as the 39.13 obtained for the "HCU1" system, as shown in Table 3. This result indicates that the performance of "HCU1" as a cross-genre text classification reached almost the same level as an intra-genre text classification.

Although the patent classification system used full texts in patents, the MAP score was rather lower than that for the cross-genre text classification system, which used titles and abstracts of research papers as queries. One of the reasons for this result could be the way terms are used in claims. As we mentioned in Section 1, more abstract and creative terms are used in patents than those used in research papers, to try to widen the scope of the claims. In particular, the terms in claims are more abstract than those used in specifications. It is therefore considered that using terms in specifications may obtain a better MAP score than using whole terms in patents for patent classification. This may also indicate that using specifications instead of full texts can improve the performance of cross-genre text classification.

Table 7. MAP for Japanese Subtask

Run ID	MAP	Run ID	MAP
HTC13	44.02	HTC04	41.65
HTC11	43.71	nttcs4	39.64
HTC12	43.61	*HCU1	39.13
HTC07	43.60	*HCU2	39.06
HTC01	43.34	HTC14	38.62
HTC06	43.29	nttcs3	35.72
HTC05	43.26	nttcs2	34.35
HTC08	43.23	nttcs1	33.03
HTC10	43.18	KECIR	27.27
HTC03	42.68	*HCU3	14.12
HTC02	42.36	nut1-1	6.98
HTC09	42.27	nut2-1	4.06

(HCU1, HCU2, and HCU3 are the task organizer's systems)

Table 8. MAP for English Subtask

Run ID	MAP	Run ID	MAP
NEUN1_S1	48.86	rali2	14.37
NEUN1_S2	47.21	ICL07	14.36
NEUN1_S3	44.53	rali1	14.23
xrce_e2j2e	42.45	ICL07_2	13.39
xrce_en_lm	42.09	BRKLY-PM-EN-02	12.65
xrce_en_filter	41.83	AINLP04	10.45
xrce_en_pp	41.49	BRKLY-PM-EN-04	9.90
nttcs2	34.79	AINLP01	9.78
nttcs1	33.74	BRKLY-PM-EN-03	9.37
KECIR	29.03	PI-5b	3.79

Table 9. MAP for Cross-lingual Subtask (J2E)

Run ID	MAP
xrce_j2e	43.80
AINLP05	10.70
AINLP06	10.41
AINLP02	9.41
AINLP03	9.34

6 Conclusion

We have given an overview of the evaluation and design of the Patent Mining Task in NTCIR-7. We focused on the "Indication of exceptions to lack of novelty" field in Japanese patent applications and thereby created 976 English and Japanese topics and their correct classifications (IPC codes). Forty-nine runs from 12 participant groups were submitted to the formal run, and the systems using the k-NN method obtained the best MAP scores in each subtask.

References

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

[2] Fujii, A., Iwayama, M., and Kando. N. 2007. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. Proceedings of the 6th NTCIR Workshop Meeting.

[3] Itoh, H., Mano, H., Ogawa, Y. 2002. Term Distillation for Cross-db Retrieval, Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task.

[4] Iwayama, M., Fujii, A., and Kando, N. 2007. Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task. Proceedings of the 6th NTCIR Workshop Meeting.

[5] Iwayama, M., Fujii, A., and Kando, N. 2005. Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task. Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.

[6] Iwayama, M., Fujii, A., Kando, N., Takano, A. 2002. Overview of Patent Retrieval Task at NTCIR-3. Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task.

[7] Kamaya, H., Nanba, H., Okumura, M., Shinmori, A., Tanigawa, H., and Suzuki, T. 2007. Paraphrasing Scholarly Terms into Patent Terms using Citation Relations between Research Papers and Patents. IPSJ SIG Notes NL-178, pp.97-102. (in Japanese)

[8] Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., and Hidaka, S. 1999. Overview of IR Tasks at the First NTCIR Workshop. Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.11-44.

[9] Kando, N., Kuriyama, K., and Yoshioka, M. 2001. Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop. Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, pp.4-37 – 4-60.

[10] Nanba, H. 2007. Query Expansion using an Automatically Constructed Thesaurus. Proceedings of the 6th NTCIR Workshop, pp.414-419.

[11] Nanba, H., Anzen, N., and Okumura, M. Automatic Extraction of Citation Information in Japanese Patent Applications. International Journal on Digital Libraries.

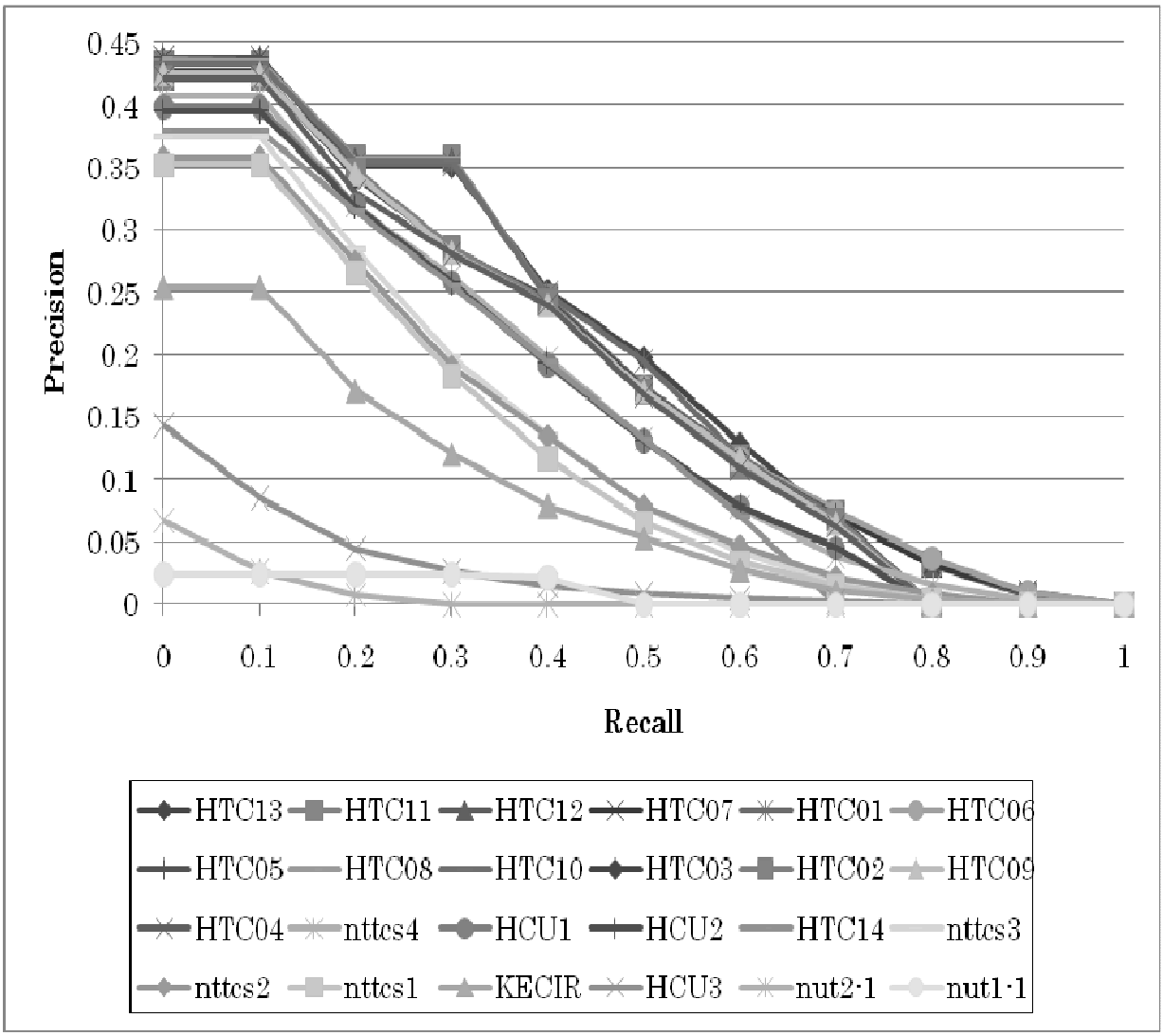

Figure 3. Recall-precision Curves for All Topics (Japanese Subtask)

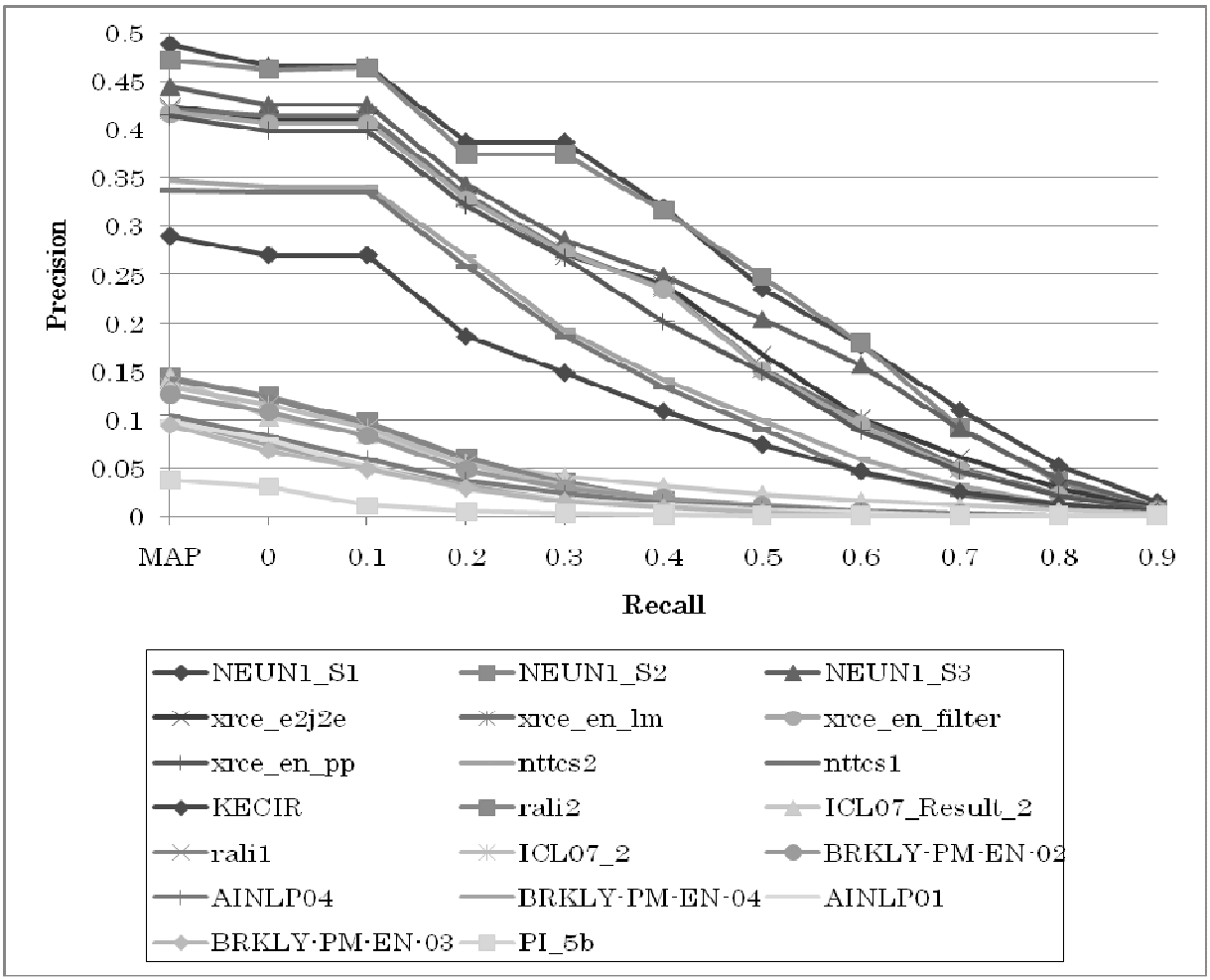

Figure 4. Recall-precision Curves for All Topics (English Subtask)

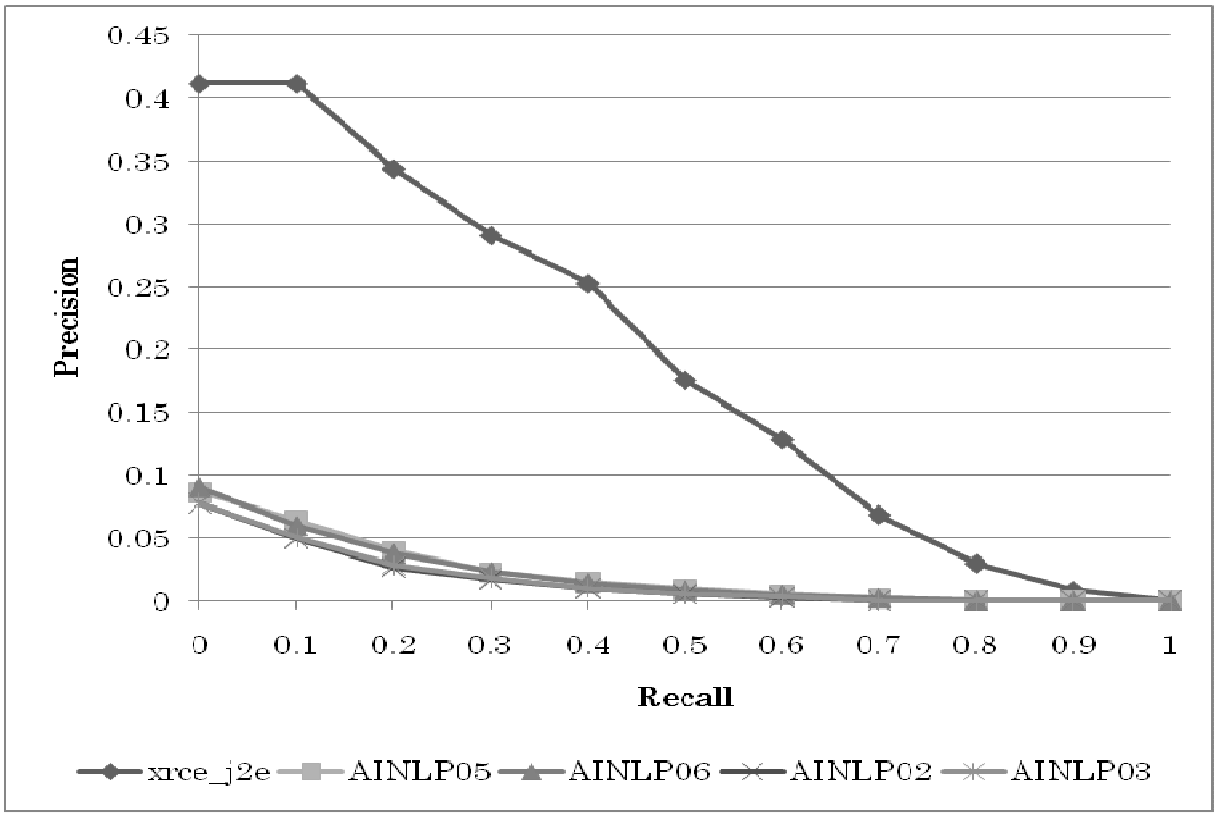

Figure 5. Recall-precision Curves for All Topics (Cross-lingual Subtask: J2E)