Classification of Research Papers into a Patent Classification System Using Two Translation Models

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194 Japan
nanba@hiroshima-cu.ac.jp

Toshiyuki Takezawa
Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194 Japan
takezawa@hiroshima-cu.ac.jp

Abstract

Classifying research papers into patent classification systems enables an exhaustive and effective invalidity search, prior art search, and technical trend analysis. However, it is very costly to classify research papers manually. Therefore, we have studied automatic classification of research papers into a patent classification system. To classify research papers into patent classification systems, the differences in terms used in research papers and patents should be taken into account. This is because the terms used in patents are often more abstract or creative than those used in research papers in order to widen the scope of the claims. It is also necessary to do exhaustive searches and analyses that focus on classification of research papers written in various languages. To solve these problems, we propose some classification methods using two machine translation models. When translating English research papers into Japanese, the performance of a translation model for patents is inferior to that for research papers due to the differences in terms used in research papers and patents. However, the model for patents is thought to be useful for our task because translation results by patent translation models tend to contain more patent terms than those for research papers. To confirm the effectiveness of our methods, we conducted some experiments using the data of the Patent Mining Task in the NTCIR-7 Workshop. From the experimental results, we found that our method using translation models for both research papers and patents was more effective than using a single translation model.

1 Introduction

Classification of research papers into patent classification systems makes it possible to conduct an exhaustive and effective prior art search, invalidity search, and technical trend analysis. However, it would be too costly and time-consuming to have the research paper's authors or another professional classify such documents manually. Therefore, we have investigated the classification of research papers into a patent classification system.

In previous studies, classification of patents was conducted as subtasks in the 5th and 6th NTCIR workshops (Iwayama *et al.*, 2005; Iwayama *et al.*, 2007). In these subtasks, participants were asked to classify Japanese patents using the File Forming Term (F-term) system, which is a classification system for Japanese patents. Here, we have focused on the classification of research papers, and we need to take into account the differences in terms used in research papers and patents because the terms used in patents are often more abstract or creative than those used in research papers in order to widen the scope of the claims. For example, the scholarly term "machine translation" can be expressed as "automatic translation" or "language conversion" in patent documents. In addition to taking the differences of genres into account, it is necessary to do exhaustive searches and analyses focusing on the classification of research papers written in various languages.

To solve these problems, we propose some classification methods using two machine translation models. When translating English research papers into Japanese, the performance of a translation model for patents is generally inferior to that for research papers, because the terms used

Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 27–35,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

in patents are different from those in research papers. However, we thought that a translation model for patents might be useful for our task, because translation results using the patent translation model tend to contain more patent terms than those obtained using the model for research papers. In this paper, we confirm the effectiveness of our methods using the data of the Cross-genre Subtask (E2J) in the 7th NTCIR Workshop (NTCIR-7) Patent Mining Task (Nanba *et al.*, 2008:b).

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 describes our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section 4 reports the experimental results. We present some conclusions in Section 5.

2 Related Work

In this section, we describe some related studies on "cross-genre information access" and "cross-lingual information access".

Cross-genre Information Access

Much research has been done in the field of cross-genre information retrieval and document classification. The technical survey task in the NTCIR-3 workshop (Iwayama *et al.*, 2002) is an example. This task aimed to retrieve patents relevant to a given newspaper article. In this task, Itoh *et al.* (2002) focused on "Term Distillation". The distribution of the frequency of the occurrence of words was known to be different between newspaper articles and patents. For example, the word "president" often appears in newspaper articles, while this word seldom appears in patents. As a result, unimportant words such as "president" were assigned high scores in patents when using tf*idf to weight words. Term Distillation is a technique that can prevent such cases by filtering out words that can be assigned incorrect weights. This idea was also used to link news articles and blog entries (Ikeda *et al.*, 2006).

Another approach for cross-genre information retrieval was that used by Nanba *et al.* (2008:a), who proposed a method to integrate a research paper database and a patent database by analyzing citation relations between research papers and patents. For the integration, they extracted bibliographic information of cited literature 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 sufficient 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.

As another approach for cross-genre information retrieval, Nanba *et al.* (2009) proposed a method to paraphrase scholarly terms into patent terms (e.g., paraphrasing "floppy disc" into "magnetic recording medium"). They focused on citation relationships between research papers and patents for the paraphrased terms. Generally, a research paper and a patent that have a citation relationship tend to be in the same research field. Therefore, they paraphrased a scholarly term into a patent term in two steps: (1) retrieve research papers that contain a given scholarly term in their titles, and (2) extract patent terms from patents that have citation relations with the retrieved papers.

The NTCIR-7 Patent Mining Task (Nanba *et al.*, 2008:b) is another example of research done on information access using research papers and patents. The aim of the Patent Mining Task was to classify research papers written in either Japanese or English using the International Patent Classification (IPC) system, which is a global standard hierarchical patent classification system. The following four subtasks were included in this task, and 12 groups participated in three of them: Japanese, English, and Cross-lingual (J2E) subtasks.

- **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.
- **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.

Because the number of categories (IPC codes) that research papers were classified into was very large (30,855), only two participating groups employed machine learning, which is the most standard approach in the NLP field. The other groups used the k-Nearest Neighbor (k-NN) method. Among all participant groups, only Mase and Iwayama's group (2008) coped with the problem of the differences in terms between re-

search papers and patents. Mase and Iwayama used a pseudo-relevance feedback method to collect related patent terms for a given research paper. First, they retrieved patents relevant to a given research paper. Next, they extracted patent terms from the top n retrieved patents. Then they retrieved patents again using the patent terms extracted in the second step. Finally, they classified research papers using the k-NN method. However, they reported that a simple k-NN based method was superior to the method based on the pseudo-relevance feedback method. In this paper, we also examined our methods using the data of the NTCIR-7 Patent Mining Task.

TREC Chemistry Track [1] is another related study involving research papers and patents. This track aims for cross-genre information retrieval using research papers and patents in the chemical field. This track started in 2009 under the Text Retrieval Conference (TREC), and the details including experimental results will be reported at the final meeting to be held in November 2009.

Cross-lingual Information Access

Much research has been done on cross-lingual information access using research papers and patents. In the NTCIR workshop, cross-lingual information retrieval tasks have been carried out using research papers (Kando *et al.*, 1999; Kando *et al.*, 2001) and patents (Fujii *et al.*, 2004; Fujii *et al.*, 2005; Fujii *et al.*, 2007). In the CLEF evaluation workshop, the cross-lingual patent retrieval task "CLEF-IP" was initiated in 2009[2]. The cross-lingual subtask in the NTCIR-7 Patent Mining Task (Nanba *et al.*, 2008:b) is another cross-lingual information access study.

Here, we describe two methods used in the cross-lingual subtask (J2E) in the Patent Mining Task (Bian and Teng, 2008, Clinchant and Renders, 2008). Bian and Teng (2008) translated Japanese research papers into English using three online translation systems (Google, Excite, and Yahoo! Babel Fish), and classified them using a k-NN-based text classifier. Clinchant and Renders (2008) automatically obtained a Japanese-English bilingual dictionary from approximately 300,000 pairs of titles from Japanese and English research papers (Kando *et al.*, 1999) using Giza[3], a statistical machine translation toolkit. Then they classified papers using this dictionary and a k-NN-based document classifier. Bian and Clinchant also participated in an English subtask and obtained almost the same mean average precision (MAP) scores as those of the J2E subtask.

Although the direction of translation of our system is different from Bian and Clinchant, we also tried our methods using the data of the cross-lingual subtask (E2J). We utilized the Giza toolkit in the same way as Clinchant, but our approach was different from Clinchant, because we solved the problem of "differences of terms used in research papers and patents" by using two translation models obtained from both research papers and patents parallel corpora.

3 Classification of Research Papers into a Patent Classification System

3.1 Our Methods

We explain here the procedure of our cross-genre, cross-lingual document classification method depicted in Figure 1. The goal of our task is to classify document *I* written in language L1 in genre G1 into a classification system (categories) using documents written in language L2 in genre G2, and classification codes were manually annotated to each of these documents. Generally, three steps are required for cross-genre, cross-lingual document classification: (1) translate document *I* into Language L2 using a translation model for genre G1 (document *O* in Figure 1), (2) paraphrase terms in document *O* into terms in genre G2 (document *O′*), and (3) classify *O′* into a classification system. Here, if a translation model for genre G2 is available, steps (1) and (2) can be resolved using this translation model, because terms in the translation results using the model are more appropriate in genre G2. However, as it is assumed that the translation model translates documents in genre G2, the translation results might contain more mistranslations than the results obtained by a model for genre G1. We therefore combine translation results (*O*+*O′*) produced by translation models for genre G1 and for G2. These results can be expected to contain terms in genre G2 and to minimize the effects of mistranslation by using the translation model for genre G1.

[1] https://wiki.ir-facility.org/index.php/TREC_Chemistry_Track

[2] http://www.ir-facility.org/the_irf/current-projects/clef-ip09-track/

[3] http://www.fjoch.com/GIZA++.html

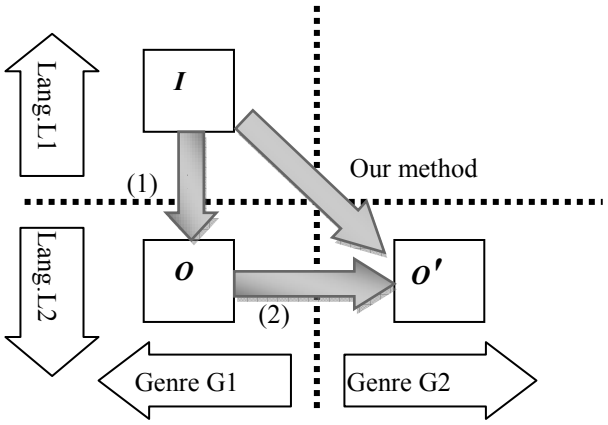

Figure 1: Overview of our method

3.2 System Configuration

The goal of our study is to classify English research papers (Language L1=English, Genre G1=research papers) into a patent classification using a patent data set written in Japanese (Language L2=Japanese, Genre G2=patents). Figure 2 shows the system configuration. Our system is comprised of a "Japanese index creating module" and a "document classification module". In the following, we explain both modules.

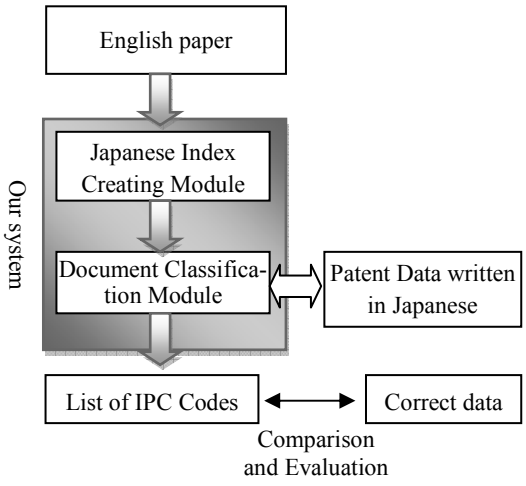

Figure 2: System configuration

Japanese Index Creating Module

When a title and abstract pair, as shown in Figure 3, is given, the module creates a Japanese index, shown in Figure 4[4], using translation models for research papers and for patents.

Here, the following two procedures (A) or (B) are possible for creating a Japanese index from an English paper: (A) translate the English title and abstract into Japanese; then create a Japanese index from them by extracting content terms[5], or (B) create an English index[6] from the English title and abstract, then translate each index term into Japanese. We conducted experiments using both procedures.

As translation tools, we used Giza and Moses[7]. We obtained translation models using a patent bilingual corpus containing 1,800,000 pairs of sentences (Fujii *et al.* 2008) and a research paper bilingual corpus containing 300,000 pairs automatically created from datasets of NTCIR-1 (Kando *et al.* 1999), and 2 (Kando *et al.* 2001) CLIR tasks.

Title: A Sandblast-Processed Color-PDP Phosphor Screen
Abstract: Barrier ribs in the color PDP have usually been fabricated by multiple screen printing. However, the precise rib printing of fine patterns for the high resolution display panel is difficult to make well in proportion as the panel size grow larger. On the other hand, luminance and luminous efficiency of reflective phosphor screen will be expected to increase when the phosphor is deposited on the inner wall of display cells. Sandblasting technique has been applied to make barrier ribs for the high resolution PDP and nonffat phosphor screens on the inner wall of display cells.

Figure 3: Example of an English title and abstract

18 形成 (formation)
18 ＰＤＰ (PDP)
18 型蛍光面 (type phosphor screen)
12 障壁形成 (barrier formation)
12 障壁 (barrier)
12 蛍光 (phosphor)
12 カラーＰＤＰ (color PDP)
12 反射型蛍光 (reflective phosphor)
12 型蛍光 (type phosphor)
12 サンドブラスト法 (Sandblasting technique)
9 サンドブラスト (Sandblasting)
(snip)

Figure 4: Example of a Japanese index

[4] Numerical values shown with index terms indicate term frequencies.

[5] As content terms, we extracted noun phrases (series of nouns), adjectives, and verbs using the Japanese morphological analyzer MeCab. (http://mecab.sourceforge.net)

[6] We used TreeTagger as a POS tagging tool. (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)

[7] http://www.statmt.org/moses/

We used two phrase tables for research papers and patents when translating English index terms into Japanese. For a given English term, we selected the Japanese term with the highest translation probability from the candidates in each table. These tables were automatically obtained in the process of constructing translation models for research papers and patents using Giza and Moses. However, there are several other ways to translate index terms, such as using bilingual dictionaries of technical terms or compositional semantics (Tonoike *et al.*, 2007), we employed a phrase table-based method because the effectiveness of this method was experimentally confirmed by Itakagi *et al.* (2007). In addition to this method, we also investigated using bilingual dictionaries of technical terms as baseline methods. Details of these methods are in Section 4.2.

Document Classification Module
We used Nanba's k-NN-based system (Nanba, 2008:c) for a Japanese subtask as a document classification module in our system. This module uses a patent retrieval engine (Nanba, 2007) which was developed for the NTCIR-6 Patent Retrieval Task (Fujii *et al.*, 2007). This engine introduced the Vector Space Model as a retrieval model, SMART (Salton, 1971) for term weighting, and noun phrases (sequence of nouns), verbs, and adjectives for index terms. The classification module obtained a list of IPC codes using the following procedure.

1. Retrieve top 170 results using the patent retrieval engine for a given 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.

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

Here, X and n indicate the IPC code and the number of patents that X was assigned to within the top 170 retrieved patents, respectively. Nanba determined the value of 170 using the dry run data and the training data of the NTCIR-7 Patent Mining Task.

3.3 Classification of Research Papers into International Patent Classification (IPC)

As a patent classification system for classification of research papers, we employed the International Patent Classification (IPC) system. The IPC system is a global standard hierarchical patent classification system. The sixth edition of the IPC contains more than 50,000 classes at the most detailed level[8]. The goal of our task was to assign one or more of these IPC codes at the most detailed level to a given research paper.

4 Experiments

To investigate the effectiveness of our method, we conducted some experiments. Section 4.1 describes the experimental procedure. Section 4.2 explains several methods that were compared in the experiments. Section 4.3 reports the experimental results, and Section 4.4 discusses them.

4.1 Experimental Method

We conducted some experiments using the data of the cross-lingual subtask (E2J) in the NTCIR-7 Patent Mining Task.

Correct data set
We used a data set for the formal run of the cross-lingual subtask in the NTCIR-7 Patent Mining Task (Nanba, *et al.*, 2008). In the data set, IPC codes were manually assigned to each 879 topics (research papers). For each topic, an average of 2.3 IPC codes was manually assigned. These correct data were compared with a list of IPC codes[9] by systems, and the systems were evaluated in terms of MAP (mean average precision). Here, the 879 topics were divided into two groups: group A, in which highly relevant IPC codes were assigned to 473 topics, and group B, in which relevant IPC codes were assigned to 406 topics. In our experiment, we evaluated several systems in two ways: using group A only and using both groups.

Document Sets
An overview of document sets used in our experiments is in Table 1. In the unexamined Japanese patent applications, manually assigned IPC codes are included together with full text patent data. These data were utilised to apply the k-NN method in our document classification module. NTCIR-1 and 2 CLIR Task test collections were used to obtain a translation model for research papers, which we mentioned in Section 3.2.

[8] Among 50,000 classes, 30,855 classes relevant to academic fields were used in the NTCIR-7 Patent Mining Task.
[9] The maximum number of IPC codes allowed to be output for a single topic was 1,000.

Data	Year	Size	No.	Lang.
Unexamined Japanese patent applications	1993 - 2002	100 GB	3.50 M	Japanese
NTCIR-1 and 2 CLIR Task	1988 - 1999	1.4 GB	0.26 M	Japanese /English

Table 1: Document sets

4.2 Alternatives

We conducted examinations using seven baseline methods, three proposed methods, and two upper-bound methods shown as follows. In the following, "SMT(X)" is a method to create a Japanese index after translating research papers using a translation model X. "Index(X)" is a method to create an English index, and to translate the index terms using a phrase table for translation model X.

Baseline methods

- SMT(Paper): Create a Japanese index after translating research papers using a translation model for research papers.
- SMT(Patent): Create a Japanese index after translating research papers using a model for patents.
- Index(Paper): First create an English index, then translate the index terms into Japanese using a phrase table for research papers.
- Index(Patent): First create an English index, then translate the index terms into Japanese using a phrase table for patents.
- SMT(Paper)+Hypernym: Paraphrase index terms created from "SMT(Paper)" by their hypernyms using a hypernym-hyponym thesaurus.
- Index(TechDic): Translate English index terms using a Japanese-English dictionary consisting of 450,000 technical terms[10].
- Index(EIJIRO): Translate English index terms using EIJIRO[11], a Japanese-English dictionary consisting of more than 1,000,000 pairs of terms.

Our methods

- Index(Paper)*Index(Patent): Product set of "Index(Paper)" and "Index(Patent)".
- Index(Paper)+Index(Patent): Union of "Index(Paper)" and "Index(Patent)".
- SMT(Paper)+Index(Patent): Union of "SMT(Paper)" and "Index(Patent)".

Upper-bound methods

- Japanese subtask: This is the same as the Japanese subtask in the NTCIR-7 Patent Mining Task. For this subtask, Japanese research papers, which are manual (ideal) translations of corresponding English papers, are input into a system.
- Japanese subtask+Index(Patent): Union of "Japanese subtask" and "Index(Patent)".

Another reason for using the baseline methods is that the terms used in patents are often more abstract or creative than those used in research papers, as mentioned in Section 1. Therefore, we paraphrased index terms in SMT(Paper) by their hypernyms using a hypernym/hyponym thesaurus (Nanba, 2007). Nanba automatically created this thesaurus consisting of 1,800,000 terms from 10 years of unexamined Japanese patent applications using a set of patterns, such as "NP_0 ya NP_1 nadono NP_2 (NP_2 such as NP_0 and NP_1)" (Hearst, 1992).

4.3 Experimental Results

Experimental results are given in Table 2. From the results, we can see that "SMT(Paper)" obtained the highest MAP scores when using topics in group A+B and in group A. Of the 10 methods used (except for the upper-bound methods), our method "SMT(Paper)+Index(Patent)" obtained the highest MAP score.

4.4 Discussion

Difference of terms between research and patents (Comparison of "Index(Paper)" and "Index(Patent)")

Although the quality of phrase tables for research papers ("Index(Paper)") and patents ("Index(Patent)") was not very different, the MAP score of "Index(Paper)" was 0.01 better than that of "Index(Patent)". To investigate this gap, we compared Japanese indices by "Index(Paper)" and "Index(Patent)". There were 69,100 English index terms in total, and 47,055 terms (47,055/69,100=0.681) were translated by the model for research papers, while 40,427 terms (40,427/69,100=0.585) were translated by the model for patents. Ten percent of this gap indicates that terms used in research papers and in patents are different, which causes the gap in MAP scores of "Index(Patent)" and "Index(Paper)".

[10] "Kagakugijutsu 45 mango taiyakujiten" Nichigai Associates, Inc., 2001.
[11] http://www.eijiro.jp/

Combination of "Index(Paper)" and "Index(Patent)"

When a term translated by the model for research papers matches a term translated by the model for patents, they seem to be a correct translation. Therefore, we examined "Index(Paper)*Index(Patent)". The method uses terms as an index when translation results by both models match. From the experimental results, this method obtained 0.1830 and 0.2230 of MAP scores when using topics in group A+B and in group A, respectively. These results indicate that the overlap of lexicons between research papers and patents is relatively large, and terms in this overlap are effective for our task. However, the MAP score of "Index(Paper)*Index(Patent)" was 0.02 lower than "Index(Paper)" and "Index(Patent)", which indicates that there are not enough terms in the overlap for our task.

In addition to "Index(Paper)*Index(Patent)", we also examined "Index(Paper)+Index(Patent)", which is a union of "Index(Paper)" and "Index(Patent)". From the experimental results, we obtained respective MAP scores of 0.2258 and 0.2596 when using topics in group A+B and in group A. These scores are 0.01 to 0.02 higher than the scores of "Index(Paper)" and "Index(Patent)". These encouraging results indicate that our method using two translation models is effective for a cross-genre document classification task.

Effectiveness of "SMT(Paper) +Index(Patent)"

In addition to "Index(Paper)", "SMT(Paper)" also obtained high MAP scores. Therefore, we combined "Index(Patent)" with "SMT(Paper)" instead of "Index(Paper)". From the experimental results, we found that this approach ("SMT(Paper)+Index(Patent)") produced MAP scores of 0.2633 when using topics in group A+B and 0.2807 when using topics in group A. These scores were the highest of all, almost approaching the results of upper-bound methods.

Comparison of "Index(TechDic)", "Index(EIJIRO)", "Index(Paper)", and "Index(Patent)"

Both "Index(TechDic)" and "Index(EIJIRO)" were worse than "Index(Paper)" and "Index(Patent)" by more than 0.05 in the MAP scores. These results were due to the lower number of terms translated by each method. Because phrase tables for research papers and patents were automatically created, they were not as correct as "TechDic" and "EIJIRO". However, the phrase tables were able to translate more English terms into Japanese in comparison with "TechDic" (30,008/69,100=0.434) and "EIJIRO" (37607/69,100=0.544), and these induced the difference of MAP scores.

Comparison of "SMT(Paper)+Hypernym" and "SMT(Paper)"

"SMT(Paper)+Hypernym" impaired "SMT(Paper)", because the method paraphrased unnecessary terms into their hypernyms. As a result, irrelevant patents were contained within the top 170 search results, and the k-NN method ranked irrelevant IPC codes at higher levels. Our methods using two translation models are different from "SMT(Paper)+Hypernym" in this point because two translation models translate into the same term when a scholarly term need not be paraphrased.

Classification of Japanese research papers using "Index(Patent)"

As we mentioned above, the "Index(Paper)+Index(Patent)" and "SMT(Paper)+Index(Patent)" models improved the MAP scores of both "Index(Paper)" and "SMT(Paper)". We further investigated whether "Index(Patent)" could also improve monolingual document classification ("Japanese subtask+Index(Patent)"). In this method, a Japanese index was created from a manually written Japanese research paper, and this was combined with "Index(Patent)". The results showed that "Japanese subtask+Index(Patent)" could slightly improve MAP scores when using topics in group A+B and in group A.

Practicality of our method

Recall values for the top n results by "SMT(Paper)+Index(Patent)", which obtained the highest MAP score, are in Table 3. In this table, the results using all topics (group A+B) and the topics in group A are shown. The results indicate that almost 40% of the IPC codes were found within top 10 results, and 70% were found within the top 100. For practical use, we need to improve recall at the top 1, but we still believe that these results are useful for supporting beginners in patent searches. It is often necessary for searchers to use patent classification codes for effective patent retrieval, but professional skill and much experience are required to select relevant IPC codes. In such cases, our method is useful to look for relevant IPC codes.

	Method	group A+B	group A
Our methods	Index(Paper)*Index(Patent)	0.1830	0.2230
	Index(Paper)+Index(Patent)	0.2258	0.2596
	SMT(Paper)+Index(Patent)	**0.2633**	**0.2897**
Baseline methods	SMT(Paper)	0.2518	0.2777
	SMT(Patent)	0.2214	0.2507
	Index(Paper)	0.2169	0.2433
	Index(Patent)	0.2000	0.2373
	SMT(Paper)+Hypernym	0.2451	0.2647
	Index(TechDic)	0.1575	0.1773
	Index(EIJIRO)	0.1347	0.1347
Upper-bound	Japanese subtask	0.2958	0.3267
	Japanese subtask+Index(Patent)	0.3001	0.3277

Table 2: Evaluation results

5 Conclusion

We proposed several methods that automatically classify research papers into the IPC system using two translation models. To confirm the effectiveness of our method, we conducted some examinations using the data of the NTCIR-7 Patent Mining Task. The results showed that one of our methods "SMT(Paper)+Index(Patent)" obtained a MAP score of 0.2897. This score was higher than that of "SMT(Paper)", which used translation results by the translation model for research papers, and this indicates that our method is effective for cross-genre, cross-lingual document classification.

rank	group A	group A+B
1	0.117 (131/1115)	0.110 (226/2051)
2	0.186 (207/1115)	0.169 (347/2051)
3	0.239 (267/1115)	0.215 (440/2051)
4	0.278 (310/1115)	0.250 (512/2051)
5	0.311 (347/1115)	0.277 (567/2051)
10	0.420 (468/1115)	0.377 (774/2051)
20	0.524 (584/1115)	0.467 (958/2051)
50	0.659 (735/1115)	0.597 (1224/2051)
100	0.733 (817/1115)	0.673 (1381/2051)
500	0.775 (864/1115)	0.728 (1494/2051)
1000	0.775 (864/1115)	0.728 (1494/2051)

Table 3: Recall for top n results (SMT(Paper)+Index(Patent))

References

Guo-Wei Bian and Shun-Yuan Teng. 2008. Integrating Query Translation and Text Classification in a Cross-Language Patent Access System, *Proceeding of the 7th NTCIR Workshop Meeting*: 341-346.

Stephane Clinchant and Jean-Michel Renders. 2008. XRCE's Participation to Patent Mining Task at NTCIR-7, *Proceedings of the 7th NTCIR Workshop Meeting*: 351-353.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. Overview of Patent Retrieval Task at NTCIR-4, *Working Notes of the 4th NTCIR Workshop*: 225-232.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2005. Overview of Patent Retrieval Task at NTCIR-5, *Proceedings of the 5th NTCIR Workshop Meeting*: 269-277.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the Patent Retrieval Task at NTCIR-6 Workshop, *Proceedings of the 6th NTCIR Workshop Meeting*: 359-365.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proceedings of the 7th NTCIR Workshop Meeting*: 389-400.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the 14th International Conference on Computational Linguistics*: 539-545.

Daisuke Ikeda, Toshiaki Fujiki, and Manabu Okumura. 2006. Automatically Linking News Articles to Blog Entries, *Proceedings of AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs*: 78-82.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic Validation of Terminology Translation Consistency with Statistical Method, *Proceedings of MT summit XI*: 269-274.

Hideo Itoh, Hiroko Mano, and Yasushi Ogawa. 2002. Term Distillation for Cross-db Retrieval, *Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task*: 11-14.

Makoto Iwayama, Atsushi, Fujii, Noriko Kando, and Akihiko Takano. 2002. Overview of Patent Re-

trieval Task at NTCIR-3, *Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task*: 1-10.

Makoto Iwayama, Atsushi Fujii, and Noriko Kando. 2005. Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task, *Proceedings of the 5th NTCIR Workshop Meeting*: 278-286.

Makoto Iwayama, Atsushi Fujii, and Noriko Kando. 2007. Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task, *Proceedings of the 6th NTCIR Workshop Meeting*: 366-372.

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. 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*: 11-44.

Noriko Kando, Kazuko Kuriyama, and Makoto Yoshioka. 2001. Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop, *Proceedings of the 2nd NTCIR Workshop Meeting*: 4-37 - 4-60.

Hisao Mase and Makoto Iwayama. 2008. NTCIR-7 Patent Mining Experiments at Hitachi, *Proceedings of the 7th NTCIR Workshop Meeting*: 365-368.

Hidetsugu Nanba. 2007. Query Expansion using an Automatically Constructed Thesaurus, *Proceedings of the 6th NTCIR Workshop Meeting*: 414-419.

Hidetsugu Nanba, Natsumi Anzen, and Manabu Okumura:a. 2008. Automatic Extraction of Citation Information in Japanese Patent Applications, *International Journal on Digital Libraries*, 9(2): 151-161.

Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto:b. 2008. Overview of the Patent Mining Task at the NTCIR-7 Workshop, *Proceedings of the 7th NTCIR Workshop Meeting*: 325-332.

Hidetsugu Nanba:c. 2008. Hiroshima City University at NTCIR-7 Patent Mining Task. *Proceedings of the 7th NTCIR Workshop Meeting*: 369-372.

Hidetsugu Nanba, Hideaki Kamaya, Toshiyuki Takezawa, Manabu Okumura, Akihiro Shinmori, and Hidekazu Tanigawa. 2009. Automatic Translation of Scholarly Terms into Patent Terms, *Journal of Information Processing Society Japan TOD*, 2(1): 81-92. (in Japanese)

Gerald Salton. 1971. *The SMART Retrieval System - Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ.

Masatsugu Tonoike. Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sakai, Takehito Utsuro, and Satoshi Sato. 2005. Translation Estimation for Technical Terms using Corpus Collected from the Web, *Proceedings of the Pacific Association for Computational Linguistics*: 325-331.