Automatic Creation of a Technical Trend Map from Research Papers and Patents

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
For a researcher in a field of great industrial relevance, retrieving and analyzing research papers and patents has 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 essential 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 0.160 and 0.491, respectively, for the analysis of research papers. We also obtained Recall and Precision scores of 0.431 and 0.545, 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 [**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 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 of great industrial relevance, retrieving and analyzing research papers and patents have become important aspects of assessing the scope of the field. Such fields include bioscience, medical science, computer science, and materials science. However, it is costly and time-consuming to collect and read all of the papers in the field. Therefore, we can see 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 illustrates 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. To investigate the effectiveness of our method, we conducted some experiments. 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 for visualizing technical trends. Figure 1 shows the technical trend map after the research field "speech recognition" was given to the system. In this figure, several elemental technologies used in the speech recognition field, such as "HMM" (Hidden Markov Model), are listed in the left-hand column. The effects of each technology, such as "精度が向上 (increase precision)", are shown in the right-hand column. These technologies and effects were extracted automatically from research papers and patents in this field, and each 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 "HMM" in Figure 1, a list of research fields for which "HMM" is an elemental technology is displayed, as shown in Figure 2. From this list, we discover that "HMM" was used in an image-recognition field (place-name recognition) in the early 1990s and that this technology was used in motion-image sequence-analysis field (gesture recognition) in the late 1990s.

3. RELATED WORK
Recently, many researchers have studied the automatic generation of survey articles from a set of research papers in a particular research field [2, 14]. 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.

PaIR'10, October 26, 2010, Toronto, Ontario, Canada.

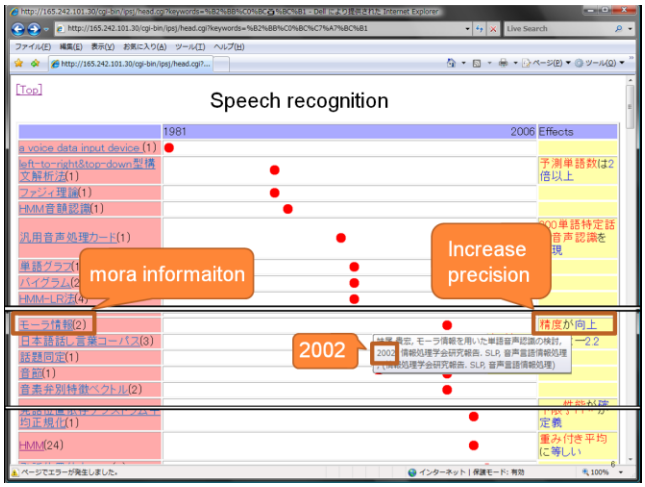

Figure 1. A list of elemental technologies used in the "speech recognition field"

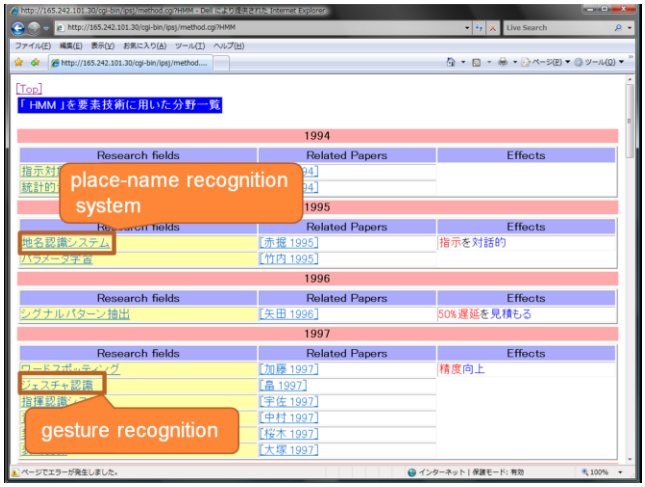

Figure 2. A list of research fields that uses "HMM" as an elemental technology

Kondo et al. [3] proposed a method that analyzes the structure of research papers' titles using a machine-learning-based information extraction technique. They extracted elemental technologies from research papers' titles in a particular field, and created a technical trend map by showing a history of such elemental technologies in the field. In addition to the elemental technologies themselves, we extracted their effects from research papers' and patents' abstracts using the data of the NTCIR-8 Patent Mining Task [7].

4. AUTOMATIC CREATION OF TECHNICAL TREND MAPS

4.1 System Overview

To create a technical trend map, such as that illustrated in Figures 1 and 2, the following two steps are required.

(Step 1) For a given field, research papers and patents are collected.

(Step 2) Elemental technologies and their effects are extracted from the documents collected in Step 1.

For Step 1, we used Nanba's system for collecting both research papers and patents in a particular field [8]. In this paper, we focus on Step 2. In the following, we describe the details of our approach to this step.

4.2 Tag Definition

We used information extraction based on machine learning to extract information such as the elemental technologies and their 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** 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.

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。
(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

4.3 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 [3].
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.
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.

4.4 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 [9] 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)". 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, and Japanese patent applications published in the 10-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 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

We use "distributional similarity" [4, 5] 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 10-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.
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.

4.5 Features Used in Machine Learning

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 4, and listed below. The phrases of the technologies, effect attributes, and effect values are encoded in the IOC2 representation [13] as shown in Figure 4. The bracketed numbers shown for each feature represent the number of cue phrases. We used values of 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 within the first, the middle, or the last third of an abstract.

5. 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.

5.1 Experimental Methods

Data sets and experimental settings

We used the data for the Patent Mining Task at the NTCIR-8 Workshop [7]. 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)

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.

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

Word	POS	F1	F2	F3	F4	F5	F6	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

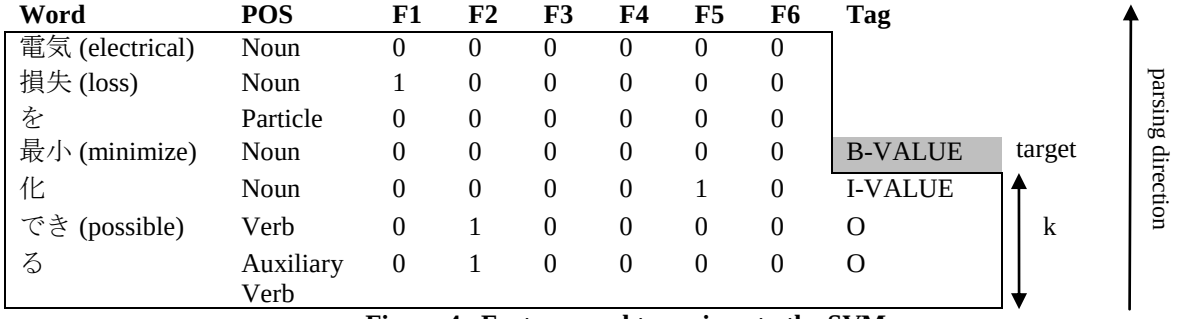

Figure 4. Features and tags given to the SVM

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

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

Alternative systems

We compared the following five systems, which were submitted to the formal run of the subtask of technical trend map creation in NTCIR-8 Patent Mining Task.

- **HCU** (our method)
- **TRL7_1** & **TRL6_2** [10]: A CRF-based approach using several features (word, its part of speech, character type, word prefix type, word suffix type, sections in patents, relative position in research papers, IPC codes manually assigned to each abstract, evaluative phrase, phrase distance in dependency trees) with domain adaptation technique FEDA [1].
- **ONT** [6]: An SVM-based approach using several features (word, its part of speech, original form of word, semantic label from the results of language analysis). Prior to machine learning, training data were divided into multiple clusters. The SVM was applied to each cluster.
- **smlab** [12]: An SVM-based approach using cue phrases such as "用い (using)" and "備え (possessing)", which were collected by using entropy-based scores.
- **HTC_1** & **HTC_1_1** [11]: Phrase extraction based on 3-tuple expressions using SVM with several features (word, part of speech, manually created cue phrase lists from "effect of the invention" fields in patent, modification relations using a Japanese dependency parser).

5.2 Experimental Results

The evaluation results for the analysis of research papers and patents are shown in Tables 1 and 2, respectively. We also show the average values of Recall, Precision, and F-measure for the five systems in Tables 3 and 4.

Table 1. Experimental results for research papers

	Recall	Precision	F-measure
TECHNOLOGY (Title)	0.656	0.656	0.656
TECHNOLOGY (Abstract)	0.131	0.495	0.206
ATTRIBUTE	0.095	0.394	0.153
VALUE	0.105	0.383	0.165
EFFECT	0.061	0.310	0.103
Average	0.160	0.491	0.241

Table 2. Experimental results for patents

	Recall	Precision	F-measure
TECHNOLOGY (Title)	0.556	0.455	0.500
TECHNOLOGY (Abstract)	0.439	0.490	0.463
ATTRIBUTE	0.371	0.544	0.440
VALUE	0.481	0.655	0.555
EFFECT	0.268	0.409	0.324
Average	0.431	0.545	0.481

Table 3. Comparison of systems for research papers (average)

	Recall	Precision	F-measure
TRL7_1	0.181	0.573	0.275
HCU (our method)	0.160	0.491	0.241
ONT	0.114	0.246	0.156
Smlab	0.081	0.354	0.132
HTC_1	0.100	0.188	0.131

Table 4. Comparison of systems for patents (average)

	Recall	Precision	F-measure
HCU (our method)	0.431	0.545	0.481
TRL_6_2	0.437	0.506	0.469
Smlab	0.272	0.547	0.363
HTC_1_1	0.233	0.346	0.278
ONT	0.178	0.271	0.215

5.3 Discussion

5.3.1 Typical Errors in the Analysis of Research Papers

There were two typical errors in the analysis of research papers. There were (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%). Among these errors, we describe error (1) for patents as follows.

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.

5.3.2 Typical Errors in the Analysis of Patents

There were three typical errors in the analysis of patents. There were (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%). Among these errors, we describe error (1) for patents as follows.

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.

5.3.3 Comparison with Other Systems

As the average sentence length of patents is longer than that of research papers, taking a wider context into account is required for precise information extraction from patents. To address this problem, Nishiyama et al. (in the TRL system) used a dependency-structure feature and confirmed its effectiveness [10]. Instead of a dependency-structure feature, we used a wider window for patents (k=4), as described in Section 4.5. As shown in Table 4, we can confirm the wider window is also effective for patents, because our method outperformed the TRL system.

On the other hand, our method performed worse than the TRL system for research papers (Table 3). As the training data for research papers, we used only 300 tagged research papers, while Nishiyama used 300 tagged patents in addition to 300 tagged research papers. In Nishiyama’s method, some features for the machine learning were used only for research papers, while some were used only for patents. To address this imbalance of features between research papers and patents, they employed a domain adaptation approach called FEDA [1]. FEDA is a feature augmentation technique that simply adds features for the source (patents) and target domains (research papers) into the original features. Even if the prospective prediction rules are different for the patent and paper domains, the weights of these augmented features will be learned correctly for each domain via FEDA. Just as for Nishiyama’s method, FEDA may improve the performance of our method for research papers.

6. CONCLUSION

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, we obtained Recall and Precision scores of 0.160 and 0.491, respectively, for the analysis of research papers. We also obtained Recall and Precision scores of 0.431 and 0.545, 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, H. 2007. Frustratingly Easy Domain Adaptation, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp.256-263.

[2] 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.

[3] 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.

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

[5] 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.

[6] Mizuguchi, H. and Kusui D. 2010. An Information Extraction Method for Multiple Data Sources, Proceedings of the 8th NTCIR Workshop Meeting, pp.348-353.

[7] 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, pp.293-302.

[8] Nanba, H. 2008. Hiroshima City University at NTCIR-7 Patent Mining Task, Proceedings of the 7th NTCIR Workshop Meeting, pp.369-372.

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

[10] Nishiyama, R., Tsuboi, Y., Unno, Y. and Takeuchi, H. 2010. Feature-Rich Information Extraction for the Technical Trend-Map Creation, Proceedings of the 8th NTCIR Workshop Meeting, pp.318-324.

[11] Sato, Y. and Iwayama, M. 2010. Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi, Proceedings of the 8th NTCIR Workshop Meeting, pp.359-363.

[12] Suzuki, Y., Nonaka, H., Sakaji, H., Kobayashi, A., Sakai, H. and Masuyama, S. 2010. NTCIR-8 Patent Mining Task at Toyohashi University of Technology, Proceedings of the 8th NTCIR Workshop Meeting, pp.364-369.

[13] Tjong Kim Sang, E.J. and Veenstra, J. 1999. Representing Text ChunksChunks. Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics, pp.173–179.

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