Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques

Hidetsugu Nanba[1], Toshiyuki Takezawa[1], Kiyoko Uchiyama[2], Akiko Aizawa[2]
[1] Hiroshima City University
[2] National Institute of Informatics
[1] 3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194 JAPAN
[2] 2-1-2 Hitotsubashi, Chiyodaku, Tokyo 101-8430 JAPAN
E-mail: nanba@hiroshima-cu.ac.jp, takezawa@hiroshima-cu.ac.jp, kiyoko@nii.ac.jp, aizawa@nii.ac.jp

Abstract

Retrieving research papers and patents is important for any researcher assessing the scope of a field with high industrial relevance. However, the terms used in patents are often more abstract or creative than those used in research papers, because they are intended to widen the scope of claims. Therefore, a method is required for translating scholarly terms into patent terms. In this paper, we propose six methods for translating scholarly terms into patent terms using two synonym extraction methods: a statistical machine translation (SMT)-based method and a distributional similarity (DS)-based method. We conducted experiments to confirm the effectiveness of our method using the dataset of the Patent Mining Task from the NTCIR-7 Workshop. The aim of the task was to classify Japanese language research papers (pairs of titles and abstracts) using the IPC system at the subclass (third level), main group (fourth level), and subgroup (the fifth and most detailed level). The results showed that an SMT-based method (SMT_ABST+IDF) performed best at the subgroup level, whereas a DS-based method (DS+IDF) performed best at the subclass level.

Keywords: patent, research paper, distributional similarity, statistical machine translation

1. Introduction

Retrieving research papers and patents is important for any researcher assessing the scope of a field with high industrial relevance. However, the terms used in patents are often more abstract or creative than those used in research papers, because they are intended to widen the scope of claims. Therefore, a method for translating scholarly terms into patent terms is required. In this paper, we propose several methods for translating scholarly terms into patent terms. Several techniques have been proposed for obtaining paraphrases or synonyms using a statistical machine translation technique (SMT) (Callison-Burch et al., 2006; Quirk et al., 2004; Zhao et al., 2008; Zhou et al., 2006) and a distributional similarity (DS) technique (Lee, 1999; Lin, 1998). We extended these techniques to the translation of scholarly terms into patent terms, and we confirm their effectiveness experimentally using the dataset from the Patent Mining Task at the NTCIR-7 Workshop (Nanba et al., 2008).

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 proposes our method for translating scholarly terms into patent terms using synonym extraction techniques. Section 4 presents our experimental investigation of the effectiveness of our method and a discussion of our results. Finally, we provide our conclusions in Section 5.

2. Related Work

There has been much research in the field of cross-genre information retrieval, such as that presented in the technical survey task of the Patent Retrieval Task at the Third NII Test Collection for Information Retrieval (NTCIR) Workshop (Iwayama et al., 2002). This task aimed to retrieve patents relevant to a given newspaper article. Itoh et al. (2002) focused on Term Distillation, where the distribution of the frequency word occurrence was considered to be different in heterogeneous databases.

Therefore, unimportant words are assigned high scores when using TFIDF to weight words. Term Distillation is a technique for preventing the incorrect assignment of weights by filtering out words.

However, some patent terms, such as *magnetic recording device*, only appear in a patent database and Term Distillation cannot be applied in such cases.

To address this problem, Nanba et al. (2009) proposed a method for paraphrasing scholarly terms into patent terms (e.g., paraphrasing *floppy disc* into *magnetic recording medium*). This method focused on citation relationships of the paraphrased terms among research papers and patents. Generally, a research paper and a patent that have citation relationships tend to belong to the same research field. Therefore, they paraphrased a scholarly term into a patent term using 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. However, their approach assumed that a researcher would sequentially input individual scholarly terms into their translation system, and the decision of "which scholarly term should be translated" belonged to the user. In general, many scholarly terms do not need to be translated into patent terms in research papers. In this paper, we evaluated our method by applying it to the task of research paper classification with the International Patent Classification (IPC) system, which is a global standard hierarchical patent classification system. In this study, the problem of "which scholarly term should be translated" was decided automatically, which was impossible with Nanba's approach. In addition to this problem, Nanba's method was also not easily applied to other languages.

In another study of cross-genre information access, the Patent Mining Task was conducted at the NTCIR-7 and -8 Workshops (Nanba et al., 2008, 2010). At these workshops, research papers were classified using the IPC system. We therefore used this dataset to confirm the effectiveness of our methods.

Another related research project is TREC Chemistry Track[1], initiated in 2009, which focuses on information access in chemistry research papers and patents.

3. Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques

We propose two translation methods: an SMT-based method and a DS-based method. We describe these methods in the following subsections.

3.1 Statistical Machine Translation-based Method

Several methods for obtaining paraphrases or synonyms using statistical machine translation techniques have been proposed (Callison-Burch et al., 2006; Quirk et al., 2004; Zhao et al., 2008; Zhou et al., 2006). If the translations of two expressions X and Y are the same expression, then the expressions X and Y are considered to be paraphrases. Based on this concept, Zhou et al. (2006) automatically obtained paraphrases from a translation model, which were created from pairs of English and Chinese sentences using the SMT technique. These paraphrases were then used for evaluating computer-produced summaries.

Here, if *高分解能* (high definition) and *high resolution* are aligned using a translation model for research papers, and *高解像度* (high resolution) and *high resolution* are aligned based on a patent model, we can translate the scholarly term *高分解能* (high definition) into *高解像度* (high resolution) as the corresponding patent term.

3.2 Distributional Similarity-based Method

Lin (1998) and Lee (1999) proposed a method for calculating the similarity between terms, which is known as DS. The underlying assumption of their approach was that semantically similar words are used in similar contexts. Therefore, the similarity between two terms can be defined as the amount of information contained in the commonality between the terms, divided by the amount of information in the contexts of the terms.

We automatically translate a scholarly term into a patent term using DS in the following procedure.

1. Analyze the dependency structures of all sentences in a research paper database using the Japanese dependency parser CaboCha[2].
2. Extract noun-phrase-verb (with a postpositional particle) pairs that have dependency relations from the dependency trees obtained in Step 1.
3. Count the frequencies of each noun-phrase-verb pair.
4. Collect verbs and their frequencies for each noun phrase, creating indices for each noun phrase.
5. Create indices from a patent database in the same way as a research paper database (Steps 1 to 4).
6. Calculate the similarities between two indices of noun phrases created in Steps 4 and 5 using the SMART similarity measure (Salton, 1971).
7. Obtain a noun phrase with the highest similarity score as a translation of a given scholarly term.

Figure 1 shows an example of two indices for a scholarly term *フロッピーディスク* (floppy disc) and for a patent term *磁気記録媒体* (magnetic recording device), both of which are created in Steps 4 and 5, respectively.

scholarly term	**patent term**
フロッピーディスク (floppy disc)	*磁気記録媒体* (magnetic recording device)
3 に_書き込む **(write_to)**	**4 に_書き込む** **(write_to)**
2 に_収める (store_to)	2 を_作る (create)
2 に_取り込む (import_to)	**2 は_読み取る** **(read)**
1 は_読み取る **(read)**	1 を_傷つける (break)

Figure 1: Example of two indices for a scholarly term and a patent term

4. Experiments

To confirm the effectiveness of our method, we conducted a series of experiments.

4.1 Experimental Conditions

Task

In this study, we used the dataset from the Patent Mining Task at the NTCIR-7 Workshop (Nanba et al., 2008). The aim of the Patent Mining Task was to classify research papers (pairs of titles and abstracts) that were written in Japanese using the IPC system at the subclass (third level), main group (fourth level), and subgroup (the fifth and most detailed) levels. The IPC system is a global-standard hierarchical patent classification system. One or more IPC codes are assigned manually to each

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

[2] http://code.google.com/p/cabocha/

patent to ensure effective patent retrieval. The sixth edition of the IPC system contains more than 50,000 classes at the most detailed (subgroup) level. Table 1 shows an example of an IPC code "A01B 1/02."

<table>
<tr><td>A</td><td>01</td><td>B</td><td>1</td><td>/02</td></tr>
<tr><td>Section</td><td colspan="4"></td></tr>
<tr><td colspan="2">Class</td><td colspan="3"></td></tr>
<tr><td colspan="3">Subclass</td><td colspan="2"></td></tr>
<tr><td colspan="4">Main group</td><td></td></tr>
<tr><td colspan="5">Subgroup</td></tr>
</table>

A	Human necessities
A01	Agriculture; forestry; hunting; etc.
A01B	Soil-working during agriculture or forestry; parts, details or accessories of agricultural machines or implements, in general.
A01B 1	Hand tools
A01B 1/02	Spades; shovels

Table 1: IPC code example for "A01B 1/02."

Experimental Data and Evaluation Measures

We used the dataset for a formal run of the Japanese subtask in the NTCIR-7 Patent Mining Task. IPC codes were manually assigned to all 879 topics in the dataset (research papers). An example of a topic is shown in Figure 2, where <TOPIC-ID> specifies the topic identification number, while <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 2. Example of a topic (translated into English).

On average, 2.3 IPC codes were manually assigned to each topic. These correct data were then compared with a list of IPC codes using the classification system, and the system was evaluated in terms of MAP (mean average precision).

Document Classification System

We used a k-NN-based document classification system. This system introduced the Vector Space Model as a retrieval model, SMART for term weighting, and noun phrases (sequence of nouns), verbs, and adjectives as index terms. The classification module produced 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 query relevance scores for each patent retrieved in step 1.
3. Rank IPC codes using the following equation.

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

where X indicates the IPC code and n is the number of patents to which X was assigned in the top 170 retrieved patents. Here, the value of 170 was determined in our previous work (Nanba, 2008).

Translation of Scholarly Terms Using Our Methods

We translated scholarly terms into patent terms using the methods described in Sections 3.1 and 3.2. We considered that it was not necessary to translate high-frequency scholarly terms such as "study" or "method" into patent terms, because their translation might impair the classification accuracy. Therefore, we tested the following two methods.

(1) Translating scholarly terms when their inverse document frequency (IDF) scores were lower than a threshold value.
(2) Translating all scholarly terms.

Translation Models

We used Giza[3] and Moses[4] as translation tools. 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 1,763,217 pairs of CiNii database[5] titles (TITLE model). In addition to these models, we also obtained a model for research papers based on 600,000 pairs of sentences found in the CiNii database abstracts (ABST model).

Distributional Similarity

To calculate the DS, we used 600 million sentences from a Japanese patent database, which covered a total of 10 years. We also used 600,000 sentences from the CiNii database abstracts.

Alternatives

We conducted tests using the following six methods and a baseline method.

[3] http://code.google.com/p/giza-pp/
[4] http://www.statmt.org/moses/
[5] http://ci.nii.ac.jp/

Our Methods

- **SMT_ABST**: Translate scholarly terms using SMT-based method with an ABST model.
- **SMT_ABST+IDF:** Do not translate scholarly terms if their IDF scores are lower than a threshold value, when using the SMT_ABST method.
- **SMT_TITLE**: Translate scholarly terms using SMT-based method with TITLE model.
- **SMT_TITLE+IDF:** Do not translate scholarly terms if their IDF scores are lower than a threshold value, when using the SMT_TITLE method.
- **DS**: Translate scholarly terms using the DS-based method.
- **DS+IDF**: Do not translate scholarly terms if their IDF scores are lower than a threshold value when using the DS method.

Baseline Method

- **PAPER:** Use scholarly terms without translation.

4.2 Results

The experimental results in Table 2 show that most of our methods using the SMT technique (SMT_ABST, SMT_ABST+IDF, SMT_TITLE, and SMT_TITLE+IDF) were superior to the baseline method. However, the DS-based method (DS and DS+IDF) slightly improved the baseline method at the subclass level.

Methods	Subgroup (5th level)	Main Group (4th level)	Subclass (3rd level)
SMT_ABST	0.3786	**0.5186**	0.6691
SMT_ABST+IDF	**0.3812**	**0.5197**	0.6709
SMT_TITLE	**0.3797**	**0.5208**	0.6688
SMT_TITLE+IDF	**0.3799**	**0.5204**	0.6710
DS	**0.3793**	0.5182	0.6717
DS+IDF	**0.3794**	0.5175	**0.6744**
PAPER (baseline)	0.3792	0.5185	0.6720

Table 2: MAP scores by our methods and a baseline method.

4.3 Discussion

Effectiveness of IDF

The DS method incorrectly translated the general (high-frequency) scholarly term 提案手法 (our method) into 残留黒鉛 (residual black lead), whereas the DS+IDF method did not translate this term, because the IDF score was very low. Most of the methods that included IDF scores performed better than those without IDF scores. Based on this result, we suggest that it is not necessary to translate general scholarly terms into patent terms.

The MAP score of the SMT_TITLE method was approximately the same as that of the SMT_TITLE+IDF method. This was because general terms, such as 提案手法 (our method) or 本研究 (this study), appear rarely in the titles of research papers so these terms were not translated by the SMT_TITLE method.

Comparison of the Two Translation Methods

The experimental results showed that the DS method tended to translate scholarly terms into related terms with the same properties, whereas the SMT_ABST and SMT_TITLE methods tended to translate them into synonymous terms. For example, the scholarly term ワードプロセッサ (word processor) was translated into ドローソフト (drawing software) using the DS method. However, although "word processor" and "drawing software" both refer to computer software, they are not synonyms. Thus, the DS-based method was suitable for retrieving a wider range of patents, whereas the SMT-based method was more applicable for retrieving a narrower range. DS performed better at the subclass level, whereas SMT_ABST performed better at the subgroup level.

5. Conclusions

In this study, we proposed six methods for translating scholarly terms into patent terms using synonym extraction techniques. We conducted experiments to confirm the effectiveness of our methods using the dataset of the Patent Mining Task at the NTCIR-7 Workshop. We found that the SMT-based method (SMT_ABST+IDF) performed best at the subgroup level, whereas the DS-based method (DS+IDF) performed best at the subclass level. We also found that most of the methods with IDF scores performed better than those without IDF scores indicating that it is not necessary to translate general scholarly terms into patent terms.

References

Callison-Burch, C., Koehn, P., and Osborne, M. (2006) Improved Statistical Machine Translation Using Paraphrases. In *Proc. NAACL 2006*, pp.17--24.

Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro T. (2008). Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting*, pp.389--400.

Itoh, H., Mano, H., and Ogawa, Y. (2002). Term Distillation for Cross-DB Retrieval. In *Proc. 3rd NTCIR Workshop Meeting*.

Iwayama, M., Fujii, A., Kando, and N., Takano, A. (2002). Overview of Patent Retrieval Task at NTCIR-3. In *Proc. 3rd NTCIR Workshop Meeting*.

Lee, L. (1999). Measures of Distributional Similarity. In *Proc. 37th ACL*, pp.25--32.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proc. 36th ACL and 17th COLING*, pp.768--774.

Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. (2008). Overview of the Patent Mining Task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting*, pp.325--332.

Nanba, H., Kamaya, H., Takezawa, T., Okumura, M., Shinmori, A., and Tanigawa, H. (2009). Automatic Translation of Scholarly Terms into Patent Terms. In *Proc. 2nd International CIKM Workshop on Patent Information Retrieval (PaIR'09)*, pp.21--24.

Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. (2010). Overview of the Patent Mining Task at the NTCIR-8 Workshop. In *Proc. 8th NTCIR Workshop Meeting*, pp.293--302.

Nanba, H. (2008). Hiroshima City University at NTCIR-7 Patent Mining Task. In *Proc. 7th NTCIR Workshop Meeting*, pp.369--372.

Quirk, C., Brockett, C., and Dolan, W. (2004) Monolingual Machine Translation for Paraphrase Generation. In *Proc. EMNLP 2004*, pp.142--149.

Salton, G. (1971). The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ.

Zhao, S., Niu, C., Zhou, M., Liu, T., and Li, S. (2008) Combining Multiple Resources to Improve SMT-based Paraphrasing Model. In *Proc. ACL-HLT 2008*, pp.1021--1029.

Zhou, L., Lin, C.-Y., Munteanu, D.S., and Hovy, E. (2006). ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In *Proc. HLT-NAACL 2006*, pp.447--454.