Extraction and Visualization of Technical Trend Information from Research Papers and Patents

Satoshi Fukuda
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
fukuda@ls.info.hiroshima-cu.ac.jp

Hidetsugu Nanba
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
+81-82-830-1584
nanba@hiroshima-cu.ac.jp

Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima, 731-3194 Japan
+81-82-830-1768
takezawa@hiroshima-cu.ac.jp

ABSTRACT
For a researcher in a field with high industrial relevance, retrieving and analyzing research papers and patents have become an important aspect of assessing the scope of the field. We propose a method for creating a technical trend map automatically from both research papers and patents. For the construction of the technical trend map, we focus on the elemental (underlying) technologies used in a particular field, and their effects. Knowledge of the history and effects of the elemental technologies used in a particular field is important for grasping the outline of technical trends in the field. Therefore, we have constructed a method that can recognize the application of elemental technologies and their effects in any research field. To investigate the effectiveness of our method, we conducted an experiment using the data in the NTCIR-8 Patent Mining Task. From our experimental results, we obtained recall and precision scores of 0.254 and 0.496, respectively, for the analysis of research papers. We also obtained recall and precision scores of 0.455 and 0.507, respectively, for the analysis of patents. Finally, we have constructed a system that creates an effective technical trend map for a given field.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Search process
H.3.4 [**System and Software**]: Performance evaluation
H.3.5 [**Online Information Services**]: Data sharing

General Terms
Measurement, Performance, Experimentation.

Keywords
Information extraction, SVM, Domain adaptation

1. INTRODUCTION
In this paper, we propose a method for creating a technical trend map automatically from both research papers and patents. This map will enable users to grasp the outline of technical trends in a particular field.

For a researcher in a field with high industrial relevance, retrieving and analyzing research papers and patents have become important aspects of assessing the scope of the field. 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 through existing patents and research papers, which could invalidate a rival company's patents or patents pending in a patent office. However, it is costly and time-consuming to collect and read all of the papers in the field. Therefore, there is a need for automatic analysis of technical trends.

For the construction of technical trend maps, we have focused on the elemental (underlying) technologies used in a particular field, and their effects. Knowledge of the history and effects of the elemental technologies used in a field is essential for analyzing technical trends in the field. Therefore, we have constructed a system that can recognize the application of elemental technologies and their effects for any research field.

The remainder of this paper is organized as follows. Section 2 shows the system behavior in terms of snapshots. Section 3 describes related work. Section 4 explains our method for analyzing the structure of research papers and patents. Section 5 reports on these experiments, and discusses the results. We present some conclusions in Section 6.

2. SYSTEM BEHAVIOR
In this section, we describe our system that visualizes technical trends. Figure 1 shows a technical trend map for the "image recognition" field. In this figure, several elemental technologies used in the image recognition field, such as "FPGA (Field Programmable Gate Array)", are listed in the left-hand column. The effects of each technology, such as "約33%と著しい速度向上 (remarkable speed improvement of about 33%)", are shown in the right-hand column. These technologies and effects were extracted automatically from research papers and patents in this field, and each research paper and patent is shown as a dot in the figure. The x-axis indicates the publication years for the research papers and patents. Moving the cursor over a dot causes bibliographic information about the research paper or the patent to be shown in a pop-up window.

If the user clicks on an elemental technology in the figure, a list of research fields in which that technology has been used is shown. For example, if the user clicks on "FPGA" in Figure 1, a list of research fields for which "FPGA" is an elemental technology is displayed, as shown in Figure 2. From this list, we discover that "FPGA" was used in the electronic device field (integrated circuit for pattern recognition) in the 1998 and that this technology was used in the information and communications engineering field (speech analysis system) in the 2005.

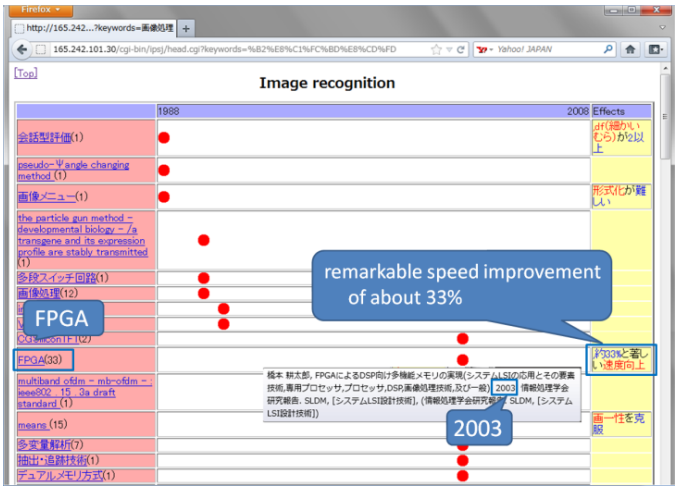

Figure 1. A list of elemental technologies used in the "Image recognition" field

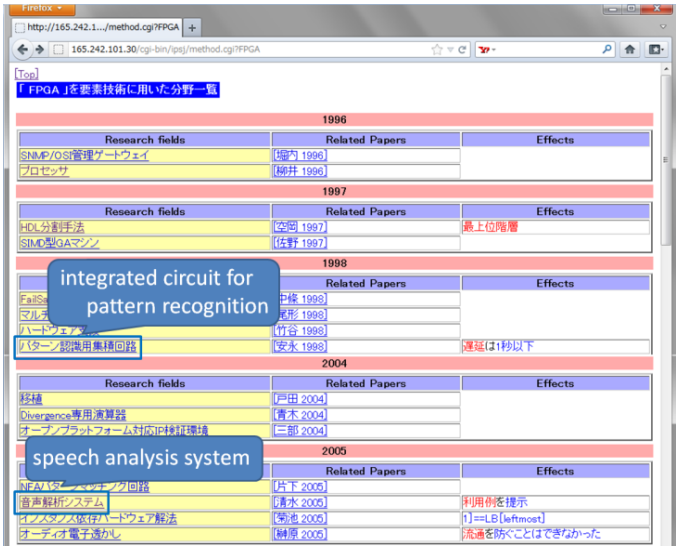

Figure 2. A list of research fields that use "FPGA" as an elemental technology

3. RELATED WORK

The interest in systems that analyze technical trends is very great. Kondo et al. [2] proposed a method that analyzes the structure of research paper titles using a machine-learning-based information extraction technique. They extracted elemental technologies from research paper titles in a particular field, and created a technical trend map by showing a history of such elemental technologies in the field.

The NTCIR-8 Patent Mining Task is another research project [3], which aims to create technical trend maps from research papers and patents. The following two steps are used to create a technical trend map.

1. For a given field, research papers and patents are collected.
2. Elemental technologies and their effects are extracted from the documents collected in step 1, and the documents are classified in terms of the elemental technologies and their effects.

For each of these steps, the following two subtasks were conducted.

1. Research Paper Classification: Classify research papers into the IPC system, a global standard hierarchical patent classification system.
2. Technical Trend Map Creation: Extract the expression of elemental technologies and their effects from research papers and patents.

We evaluated our method using the dataset used for the subtask of Technical Trend Map Creation.

We used Nanba's approach [4] for the basic framework of Technical Trend Map Creation subtask. Nanba et al. were one of participant groups in this subtask, and employed machine learning with several features based on cue phrases, which we will describe in Section 4.2. Although they obtained the best performance of all the participating groups, the recall score was low, due to the lack of cue phrases and the insufficiency of training data. We improved the recall score using the following two methods.

1. Using a unit list as an additional feature for machine learning
2. Applying domain adaptation techniques

Nishiyama et al. [5] also used a domain adaptation technique, FEDA [1], for the subtask of Technical Trend Map Creation, and reported its effects. We also examined FEDA, and confirm its effectiveness. In addition to FEDA, we propose some domain adaptation methods, and show that our methods are superior to FEDA.

4. AUTOMATIC CREATION OF THE TECHNICAL TREND MAP

4.1 Task Definition

To create a technical trend map, such as that show in Figures 1 and 2, we extract elemental technologies and their effects from research papers and patents using information extraction based on machine learning. We formulated the information extraction as a sequence-labeling problem, then analyzed and solved it using machine learning. The tag set is defined as follows.

- **TECHNOLOGY** includes algorithms, materials, tools, and data used in each study or invention.
- **EFFECT** includes pairs of ATTRIBUTE and VALUE tags.
- **ATTRIBUTE** and **VALUE** include effects of a technology that can be expressed by a pair comprising an attribute and a value.

A tagged example is given in Figure 3.

[original]
PM 磁束制御用コイルを設けて<TECHNOLOGY>閉ループフィードバック制御</TECHNOLOGY>を施すため、<EFFECT><ATTRIBUTE>電力損失</ATTRIBUTE>を<VALUE>最小化</VALUE></EFFECT>できる。
[translation]
Through <TECHNOLOGY>closed-loop feedback control</TECHNOLOGY>, the system could <EFFECT><VALUE>minimize</VALUE> the <ATTRIBUTE>power loss</ATTRIBUTE></EFFECT>.

Figure 3. A tagged example

4.2 Strategies for Extraction of Elemental Technologies and Their Effects

We used Nanba's approach [4] as a basic framework for extracting elemental technologies and their effects from research

papers and patents. As for the machine learning method used, Nanba investigated the Support Vector Machine (SVM) approach, which obtained the higher precision in comparison to Conditional Random Field (CRF) via pilot studies. The SVM-based method identifies the class (tag) of each word. The features and tags given by the SVM method are shown in Figure 4, and are as follows. The phrases of the technologies, effect attributes, and effect values are encoded in the IOB2 representation [6]. The bracketed numbers shown for each feature represent the number of cue phrases. They used window sizes k=3 and k=4 for research papers and patents, respectively, which were determined via a pilot study.

- A word.
- Its part of speech.
- ATTRIBUTE-internal (F1): Whether the word is frequently used in ATTRIBUTE tags; e.g., "精度(precision)". (1210)
- EFFECT-external (F2): Whether the word is frequently used before, or after the EFFECT tags; e.g., "できる (possible)". (21)
- TECHNOLOGY-external (F3): Whether the word is frequently used before, or after the TECHNOLOGY tags; e.g., "を用いた (using)". (45)
- TECHNOLOGY-internal (F4): Whether the word is frequently used in TECHNOLOGY tags; e.g., "HMM" and "SVM". (17)
- VALUE-internal (F5): Whether the word is frequently used in VALUE tags; e.g., "増加 (increase)". (408)
- Location (F6): Whether the word is contained in the first, the middle, or the last third of an abstract.

In addition to these features, we examine a unit list as another feature "F7" for machine learning. We will describe the details of this feature in the next section.

4.3 Creation of a Unit List

We created a unit list for VALUE tag annotation semi-automatically. Most nouns (counter suffix) immediately after numerical values, such as "100 cm" or "20 MB/s", are considered units. Therefore, we collected these nouns from the CiNii research paper corpus[1] automatically, then manually created a unit list from them. Finally, we obtained 274 unit terms, some of which are shown in Figure 5.

4.4 Domain Adaptation

In the Technical Trend Map Creation subtask, 300 research papers and 300 patents with manually assigned "TECHNOLOGY", "EFFECT", "ATTRIBUTE", and "VALUE" tags were prepared. For extracting elemental technologies and their effects from research papers, Nanba et al. [4] used 300 research papers as the training data, while Nishiyama et al. [5] used both 300 research papers and 300 patents by introducing a domain adaptation method, FEDA [1], and reported on the effectiveness. FEDA is a feature augmentation technique that simply adds features for the source and target domains to the original feature list. The augmented feature vector for the paper domain is $f_{paper}(x)=\langle f(x), f(x), \mathbf{0}\rangle$ and that for the patent domain is $f_{patent}(x)=\langle f(x), \mathbf{0}, f(x)\rangle$ where $\mathbf{0}=\langle 0, 0, \ldots, 0\rangle \in R^m$ is the zero vector. Then this augmented data in both domains is used for predictive modeling, and the weights of the shared features are estimated using the training data from both domains. We also examine FEDA, and confirm its effectiveness.

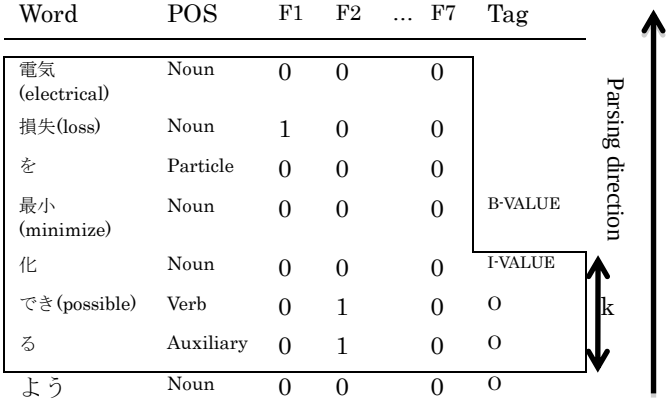

Figure 4. Features and tags given to the SVM

%	日(day)	歳(age)	回(times)	℃	wt	byte	kbyte	
cm	mg	mm	J	kg	bit	パーセント(percent)	sec	

Figure 5. An example of a sorted out unit term list

In addition to FEDA, we propose the following method.

Method 1: SEQ

1. Obtain model A using 300 research papers as training data.
2. Obtain model B using 300 patents as training data.
3. Annotate research papers with tags obtained from model A, then annotate the papers with additional tags obtained from model B.

In addition to this method, we propose another one. Generally, elemental technologies are written descriptively in patents. As a result, the average length of "TECHNOLOGY" tags in patents is much greater than that of research papers. This indicates that model B tends to annotate longer "TECHNOLOGY" tags, even though the target documents are research papers. To improve this problem, we propose the following method.

Method 2:SEQ(T)

1. Obtain model A using 300 research papers for training data.
2. Obtain model B' using 300 patents, whose "TECHNOLOGY" tags were preliminarily removed, for training data. In this step, features F3 and F4 are not used.
3. Annotate research papers with tags obtained from model A, then annotate the papers with additional tags obtained from model B'.

In case of patent analysis, models B or B' are first applied to patents in step 3.

5. EXPERIMENTS

To investigate the effectiveness of our method, we conducted some experiments. We describe the experimental methods and the results in Sections 5.1 and 5.2, respectively.

5.1 Experimental Methods

Datasets and experimental settings

We used the data for the Technical Trend Map Creation subtask of the patent mining task from the NTCIR-8 Workshop [3]. In this subtask, sets of the following documents with manually assigned "TECHNOLOGY", "EFFECT", "ATTRIBUTE", and "VALUE" tags were prepared.

- 500 Japanese research papers (abstracts).

[1] http://ci.nii.ac.jp

- 500 Japanese patents (abstracts).

For each type of document, 300 were provided as training data, with the remaining 200 being used as test data in the Patent Mining Task.

Evaluation

We used the following measures for evaluation.

$$Recall = \frac{The\ number\ of\ correctly\ extracted\ tags}{The\ number\ of\ tags\ \ that\ should\ be\ extracted}$$

$$Precision = \frac{The\ number\ of\ correctly\ extracted\ tags}{The\ number\ of\ tags\ that\ the\ system\ extracted}$$

$$F-measure = \frac{2\ \cdot Recall\ \cdot Precision}{Recall + Precision}$$

Alternative methods

We conducted experiments using the following methods.

Baseline Methods:

- **HCU** [4]: An SVM-based approach using eight features (word, its part of speech, and features F1 to F6).
- **TRL_7_1 & TRL_6_2** [5]: A CRF-based approach using several features (word, its part of speech, character type, word prefix type, sections in patents, relative position in research papers, IPC codes manually assigned to each abstract, evaluative phrase, phrase distance in dependency trees) along with domain adaptation technique FEDA [1].

Our Methods:

- **UNIT**: An SVM-based approach using features F1 to F7.
- **UNIT_FEDA**: An SVM-based approach using features F1 to F7. Both research papers and patents are used as training data by applying FEDA.
- **SEQ**: An SVM-based approach using features F1 to F7. (Method 1 in Section 4.4).
- **SEQ(T)**: An SVM-based approach using features F1, F2, F5, F6, and F7. (Method 2 in Section 4.4).

5.2 Experimental Results

The average scores of recall, precision, and F-measure for the analysis of research papers and patents are shown in Tables 1 and 2, respectively. As can be seen from Table 1, our method SEQ(T) significantly improved recall scores compared with the baseline methods used in research paper analysis. On the other hand, our methods did not improve on the baseline methods in patent analysis (Table 2), because the performance of the baseline systems is so high there was little room for improvement.

5.3 Discussion

Effectiveness of domain adaptation and a unit list

To investigate the effects of a unit list and domain adaptation methods, we calculated recall, precision, and F-measure of methods HCU, UNIT, and SEQ(T) for each tag. We found that recall scores of UNIT for ATTRIBUTE and VALUE were 1.3 to 1.7% higher than those of HCU, which recall scores of SEQ(T) for ATTRIBUTE and VALUE were 12.8 to 13.6% higher than those of HCU. These results indicate that a unit list is useful, but the contribution by our domain adaptation method is much greater than a unit list.

Comparison of domain adaptation methods

The recall score of UNIT_FEDA is higher than that of UNIT, and we could also confirm the effectiveness of FEDA, even though the features used in our method is differ from those in Nishiyama's method. However, we can conclude that our domain adaptation methods (SEQ(T) and SEQ) are more useful than FEDA in research paper analysis.

Table 1. Experimental results for research papers

	Recall	Precision	F-measure
HCU	0.184	0.686	0.290
TRL_7_1	0.181	0.573	0.275
UNIT	**0.191**	0.669	**0.298**
UNIT_FEDA	**0.211**	0.547	**0.305**
SEQ	**0.246**	0.411	**0.308**
SEQ(T)	**0.254**	0.496	**0.336**

Table 2. Experimental results for patents

	Recall	Precision	F-measure
HCU	0.441	0.537	0.485
TRL_6_2	0.437	0.506	0.469
UNIT	0.441	0.537	0.484
UNIT_FEDA	0.429	**0.540**	0.478
SEQ	**0.454**	0.493	0.473
SEQ(T)	**0.455**	0.507	0.480

6. CONCLUSIONS

In this paper, we have proposed a method that extracts elemental technologies and their effects from the abstracts of research papers and patents. From our experimental results, our method SEQ(T) obtained recall and precision scores of 0.254 and 0.496, respectively, for the analysis of research papers. The SEQ(T) method also obtained recall and precision scores of 0.455 and 0.507, respectively, for the analysis of patents. Therefore, we have constructed a system that creates an effective technical trend map for a given field.

7. REFERENCES

[1] Daumé, III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 256-263.

[2] Kondo, T., Nanba, H., Takezawa, T., and Okumura, M. 2009. Technical trend analysis by analyzing research papers' titles. In *Proceedings of the 4th Language & Technology Conference* (LTC'09), 234-238.

[3] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. 2010. Overview of the patent mining task at the NTCIR-8 workshop. In *Proceedings of the 8th NTCIR Workshop Meeting*.

[4] Nanba, H., Kondo, T., and Takezawa, T. 2010. Automatic creation of a technical trend map from research papers and patents. In *Proceedings of the 3rd International CIKM Workshop on Patent Information Retrieval (PalR'10)*, 11-15.

[5] Nishiyama, R., Tsuboi, Y., Unno, Y., and Takeuchi, H. 2010. Feature-rich information extraction for the technical trend map creation. In *Proceedings of the 8th NTCIR Workshop Meeting*.

[6] Tjong Kim Sang, E.F. and Veenstra, J. 1999. Representing text chunks. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics (EACL)*, 173-179.