Hiroshima City University at NTCIR-8 Patent Mining Task

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi
Hiroshima 731-3194 JAPAN

Tomoki Kondo
Hiroshima City University
3-4-1 Ozukahigashi
Hiroshima 731-3194 JAPAN

Toshiyuki Takezawa
Hiroshima City University
3-4-1 Ozukahigashi
Hiroshima 731-3194 JAPAN

ABSTRACT
Our group participated in the subtask of technical trend map creation for the NTCIR-8 Patent Mining Task. We prepared five types of cue phrase list using statistical methods, and used them in the analysis of research papers and patents based on the Support Vector Machines. From the experimental results, we obtained Recall of 0.110 and Precision of 0.424 for research papers, and Recall of 0.430 and Precision of 0.563 for patents.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**: Search process

H.3.4 [**Systems and Software**]: Performance evaluation

H.3.5 [**Online Information Services**]: Data sharing

General Terms
Measurement, Performance, Experimentation

Keywords
Information extraction, SVM, distributional similarity

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

For a researcher in a field of high industrial relevance, retrieving and analyzing both research papers and patents have become an important aspect of assessing the scope of the field. Such fields include bioscience, medical science, computer science, and materials science. 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 and research papers, which could invalidate a rival company's patents or patents under application in a Patent Office. Therefore, we participated in the subtask of technical trend map creation form the NTCIR-8 Patent Mining Task to develop techniques for analyzing both research papers and patents.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 explains our method for analyzing the structure of research papers and patents. To investigate the effectiveness of our method, we conducted some experiments. Section 4 reports on these experiments, and discusses the results. We present some conclusions in Section 5.

2. RELATED WORK
Recently, many researchers have studied the automatic generation of survey articles from a set of research papers in a particular research field [8,1,11,13]. Our present task may be considered a type of multi-paper summarization, expressed in terms of elemental technologies and their effects, although our method generates technical trend maps instead of summary documents.

The interest in systems that analyze technical trends is very high. However, few systems are actually in use. Aureka[1] from Thomson Reuters is one such system. Aureka is fundamentally a patent analysis system. One of its functions is to express quotation relations as a tree. Alternatively, they can be displayed in an aerial view, called a ThemeScape map, which relates the patent to a given patent set. The importing of paper data in various formats, such as PDF and MS Word, is possible with this system. In this way, a paper can be mapped and analyzed via the ThemeScape map for a patent.

3. AUTOMATIC CREATION OF TECHNICAL TREND MAPS
3.1 Tag Definition
We used information extraction based on machine learning to extract information such as the elemental technologies and effects from research papers and patents. 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, tools, materials, and data used in each study or invention.
- **EFFECT** includes pairs of ATTRIBUTE and VALUE tags.
- **ATTRIBUTE** and **VALUE** includes effects of a technology that can be expressed by a pair comprising an attribute and a value.

3.2 Strategies for Creating Cue Phrase Lists
We investigated randomly selected research papers and patents, seeking useful cues for the automatic assignment of TECHNOLOGY, ATTRIBUTE, and VALUE tags, and found the following three features of cues.

1. Noun phrases before particular phrases, such as "を用いた (using)" or "を具備する (equipped)", tend to be assigned a TECHNOLOGY tag. There are few such phrases, and the phrases are domain independent[5].
2. Particular phrases, such as "信頼性 (credibility)" or "精度 (precision)", tend to be assigned an ATTRIBUTE tag. There are many such phrases, and they differ according to their domains. For example, "稼働率 (capacity operating rate)" or "駆動周波数 (drive frequency)" tend to be used in one particular domain.

[1] http://science.thomsonreuters.com/training/aureka/

3. Particular words, such as "改善 (improvement)" or "高速化 (speeding up)", tend to be assigned a VALUE tag. There are many such phrases. Although some of these phrases are domain independent, there are many phrases, such as "平滑化 (smoothing)", which tend to be used in particular domains.

From the results of this investigation, we employed the following strategy for creating cue phrase lists.

- Manually create a cue phrase list for a TECHNOLOGY tag.
- Create cue phrase lists for ATTRIBUTE and VALUE tags semi-automatically.

In the next section, we describe how to create cue phrase lists for ATTRIBUTE and VALUE tags.

3.3 Creating Cue Phrase Lists

We created cue phrase lists for ATTRIBUTE and VALUE tags using the following three steps.

- (Step 1) Collect cue phrases for a VALUE tag using patterns.
- (Step 2) Collect cue phrases for an ATTRIBUTE tag using dependency parsing.
- (Step 3) Collect cue phrases for ATTRIBUTE and VALUE tags using distributional similarity.

In the following, we describe the details of each step.

(Step 1) Collect cue phrases for a VALUE tag using patterns

Nanba[10] extracted hypernym/hyponym relations for words (or phrases) from Japanese patent applications using a set of patterns, such as "NP_1 (や|と|,) NP_2 (等の|などの) NP_0 (NP_0, such as NP_1, NP_2, (and/or) NP_n)", which was originally devised by Hearst[2] for English text corpora. By using "効果 (effect)" or "特徴 (feature)" instead of NP_0 in the above pattern, we can collect cue phrases for a VALUE tag from research papers and patents. For example, we can extract "軽減 (reduction)" from the following sentence using the pattern:

> ...炉壁熱負荷の**軽減**等の効果が得られる。
>
> (..obtain an effect, such as **reduction** of heat load of furnace wall.)

We applied this method to 255,960 research papers' abstracts, which were used at the first and second NTCIR Workshops[3,4], and Japanese patent applications published in the ten-years period 1993-2002, and obtained a set of candidate cue phrases. Then we manually eliminated inappropriate phrases from the candidates, finally obtaining 300 cue phrases for a VALUE tag.

(Step 2) Collect cue phrases for an ATTRIBUTE tag using dependency parsing

Many noun phrases that have dependency relations with the cue phrases for a VALUE tag obtained in Step 1 are cue phrases for an ATTRIBUTE tag. Therefore, we applied the Japanese syntactic parser CaboCha[2] to the research papers' abstracts and the Japanese patent applications to obtain a set of candidate cue phrases. Then we manually eliminated inappropriate phrases from the candidates, obtaining 700 cue phrases for an ATTRIBUTE tag.

(Step 3) Collect cue phrases for ATTRIBUTE and VALUE tags using distributional similarity

Lin[7] and Lee[6] proposed a method for calculating the similarity between terms, which they called "distributional similarity". The underlying assumption of their approach is that semantically similar words are used in similar contexts. They therefore defined the similarity between two terms as the amount of information contained in the commonality of the terms, divided by the amount of information in the contexts of the terms. In our work, we use "distributional similarity" as a method for acquiring cue phrases for ATTRIBUTE and VALUE tags via the following procedure.

1. Analyze the dependency structures of approximately 600 million sentences in Japanese patent applications over a ten-year period, using the Japanese parser CaboCha.
2. Extract noun phrase-verb 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. Calculate the similarities between two indices for nouns using the SMART similarity measure[12].
6. Obtain a list of pairs of related noun phrases.
7. For each phrase in the cue phrase lists for ATTRIBUTE and VALUE tags, obtain its counterpart in the list obtained in the previous step as a new cue phrase.

3.4 Features used in Machine Learning

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

For the machine learning method, we investigated the Support Vector Machine (SVM) approach. The SVM-based method identifies the class (tag) of each word. The features and tags given by the SVM method are shown in Figure 1. The numbers shown together in each feature are the number of cue phrases. We used values of k=3 and k=4 for research papers and patents, respectively, which were determined from a pilot study.

- A word.
- Its part of speech[3].
- ATTRIBUTE-internal (F1): Whether the word is frequently used in ATTRIBUTE tags, e.g., "処理量 (throughput)" or "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)" and "実現する (realize)". (21)
- TECHNOLOGY-external (F3): Whether the word is frequently used before, or after the TECHNOLOGY tags, e.g., "を用いた (using)" and "に基づいた (based on)". (45)

[2] http://chasen.org/~taku/software/cabocha/

[3] We used MeCab as a Japanese morphological analysis tool. (http://mecab.sourceforge.net)

Word	POS	F1	F2	F3	F4	F5		Tag
電気 (electrical)	Noun	0	0	0	0	0	0	
損失 (loss)	Noun	1	0	0	0	0	0	
を	Particle	0	0	0	0	0	0	
最小 (minimize)	Noun	0	0	0	0	0	0	B-VALUE
化	Noun	0	0	0	0	1	0	I-VALUE
でき (possible)	Verb	0	1	0	0	0	0	O
る	Auxiliary Verb	0	1	0	0	0	0	O
よう	Noun	0	0	0	0	0	0	O
に	Particle	0	0	0	0	0	0	O
なる	Verb	0	0	0	0	0	0	O

target

k

parsing direction

Figure 1. Features and tags given to the SVM

- 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)" and "抑止 (determent)". (408)
- Location (F6): Whether the word is within the first, the middle, or the last third of an abstract[4].

4. EXPERIMENTS

To investigate the effectiveness of our method, we conducted some experiments. For the formal run of the Japanese subtask, we submitted "HCU". We describe the experimental methods and the results in Sections 5.1 and 5.2, respectively.

4.1 Experimental Methods

Data sets and experimental settings

We used the data for the Patent Mining Task at the NTCIR-8 Workshop[9]. In this task, sets of the following documents with manually assigned "TECHNOLOGY", "EFFECT", "ATTRIBUTE", and "VALUE" tags were prepared.

- 500 Japanese research papers (abstracts)
- 500 Japanese patents (abstracts)[5]

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

Evaluation

We used the following measures for evaluation.

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

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

4.2 Experimental Results

The evaluation results for the analysis of research papers and patents are shown in Tables 1 and 2, respectively.

Table 1. Experimental results for research papers

	Recall	Precision
TECHNOLOGY (Title)	0.656	0.656
TECHNOLOGY (Abstract)	0.131	0.495
ATTRIBUTE	0.095	0.394
VALUE	0.105	0.383
EFFECT	0.061	0.310
Average	0.160	0.491

Table 2. Experimental results for patents

	Recall	Precision
TECHNOLOGY (Title)	0.556	0.455
TECHNOLOGY (Abstract)	0.439	0.490
ATTRIBUTE	0.371	0.544
VALUE	0.481	0.655
EFFECT	0.268	0.409
Average	0.431	0.545

[4] Generally, the purpose of the research paper, the elemental technologies, and the effects are written in the first, the middle, and the last third of an abstract, respectively. In contrast, the patent abstracts used in our work comprise three fields, namely "technical problem", "the means for solving the technical problem", and "the effects of the invention". We regard these fields as the first, the middle, and the last third of a patent abstract.

[5] Tags were assigned to the fields of "technical problem", "the means for solving a technical problem", and "effect of the invention" in each abstract.

4.3 Discussions

4.3.1 Typical Errors in the Analysis of Research Papers

There were two typical errors in the analysis of research papers: (1) effects of ambiguous expressions "の (of)" and "による (by)" for ATTRIBUTE tag assignment (14%) and (2) lack of TECHNOLOGY-internal cue phrases (13%). We describe these errors as follows.

(1) Effects of ambiguous expressions "の (of)" and "による (by)" for ATTRIBUTE tag assignment

For an expression "指向性の影響を低減 (reduction of an effect of directionality)", ATTRIBUTE and VALUE tags should be assigned to "指向性の影響 (an effect of directionality)" and "低減 (reduction)", respectively, but our method could not assign any tags to this expression. The expression "の (of)" is often used between ATTRIBUTE and VALUE tags, but it is sometimes used within the ATTRIBUTE tag. In addition to this, both "低減 (reduction)" and "影響 (effect)" are contained in VALUE-internal cues. In this case, there are three possibilities as follows, and our system selected the third one.

1. Assign ATTRIBUTE and VALUE tags to "指向性の影響 (an effect of directionality)" and "低減 (reduction)", respectively.
2. Assign ATTRIBUTE and VALUE tags to "指向性 (directionality)" and "影響 (an effect)", respectively.
3. Assign no tags to this expression.

(2) Lack of TECHNOLOGY-internal cues (13%)

For an expression "SAW 素子を用いた (using SAW element)", our method could not assign the TECHNOLOGY tag to "SAW 素子 (SAW element)", because "SAW 素子 (SAW element)" is not contained in the TECHNOLOGY-internal cues.

4.3.2 Typical Errors in the Analysis of Patents

There were three typical errors in the analysis of patents: (1) patent-specific expressions (33%), (2) effects of ambiguous expressions "の (of)" and "による (by)" for ATTRIBUTE tag assignment (7%) and (3) order of ATTRIBUTE and VALUE tags (7%). We describe errors (1) and (3) as follows.

(1) Patent-specific expressions

Elemental technologies are often expressed with longer or multiple noun phrases in patents. Typical patterns are "[elemental technology A]と、[elemental technology B]と、[elemental technology C]とを設け (comprising [elemental technology A], [elemental technology B], and [elemental technology C])", and our method uses cues, such as "と、 (, and)", for the TECHNOLOGY tag assignment. However, the expression "と、 (, and)" is also used except for listing elemental technologies. Even in such cases, our method mistakenly assigns the TECHNOLOGY tag.

(3) Order of ATTRIBUTE and VALUE tags

For an expression "高い認識率 (high recognition rate)", ATTRIBUTE and VALUE tags should be assigned to "認識率 (recognition rate)" and "高い (high)", respectively, but our system did not assign any tags to this expression. Most of the order of these two tags in the training data was "ATTRIBUTE -> VALUE". As a result, our system could not assign any tags if an expression, in which the ATTRIBUTE tag should be assigned, appears just after an expression, in which the VALUE tag should be assigned.

5. CONCLUSION

In this paper, we proposed a method that extracts elemental technologies, and their effects from research papers' abstracts and patents. From the experimental results, we obtained Recall and Precision scores of 0.110 and 0.424, respectively, for the analysis of research papers. We also obtained Recall and Precision scores of 0.430 and 0.563, respectively, for the analysis of patents.

6. REFERENCES

[1] Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, E. 2007. Blind men and Elephants: What do Citation Summaries Tell Us about a Research Article? Journal of the American Society for Information Science and Technology, 59 (1), pp.51-62.

[2] Hearst, M.A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545.

[3] 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.

[4] 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.

[5] Kondo, T., Nanba, H., Takezawa, T., and Okumura, M. 2009. Technical Trend Analysis by Analyzing Research Papers' Titles, Proceedings of the 4th Language & Technology Conference (LTC'09), pp.234-238.

[6] Lee, L. 1999. Measures of Distributional Similarity. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.25-32.

[7] Lin, D. 1998. Automatic Retrieval and Clustering of Similar Words, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp.768-774.

[8] Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., and Zajic, D. 2009. Generating Surveys of Scientific Paradigms, Proceedings of HLT-NAACL 2009, Boulder, CO.

[9] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. 2010. Overview of the Patent Mining Task at the NTCIR-8 Workshop Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access.

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

[11] Nanba, H. and Okumura, M. 1999. Towards Multi-paper Summarization Using Reference Information, Proceedings of the 16th International Joint Conferences on Artificial Intelligence, pp.926-931.

[12] Salton, G. 1971. The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ.

[13] Teufel, S. and Moens, M. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, Computational Linguistics: 28 (4), pp.409-445.