International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2020, Vol. 4, No. 2, 19 – 38

Extracting Characteristic Terms from Patent Documents

Kaito Takano [*], Miryu Tanaka [*],
Hiroyuki Sakai [*], Ryozo Kitajima [†],
Takahisa Ota [‡], Chinatsu Tanabe [‡], Hiroki Sakaji [§]

Abstract

We propose a method to automatically extract "characteristic terms" from several patent documents belonging to a certain technical field. The characteristic terms extracted by the proposed method are useful for technology trend analysis. For example, in the case of patent documents relating to organic electroluminescence, the characteristic term expresses the characteristic of the technology shown in the patents, e.g., "発光効率" (*hakkoukouritu*: luminous efficiency) or "輝度" (*kido*: brightness). The proposed method was evaluated and good results were obtained.

Keywords: characteristic terms, natural language processing, patent documents, text mining.

1 Introduction

When formulating future growth strategies, it is important for companies to analyze patents and predict customer demand and technological advancements. However, according to statistics issued by the patent office, more than 300,000 patent applications have been submitted each year over the past 10 years (2008 to 2017). Analyzing such an enormous number of patent documents requires a significant manual labor: therefore, interest in tools that assist patent analysis has increased, and various patent analysis methods have been developed to analyze large quantities of patent documents automatically[1][2][3].

One patent analysis method uses key phrases extracted from patent documents to facilitate analyze. For example, Hong et al. proposed a trend analysis method that uses named entities extracted from patent documents [4], and Okamoto et al. proposed a method for structural analysis of claims using named entities extracted from claims and dependency parsing [5]. However, in technical trend analysis, it is not necessarily appropriate to use a named entity as a key phrase. Generally, in the key phrase extraction method for patent analysis, there is a tendency to extract terms that are specific to a patent as key phrases [6]; thus, the extracted terms may not represent technical characteristics.

[*] Seikei University, Tokyo, Japan
[†] Tokyo Polytechnic University, Kanagawa, Japan
[‡] Showa Denko K.K., Tokyo, Japan
[§] The University of Tokyo, Tokyo, Japan

In this paper, we propose a method to automatically extract “characteristic terms” as key phrases suitable for technology trend analysis from patent documents belonging to a certain technical field. Here, a characteristic term represents the characteristics of the technology described in the patent. For example, in the case of a patent relating to “有機EL” (*yuuki EL*: organic EL), it is a term expressing the characteristics of the technology shown in the patent, e.g., “発光効率” (*hakkou kouritu*: luminous efficiency)or “輝度” (*kido*: brightness).

In other words, characteristic terms are specific to the technical field to which patents belong rather than words unique to certain patents. Therefore, the characteristic terms are suitable for comparing and analyzing multiple patents. As a research study focused solely on extracting key phrases suitable for patent analysis, Suzuki et al. proposed a method to extract words related to the novelty or inventive step of a patent from claims [6]. In contrast, the proposed method extracts characteristic terms, and the extraction targets differ.

Research in the field of chemistry is very active, new substances are developed every day, and new terms are born. A database of chemical terms is an important language resource in research, but at present, experts are manually deciphering articles and patents to construct a database. This work is very costly and time consuming, so there are many studies aimed at automating this work [7]. There are several studies that extract chemical terms by named entity recognition [8]. In particular, Ling et al. propose a method extracting chemical terms using BiLSTM-CRF approach and shows high performance [9]. In addition to terms, it is necessary to record the characteristics of new substances in the database, and the extraction of characteristic terms in our research is useful in such situations.

2 Proposed Method

To extract characteristic terms, we must first extract sentences that contain the characteristic terms from patent documents. Here, we define sentences that contain characteristic terms as “characteristic sentences.” We employ a deep learning model to extract characteristic sentences. To extract characteristic sentences with high accuracy, a large amount of training data is required; however, it is difficult to manually generate large amounts of training data to extract characteristic sentences containing characteristic terms, i.e., technical terms. Therefore, the proposed method generates training data automatically.

Next, the proposed method extracts characteristic term candidates from characteristic sentences using word2vec [10]. The characteristic term candidates are extracted by inputting words with high probability as characteristic terms (words with “性” (*sei*: characteristic)) to a word2vec model generated from patent documents.

Finally, the proposed method extracts characteristic terms from characteristic term candidates for each patent applicant.

The process of the proposed method is summarized follows.

Step 1: The method extracts characteristic sentences that contain characteristic terms from patent documents using deep learning.

Step 2: The method generates a word2vec model using patent documents and extracts characteristic term candidates from characteristic sentences using the generated word2vec model.

Step 3: The method extracts characteristic terms from characteristic term candidates for each patent applicant.

Figure 1 shows the general flow of the proposed method from input of patent documents to extraction of characteristic terms.

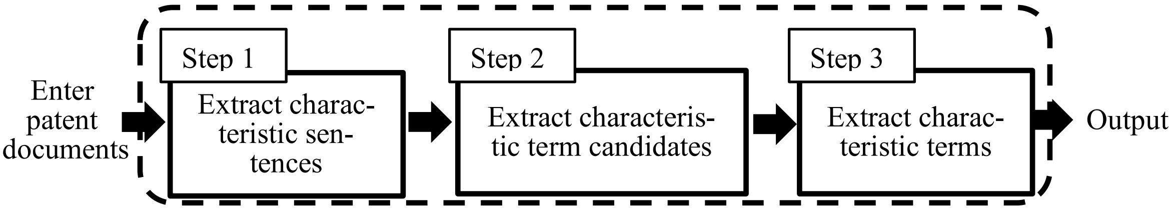

Enter patent documents
Step 1
Extract characteristic sentences
Step 2
Extract characteristic term candidates
Step 3
Extract characteristic terms
Output

Figure 1: Flow of the proposed method

2.1 Step 1: extraction of characteristic sentences

2.1.1 Automatic generation of training data

We use a deep learning model [11] to extract characteristic sentences from patent documents. Here, the training data used to construct the model are generated automatically. The proposed method generates training data form patent documents on “放熱材料” (*hounetu zairyou*: heat-dissipating material) and “ 有機 EL” (*yuuki EL*: organic EL). The method to automatically generate training data is summarized as follows.

Step 1.1: The method extracts sentences that contain phrases that often appear in characteristic sentences from sentences included in patent documents. These phrases are referred to as “clue phrases.”

Step 1.2: The method generates training data by dividing the extracted sentences into sentences that contain characteristic terms as positive examples and sentences that do not contain characteristic terms as negative examples. Here, we manually prepared characteristic terms in advance from patent documents on “放熱材料” (*hounetu zairyou*: heat-dissipating material)and “有機 EL” (*yuuki EL*: organic EL).

We use 57 phrases as clue phrases in Step 1, including 36 positive phrases described in Sakai et al. [3] (e.g., “できる” (*dekiru*: can)) and 21 negative phrases (e.g., “防止する” (*bousi suru*: curb)). Examples of positive phrase are given as follows.

できるので(*dekirunode*), できるとともに(*dekirutotomoni*), できて(*dekite*), 可能になり(*kanouninari*), 可能と(*kanouto*), できるようになる(*dekiruyouninaru*), できる(*dekiru*), できるようになった(*dekiruyouninatta*), 可能な(*kanouna*), でき(*deki*), 可能である(*kanoudearu*), 可能になる(*kanouninaru*)

Examples of negative phrases are given as follows.

防ぐ(*husegu*: prevent), 防止する(*bousisuru*: prevent), 軽減する(*keigensuru*: reduce), 緩和する(*kanwasuru*: ease), 和らげ(*yawarage*: soften), 鎮め(*sizume*: calming), 抑制する(*yokuseisuru*: suppress), 節約する(*setuyakusuru*: save), 削減する(*sakugensuru*: reduce), 低減する(*teigensuru*: reduce), 予防する(*yobousuru*: anisotropy)

In Step 1.2, we use characteristic terms prepared by chemistry experts. The proposed method uses 93 characteristic terms, including 37 characteristic terms of "放熱材料" (*hounetu zairyou*: heat-dissipating material) and 60 characteristic terms of "有機 EL" (*yuuki EL*: organic EL). Examples of characteristic terms for "放熱材料" (*hounetu zairyou*: heat-dissipating material) are given as follows.

異方性(*ihousei*: anisotropy), 加工性(*kakousei*: workability), 成形性(*seikeisei*: formability), 硬度(*koudo*: hardness), 強度(*kyoudo*: strength), 高分離能(*kou bunrinou*: high resolution), 耐腐食性(*tai husyokusei*: corrosion resistance), 耐食性(*taisyokusei*: corrosion resistance), 熱伝導性(*netu dendousei*: thermal conductivity), 意匠性(*iyhousei*: creativity), 機密性(*kimitusei*: confidentiality), 金属光沢(*kinzoku koutaku*: metallic luster), 軽量性(*keiryousei*: Lightness), 柔軟性(*jyuunansei*: flexibility), 潤滑性(*jyunkatusei*: lubricity), 信頼性(*sinraisei*: reliability), 絶縁性(*zetuensei*: insulation), 耐熱性(*tainetusei*: heat resistance), 耐摩耗性(*taimamousei*: wear resistance), 耐溶剤性(*taiyouzaisei*: solvent resistance), 耐薬品性(*taiyakuhinsei*: chemical resistance), 弾力性(*danryokusei*: elasticity), 断熱性(*dannetusei*: thermal insulation properties), 通気性(*tuukisei*: breathability), 電気伝導性(*denki dendousei*: electrical conductivity), 難燃性(*nannensei*: flame retardance), 発熱性(*hatunetusei*: fever), 放熱性(*hounetusei*: heat dissipation), 密着性(*mittyakusei*: adhesion)

Examples of characteristic terms for "有機 EL" (*yuuki EL*: organic EL) are given as follows.

発光効率(*hakkou kouritu*: luminous efficiency), 輝度(*kido*: brightness), 耐久性(*taikyuusei*: durability), 効率(*kouritu*: efficiency), 抑制(*yokusei*: suppression), 透明(*toumei*: clear), 高輝度(*kou kido*: high brightness), 安定性(*anteisei*: stability), 高効率(*kou kouritu*: high efficiency), 劣化(*rekka*: deterioration), 低電圧(*tei denatu*: low voltage), 均一(*kinitu*: uniform), 強度(*kyoudo*: Strength), 耐熱性(*tainetusei*: heat resistance), 導電性(*dendousei*: conductivity), 高発光効率(*kou hakkou kouritu*: high luminous efficiency), 信頼性(*sinraisei*: reliability), 反射(*hansya*: reflection), 屈折率(*kusseuturitu*: refractive index), 長寿命化(*tyou jumyouka*: life prolongation), 生産性(*seisansei*: productivity), 輸送性(*yusousei*: transportability), 消費電力(*syouhi denryoku*: power consumption), 高温(*kouon*: high temperature), 透明性(*toumeisei*: transparency), 低コスト(*tei kosuto*: low cost), 均一性(*kinitusei*: Uniformity), 透光性(*toukousei*: translucency), ガスバリア性(*gasubariasei*: gas barrier), 欠陥(*kekkan*: defect), 透過率(*toukaritu*: transmittance), 移動度(*idoudo*: mobility), 隔壁(*kakuheki*: partition)

Using the above method, training data containing 173,814 sentences were generated.

2.1.2 Feature selection

Next, we select the features used to generate the deep learning model. Here, the proposed method assigns weight $W_p(t, S_p)$ to content words (nouns, verbs, and adjectives) contained in positive examples in the training data using Equation (1).

$$W_p(t, S_p) = TF(t, S_p) \times H(t, S_p), \tag{1}$$

where S_p is a set of characteristic sentences belonging to positive examples in the training data, $TF(t, S_p)$ is the occurrence frequency of word *t* in positive example S_p, and $H(t, S_p)$ is the entropy based on the occurrence probability of word *t* in positive example S_p. $H(t, S_p)$ is obtained using Equation (2).

$$H(t, S_p) = \sum_{s \in S_p} P(t, s) \log_2 P(t, s) \tag{2}$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)}, \tag{3}$$

where $P(t, s)$ is the probability of word *t* appearing in characteristic sentence *s*, and $tf(t, s)$ is the frequency of word t in characteristic sentence *s*.

Next, the proposed method assigns weight $W_n(t, S_n)$ to content words in negative examples in the training data using Equation (4).

$$W_n(t, S_n) = TF(t, S_n) \times H(t, S_n) \tag{4}$$

Here, S_n is a set of characteristic sentences belonging to negative examples in the training data. In addition, *w* is extracted as a feature if word *w* satisfies the following condition.

$$W_p(t, S_p) \ > 2 \times W_n(t, S_n) \tag{5}$$

$$W_n(t, S_n) > 2 \times W_p(t, S_p) \tag{6}$$

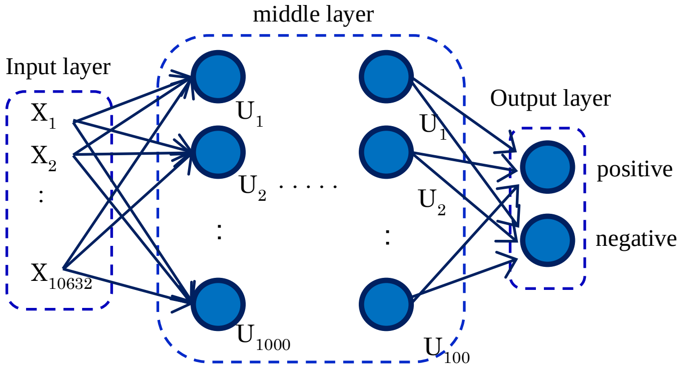
By imposing the above conditions, only important words in the positive and negative examples are selected as features, and general words that appear frequently in both the positive and negative examples are removed as features.

In this study, 3,627 patents containing 3,372 patents for “有機 EL” (*yuuki EL*: organic EL) and 255 patents for “放熱材料” (*hounetu zairyou*: heat-dissipating material) were used as training data for deep learning. Using the above method, 10,632 words were selected as features from 173,814 sentences in the training data. Examples of words extracted as features are given as follows.

用いる(*motiiru*: use), 化合(*kagou*: compound), 方法(*houhou*: method), 電子(*densi*: electron), 輸送(*yusou*: transport), 構成(*kousei*: composition), 使用(*siyou*: use), 実施(*zissi*: embodiment), 形態(*keitai*: form), 抑制(*yokusei*: suppression), 効率(*kouritu*: efficiency), 向上(*koujou*: improvement)

2.1.3 Extraction of characteristic sentences

In the proposed method, we use the multilayer perceptron (MLP) as the deep learning model. Here, number of nodes in the input layer is the same as the number of selected features (10,632). The middle layers comprises 12 layers, and the number of nodes in three layers is 1000, the number of nodes in three layers is 500, the number of nodes in three layers is 200, and the number of nodes in three layers is 100, respectively. The deep learning model is illustrated in Figure 2.



※The middle layer is 12 layers.

Figure 2: Deep learning model (MLP)

The proposed method extracts characteristic sentences from patent documents by the model trained using the training data. In the following, we demonstrate some characteristic sentences extracted from chemistry patent documents using the proposed method.

・架橋させることにより、分子間の架橋密度が増加し、耐熱性や機械物性を向上させることができる。(*kakyou saseru kotoniyori, bunnsikan no kayou mitudo ga zouka si, taonetusei ya kikai bussei wo koujou sasetu kotoga dekiru:* Crosslinking increases the cross-link density between molecules and can improve heat resistance and mechanical properties.)

・C−F結合エネルギーの高さゆえに耐熱性、耐薬品性等を付与することができる。(*C-F ketugou enerugi no takasa yueni tainetusei, taiyakuhinnsei nado wo huyo surukotoga dekiru*: Heat resistance, chemical resistance, etc. can be added because C-F bond energy is high.)

2.2 Step 2: Extracting characteristic term candidates

2.2.1 Generating word2vec model

We assume characteristic terms are used in the same context throughout patent documents, and the proposed method extracts characteristic term candidates from words appearing before and after characteristic terms (pseudo context). Here, we use the continuous bag-of-word model (CBOW) as the word2vec model for extraction [10]. The CBOW model is a neural network comprising three layers, i.e., an input layer, a hidden layer, and an output layer. The CBOW model is shown in Figure 3.

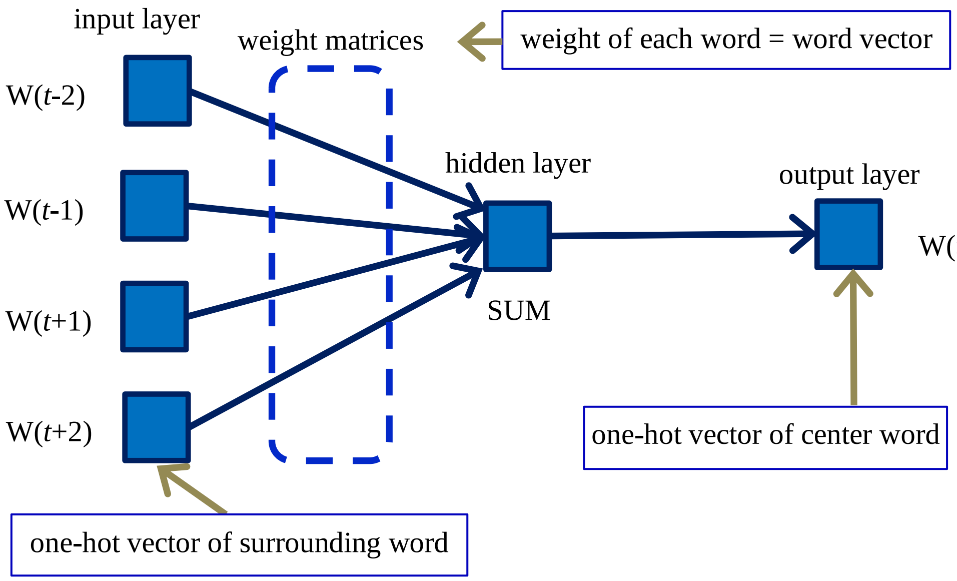

input layer
weight matrices
weight of each word = word vector
W(t-2)
W(t-1)
W(t+1)
W(t+2)
hidden layer
SUM
output layer
W(t)
one-hot vector of center word
one-hot vector of surrounding word

Figure 3: CBOW model

The training data can be generated automatically from a set of sentences, and the output is a one-hot vector of a target word in the sentence, and input is a one-hot vector of the surrounding words of the target word. The weights of the linear transformation from the input layer to the middle layer in this model can be obtained as a distributed representation of the words.

2.2.2 Extracting characteristic term candidates

We use all words and phrases contained in input patent documents to train the word2vec model. When a word is given, a word vector of the given word can be obtained. By calculating the cosine similarity between word vectors, it is possible to extract words used in the same context as the input words. The similarity $cos(V_{in}, V_t)$ between word vectors is calculated using Equation (7).

$$cos(V_{in}, V_t) = \frac{V_{in} \cdot V_t}{\|V_{in}\| \times \|V_t\|} = \frac{\sum_{i=1}^{n} x_i \, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (7)$$
$$V_{in} = \{x_1, x_2, \cdots, x_n\}, V_t = \{y_1, y_2, \cdots, y_n\},$$

where V_{in} is the word vector of the input word stored in the word2vec model trained by the input patent documents, and V_t is a word vector stored in the word2vec model.

We input words with a high likelihood of being characteristic terms into the trained word2vec model, and the proposed method extracts words output from the word2vec model as characteristic term candidates. Words with a high likelihood of being characteristic terms are extracted from the characteristic sentences extracted by deep learning (Step 1). The method used to extract words with a high likelihood of being characteristic terms is described as follows.

Step 2.1: The proposed method extracts words with “性” (*sei*: characteristic) that are attached to the ending contained in characteristic sentences, e.g., “耐熱性” (*tainetu sei*: heat resistant), for each applicant and counts their appearance frequency.

Step 2.2: The proposed method extracts 10 terms in descending order of appearance frequency as terms with a high likelihood of being characteristic terms for each applicant.

The proposed method inputs the words with high likelihood of being characteristic terms to the trained word2vec model. Among the words output from the model, words with similarity $cos(V_{in}, V_t)$ of 0.5 or more are extracted as characteristic term candidates for each applicant.

2.3 Step3: Extracting characteristic terms

The proposed method extracts characteristic terms in characteristic sentences for each applicant of patents as characteristic terms. Here, characteristic sentences are extracted for each applicant of patents; thus, it is possible to extract characteristic terms for each applicant by extracting only the characteristic term candidates in the characteristic sentences. In addition, the proposed method extracts characteristic term candidates whose document frequency is greater than threshold *v* as characteristic terms to prevent the extraction of inappropriate characteristic terms.

Some characteristic terms extracted by the proposed method from patent documents related to “resist material” of “FUJIFILM” are shown in the following.

硬化性(*koukasei*: curability), 感度(*kando*: sensitivity), 密着性(*mittyakusei*: adhesion), 耐熱性(*tainetusei*: heat resistance), 解像性(*kaizousei*: resolution), 現像性(*genzousei*: developability), パターン形状(*pata-n keijyou*: pattern shape), エッチング耐性(*ettingu taisei*: etching resistance), 保存安定性(*hozon annteisei*: storage stability)

Some characteristic terms extracted by the proposed method from patent documents on “resist material” of “JSR” are shown in the following.

接着性(*settyakusei*: adhesion), 透明性(*toumeisei*: transparency), アルカリ現像性(*arukari gennzousei*: alkali developability), MEEF(mask error enhancement factor), 放射線感度(*housyasen kando*: radiation sensitivity), 解像性能(*kaizou seinou*: resolution), 表面硬度(*hyoumen koudo*: surface hardness), 耐溶剤性(*taiyouzaisei*: solvent resistance)

3 Evaluation

Characteristic terms cannot be evaluated properly unless they are evaluated by experts, however, the proposed method occasionally extracts characteristic terms in large numbers, and it is difficult for experts to evaluate many characteristic terms. It is challenging to measure the goodness of precision and recall of a patent, even with only a partial evaluation of the data, because it requires experts to carefully read the patent specification and evaluate the characteristic terms. In the case of plural experts' evaluations, different experts' results are not always consistent. Even if an expert in the field evaluates a patent, the evaluation is not reproducible for the above reasons.

Therefore, we mechanically evaluated the proposed method using F terms attached to patents. The F term is a detailed classification code manually assigned to each characteristic of the technology shown in the patent. The classification code has a hierarchical structure indicated by dots. Figure 4 shows an example of the hierarchical structure of F terms.

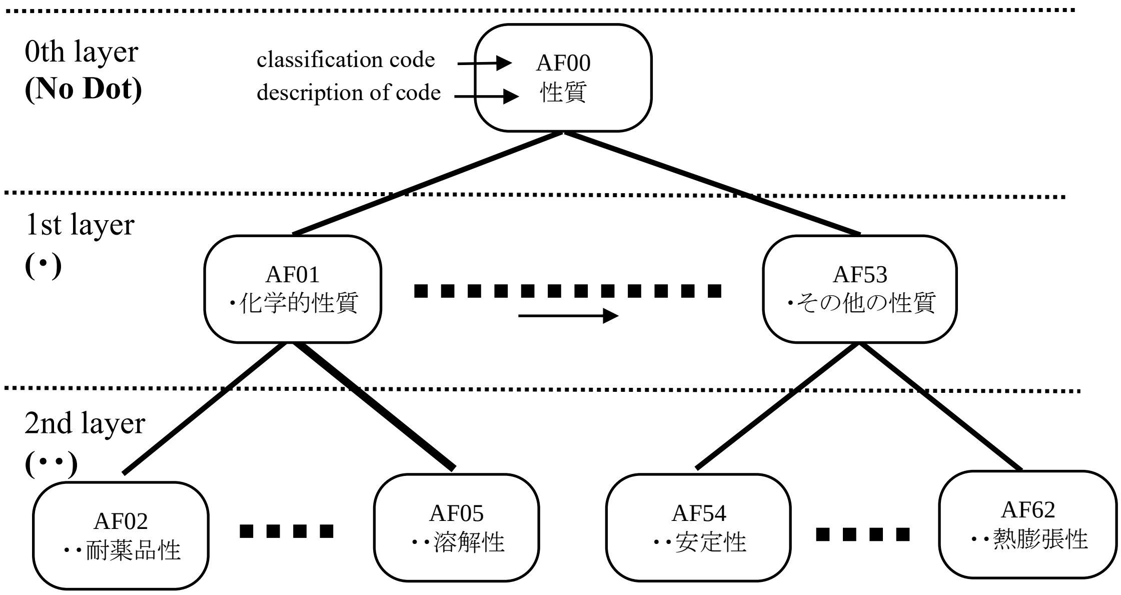


Figure 4: Hierarchical structure of F terms

As examples, F terms attached to patents classified as “包装体” (*housoutai*: package) are shown in the following.

熱収縮性(*netu kyuusyuusei*: heat shrinkable), ストレッチ性(*sutorettisei*: stretchability), 通気性(*tuukisei*: breathable), 気体透過防止(*kitai touka bousi*: gas permeation prevention), 防湿防湿(*bousitu*: Moisture proof), 酸素透過防止(*sanso touka bousi*: oxygen barrier), 耐水性 (*taisuisei*: water resistant), 透水性(*tousuise*: permeability), 吸湿性(*kyuusitusei*: hygroscopic), 蒸気透過性(*jyouki toukasei*: vapor permeability), 透明性 (*toumeisei*: transparency)

Note that, since F terms were assigned to patents manually, the number of F terms is small. In contrast, the proposed method extracts many characteristic terms. Specifically, we input patent documents with F terms to the proposed method and use the method to extract characteristic terms from patent documents. Using the F terms, we calculate recall *R* using Equation (8).

$$R = \frac{|M \cap N|}{|N|}, \tag{8}$$

where *M* is a set of characteristic terms extracted from patent documents by the proposed method, and *N* is a set of F terms attached to the patent documents.

We evaluated the proposed method using three theme patents in patent publications from 2013 and 2016. Table 1 shows the details of the patent documents used for this evaluation.

Table 1: Patents used for evaluation

Code	Explanation	Num. of patents	Num. of F terms
4F071/AF	Production of polymer moldings	3,136	62
4J038/NA	Effect of paint and depleting agent	5,544	27
3E067/CA	Properties of the material of the package	2,469	24

Table 2 shows examples of the F terms attached to patents with code 4F071/AF.

Table 2: Examples of F terms attached to patents with code 4F071/AF

Code	Explanation
AF00	性質(*seisitu*: characteristic)
AF01	・化学的性質(*kagakuteki seisitu*: chemical property)
AF02	・・耐薬品性(*taiyakuhinsei*: chemical resistance), 耐アルカリ(*taiarukari*: alkali resistance)
AF03	・・耐酸化性(*taisankasei*: oxidation resistance)
AF04	・・親水性(*sinsuisei*: hydrophilic), 疎水性(*sosuisei*: hydrophobic)
AF05	・・溶解性(*youkaisei*: solubility), 不溶性(*huyousei*: insoluble)
AF06	・物理的性質(*buturiteki seisitu*: physical property)
AF07	・・透過性(*toukasei*: transparency), 遮断性(*syadansei*: barrier)
AF08	・・・気体の通気性(*kitai no toukisei*: gas breathability)
AF09	・・・液体の防水性(*ekitai no bousuisei*: liquid waterproof)
…	…
AF53	・その他の性質(*sonota no seisitu*: other properties)
AF54	・・寸法安定性(*sunpou anteisei*: dimensional stability), 形状保持性(*keijyouhozisei*: shape retention)
AF62	・・熱膨張性(*netuboutyousei*: thermal expansion)

The evaluation results are shown in Table 3.

Table 3: Evaluation results

Code	$\|M \cap N\|$	Num. of F terms	R
4F071/AF	40	62	0.645
4J038/NA	23	27	0.852
3E067/CA	16	24	0.667

Next, we evaluated the proposed method in more detail for patent documents with code 4F071/AF. Here, we created a dictionary (2,946 terms) to convert characteristic terms to F terms manually. Table 4 shows part of the dictionary created to convert characteristic terms to F terms.

Table 4: Part of dictionary to convert characteristic terms to F terms

Characteristic term	Code
耐熱性(*tainetusei*: heat resistance)	AF45
透明性(*toumeisei*: transparency)	AF30
柔軟性(*jyunansei*: flexibility)	AF26
機械的強度(*kikaiteki kyoudo*: mechanical strength)	AF14
耐久性(*taikyuusei*: durability)	AF57
成形性(*seikeisei*: formability)	AF53
寸法安定性(*sunpou anteisei*: dimensional stability)	AF54
耐衝撃性(*taisyougekisei*: impact resistance)	AF23

加工性(*kakousei*: workability)	AF13
接着性(*settyakusei*: adhesion)	AF58
剛性(*gousei*: rigidity)	AF14
意匠性(*iyhousei*: creativity)	AF53
耐候性(*taikousei*: weather resistance)	AF57
光学特性(*kougaku tokusei*: optical characteristics)	AF29
難燃性(*nannensei*: flame retardance)	AF47
耐薬品性(*taiyakuhinsei*: chemical resistance)	AF02
導電性(*dendousei*: conductivity)	AF37
熱伝導性(*netu dendousei*: thermal conductivity)	AF44
絶縁性(*zetuensei*: insulation)	AF39
耐水性(*taisuisei*: water resistance)	AF09
耐摩耗性(*taimamousei*: wear resistance)	AF22
透明度(*toumeido*: transparency)	AF30
光透過性(*hikaritoukasei*: light transmission)	AF30
接着力(*settyakuryoku*: adhesion)	AF58
滑り性(*suberisei*: lubricity)	AF27
耐溶剤性(*taiyouzaisei*: solvent resistance)	AF02
帯電防止性(*taiden bousisei*: antistatic)	AF38
靭性(*zinsei*: toughness)	AF13
平滑性(*heikatusei*: smoothness)	AF27

We calculated the precision $P'(v)$, recall $R'(v)$, and F value $F'(v)$ for the characteristic terms contained in more patent documents than threshold v using Equations (9), (10), and (11). Here, threshold v is the document frequency when the proposed method extracts characteristic terms (Section 2.3).

$$P'(v) = \frac{G(M'(v))}{|M'(v)|}, \tag{9}$$

$$R'(v) = \frac{|M'(v) \cap N|}{|N|}, \tag{10}$$

$$F'(v) = \frac{2 \times P'(v) \times R'(v)}{P'(v) + R'(v)}, \tag{11}$$

Using the created dictionary, the proposed method converts the extracted characteristic terms to F terms. Then, we define a characteristic term converted to an F term using the dictionary as "F term based on characteristic term," and we define an F term originally assigned to the patent as a "substantial F term."

Here, $M'(v)$ is a set of "F terms based on characteristic terms" extracted by the proposed method whose document frequency (i.e., number of patents) is not less than threshold v. In addition, $G(M'(v))$ is the number of F terms correctly extracted among F terms in $M'(v)$, and N is a set of F terms attached to the patent documents.

Figure 5 shows a graph of precision $P'(v)$, recall $R'(v)$, and F value $F'(v)$ obtained when the threshold value was varied by 9, and Figure 6 shows a graph of precision $P'(v)$, recall $R'(v)$, and F value $F'(v)$ when the threshold value was varied by 1.

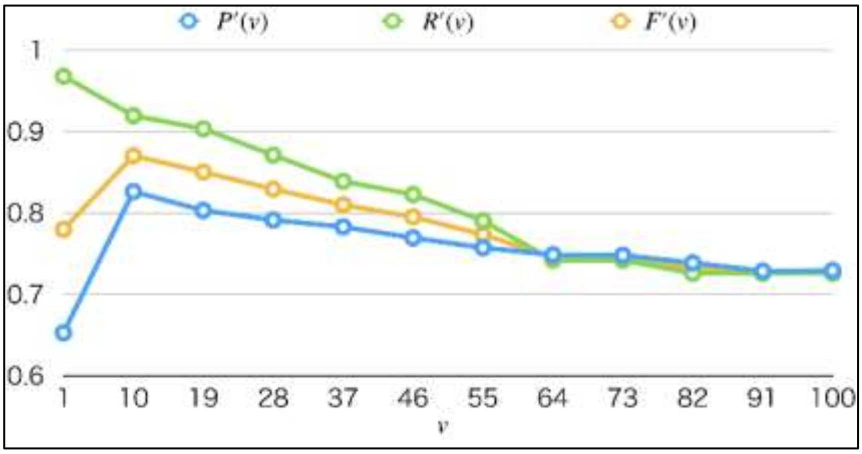


Figure 5: Change of $P'(v)$, $R'(v)$, and $F'(v)$ ($1 \leq v \leq 100$)

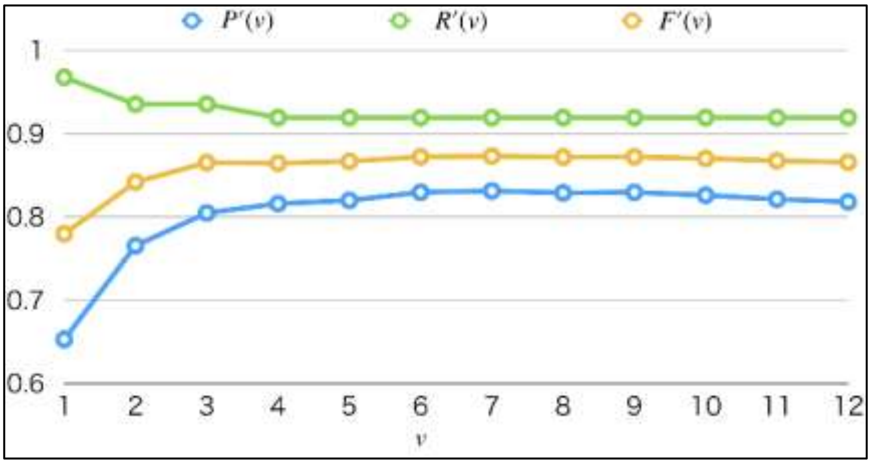


Figure 6: Change of $P'(v)$, $R'(v)$, and $F'(v)$ ($1 \leq v \leq 12$)

Using a dictionary to convert the above characteristic term to an F term, we performed a time-series evaluation for each company (applicant). Here, we selected companies with 100 or more patent documents from the evaluation patents with code 4F071/AF. We selected 11 companies as patent applicants. Examples of the selected companies are given as follows.

コニカミノルタ(Konica Minolta), 三菱ケミカル(Mitsubishi Chemical),
住友化学(Sumitomo Chemical), 富士フイルム(Fujifilm) , 日東電工(Nitto Denko),
旭化成(Asahi Kasei), 東レ(Toray), 東洋紡(Toyobo), カネカ(Kaneka),
クラレ(Kuraray), 積水化学工業(Sekisui Chemical)

We compared the characteristic terms extracted by the proposed method ($v = 1$) for each selected company with F terms attached to the evaluation patents. For comparison, we first summarized the "F terms based on characteristic term" extracted by the proposed method and F terms contained in each patent in the first F term hierarchy. We then calculate the correlation coefficient for each company and year. Regarding the correlation coefficient, we ranked the "F terms based on characteristic terms” and “substantial F terms” in order of document frequency (i.e., the number of patents containing “F terms based on characteristic terms” and “substantial F terms”) for each company and year. In addition, we calculated Spearman’s rank correlation coefficient $\rho(s,d)$ in year d of company s using Equation (12).

$$\rho(s,d) = 1 - \frac{6\sum D(s,d)^2}{N(s,d)^3 - N(s,d)}\ , \tag{12}$$

where

$D(s,d)$ is the difference between the order of “F term based on characteristic term” and “substantial F terms” corresponding to it in year d of company s, and $N(s,d)$ is the number of pairs (“substantial F terms” and “F terms based on characteristic terms”) in year d of company s.

The rank correlation coefficient $\rho(s,d)$ was then calculated. The results for the case where it appeared more than once in “substantial F terms” are shown in Figures 7 and 8. In addition, the results for the case where it appeared more than once in “substantial F terms” and “F terms based on characteristic terms” are shown in Figures 9 and 10.

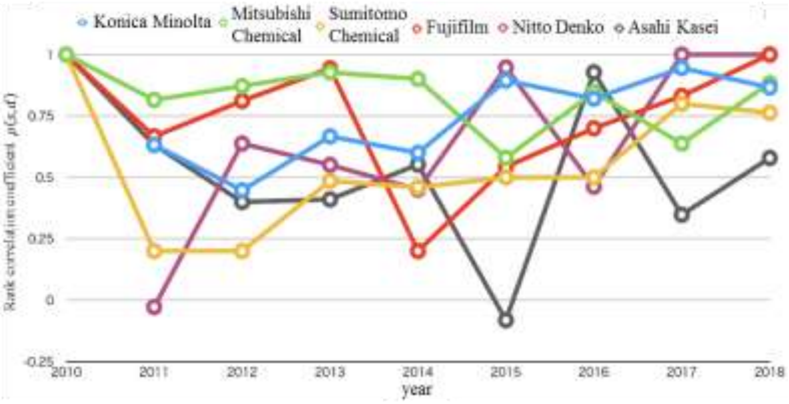


Figure 7: Results of Spearman's rank correlation coefficient (1-1)

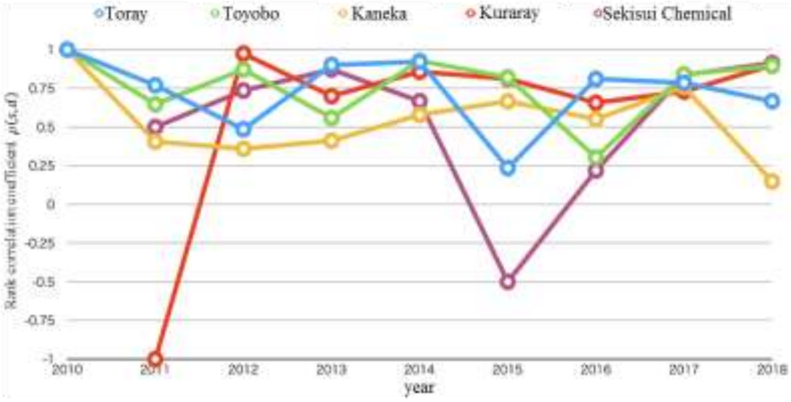


Figure 8: Results of Spearman's rank correlation coefficient (1-2)

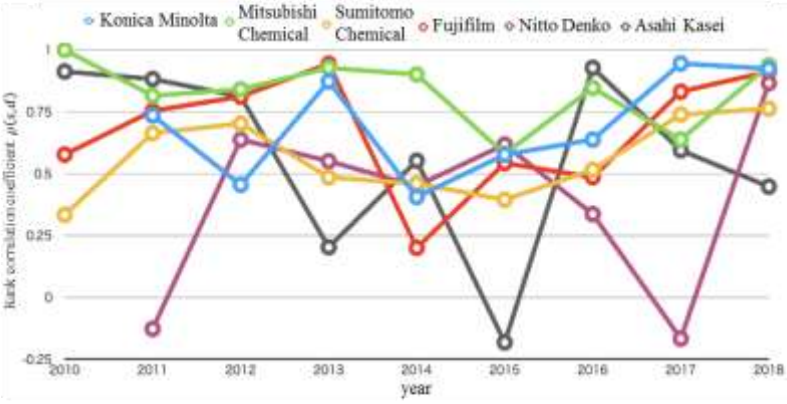


Figure 9: Results of Spearman's rank correlation coefficient (2-1)

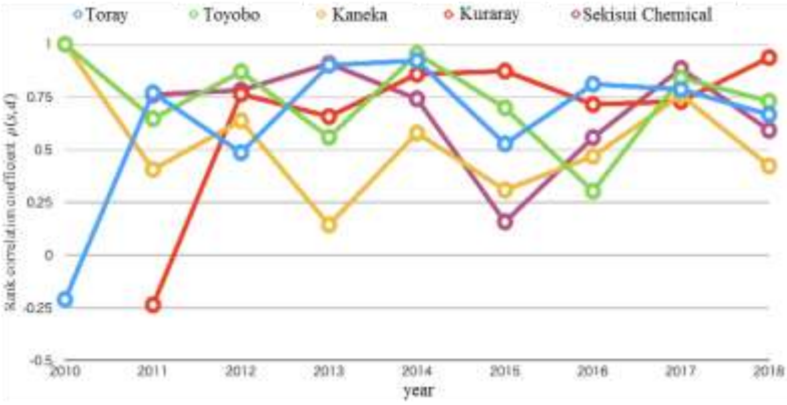


Figure 10: Results of Spearman's rank correlation coefficient (2-2)

We also calculated the average coverage $\bar{C}(s,d)$ and overall coverage $C'(s,d)$ of company s in year d to evaluate coverage. Here, $\bar{C}(s,d)$ and $C'(s,d)$ are expressed as Equations 13 and 14, respectively.

$$\begin{gathered}\bar{C}(s,d) = \frac{\sum_{w \in Q} C(s,d;w)}{|Q|} \qquad (|Q| \neq 0), \\ Q = F_x(s,d) \cap F_y(s,d),\end{gathered} \tag{13}$$

$$\begin{gathered}C'(s,d) = \frac{\sum_{w \in Q} |A|}{\sum_{w \in Q} |B|} \qquad \left(\sum\nolimits_{w \in Q} |B| \neq 0\right), \\ A = I(F_x(s,d),w) \cap I\big(F_y(s,d),w\big), \\ B = I\big(F_y(s,d),w\big),\end{gathered} \tag{14}$$

where

$F_x(s,d)$ is a set of "F terms based on characteristic terms" with one or more patent documents in year d of company s, and $F_y(s,d)$ is a set of "substantial F terms" with one or more patent documents in year d of company s. In addition, $I(F_x(s,d),w)$ is a set of patent documents with F term $w \in F$ $(s,\ d)$ in year d of company s, and $C(s,d;w)$ is the coverage for F term w in company s for year d. $C(s,d;w)$is calculated as follows.

$$C(s,d;w) = \frac{|A|}{|B|} \quad (|B| \neq 0) \tag{15}$$

The created graphs are shown in Figures 11, 12, 13 and 14, respectively.

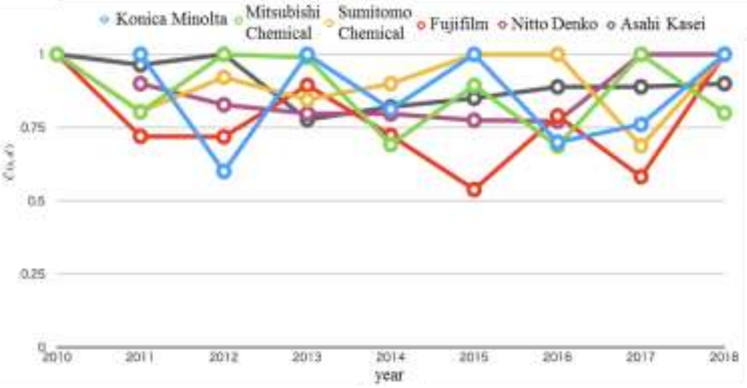


Figure 11: Average $\bar{C}(s,d)$ (1/2)

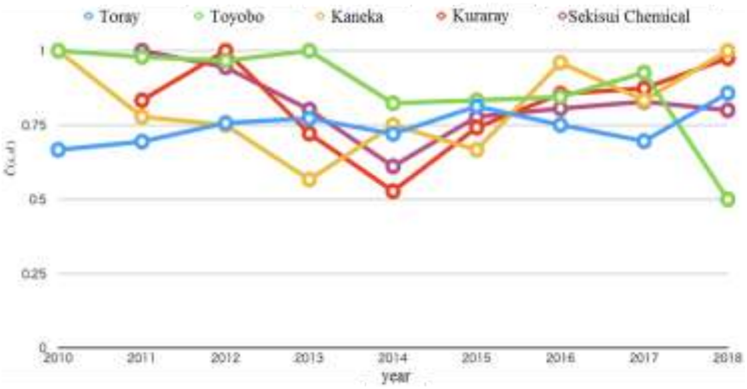


Figure 12: Average $\bar{C}(s,d)$ (2/2)

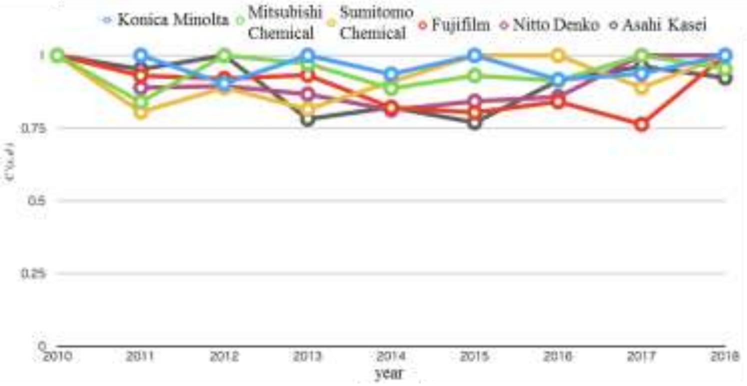


Figure 13: Overall $C'(s,d)$ (1/2)

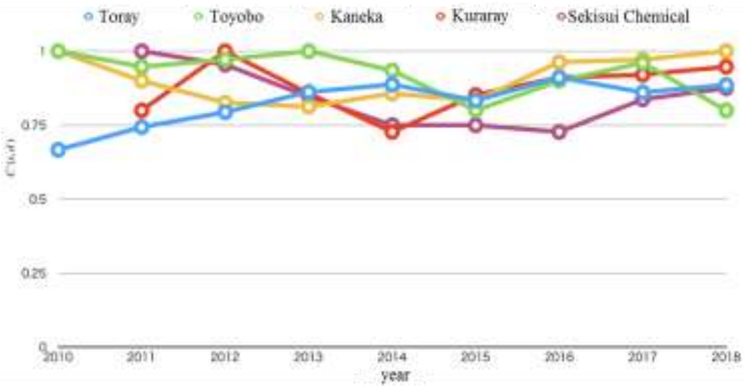


Figure 14: Overall $C'(s,d)$(2/2)

4 Discussion

As shown in Table 3, the proposed method achieved high recall when patent documents of code 4J038/NA were input. However, patent documents of code 4J071/AF and 3E067/CA were significantly less than the recall of patent documents of code 4J038/NA because, in some cases, synonyms of F term are occasionally used in patents. The follow shows F terms that could not be extracted by the proposed method from patent documents of code 3E067/CA.

For example, “ストレッチ性” (*sutorettisei*: stretchability) is a synonym for “柔軟性” (*jyuunannsei*: flexibility), “気体透過防止” (*kitai touka bousi*: gas permeation prevention) is a synonym for “気体遮断性” (*kitai syadansei*: gas barrier), and “酸素透過防止” (*sanso touka bousi*: oxygen barrier)is a synonym for “酸素バリア性” (*sanso baria sei*: oxygen permeation prevention). Note that these terms were extracted by the proposed method. In this evaluation, the word in which the correct answer F term and the character string perfectly agreed was taken as a mechanically correct answer. Therefore, even if synonyms of F term were extracted, recall R decreased because it is incorrect since it does not coincide with the character string of the F term that is the correct answer. The evaluation result was 0.95 for $R'(v)$ when $v = 1$, which is a significant improvement from 0.645. This is apparent from Figures 5 and 6, which show the results when “F terms based on characteristic terms” was the correct answer.

As shown in Figures 5 and 6, the proposed method achieved high precision $P'(v)$, recall $R'(v)$, and F value $F'(v)$ on average even if the threshold v changed when patent documents of code 4F071 / AF were input. However, when threshold v was 1, precision $P'(v)$ was approximately 0.65, which is extremely low compared to the other results. The F value $F'(v)$ remained nearly flat while threshold value v was 3-12 and dropped gently from 10 to 100. Therefore, unless we want to actively investigate characteristics that are not assigned as “substantial F terms,” we consider that it is more likely to obtain better results by using characteristic terms with a document frequency of 3 or greater ($3 \le v \le 12$).

As shown in Figures 7 and 8, although there was some variability among applicants and years, the average Spearman's rank correlation coefficient $\rho(s,d)$ was approximately 0.5. These results show a weak correlation between the rankings using “F terms based on characteristic terms” and “substantial F terms.” Furthermore, these results indicate a weak correlation between “F terms based on characteristic terms” and “substantial F terms.”

We consider that the reason for this result was that “substantial F terms” were not assigned exhaustively to all characteristics indicated by the patent. Therefore, we consider that the ranking likely differed and the correlation weakened. In addition, the values of rank correlation coefficient $\rho(s,d)$ are greater in Figures 7 and 8 than in Figures 9 and 10. In Figures 7 and 8, the characteristic terms that appear in only "F terms based on characteristic terms" were not used; however, in Figures 9 and 10, the characteristic terms that appear in both "substantial F terms" and "F terms based on characteristic terms" were used because the difference in ranking between "F terms based on characteristic terms" and "substantial F terms" is unlikely to occur.

As shown in Figures 11, 12, 13, and 14, the average coverage $\bar{C}(s,d)$ and overall coverage $C'(s,d)$ were approximately 0.8 on average. From these results, we found that the characteristic terms automatically extracted by the proposed method and given to the patent roughly include the characteristic given as "substantial F terms."

In addition to calculating the correlation coefficient and coverage, we calculated the number of patent documents (document frequency) containing each F term for each year. Figure 15 shows the graph for "substantial F terms," and Figure 16 shows the graph for "F terms based on characteristic terms."

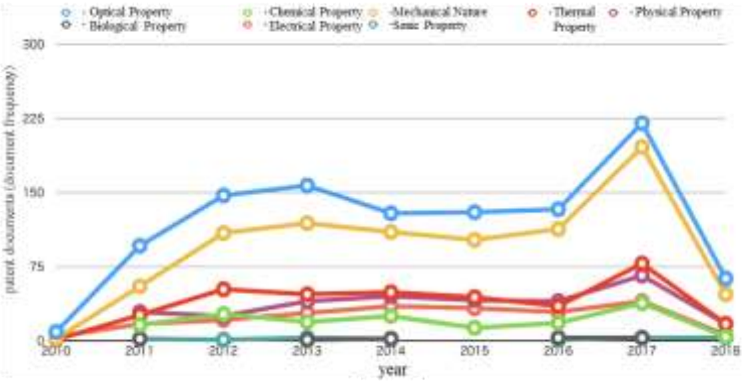


Figure 15: Document frequency of "substantial F terms"

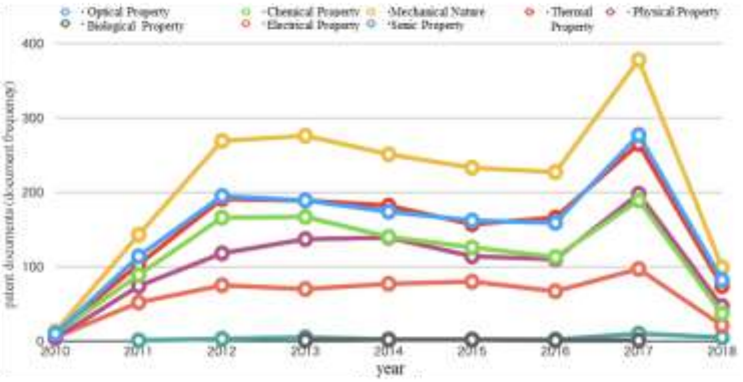
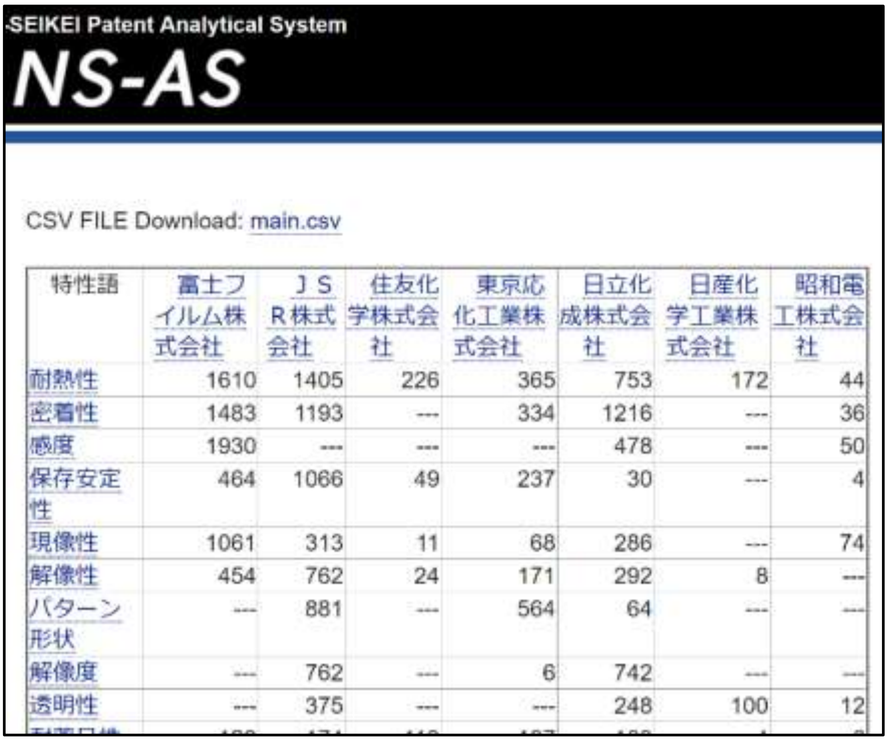


Figure 16: Document frequency of "F terms based on characteristic terms"

As shown in Figures 15 and 16, the number of F terms assigned to patents was greater on average in "F terms based on characteristic terms" than "substantial F terms." Therefore, we consider that the characteristic terms extracted by the proposed method are a useful information source when we seek to investigate characteristics given as "substantial F term" and characteristics not given as "substantial F term" by examining the characteristics given as "F terms based on characteristic terms."

To demonstrate a practical application of characteristic terms extracted by the proposed method, we implemented a WEB-based system for technology trend analysis using characteristic terms. Figure 17 shows an example of technology trend analysis using this system.



CSV FILE Download: main.csv

特性語	富士フイルム株式会社	ＪＳＲ株式会社	住友化学株式会社	東京応化工業株式会社	日立化成株式会社	日産化学工業株式会社	昭和電工株式会社
耐熱性	1610	1405	226	365	753	172	44
密着性	1483	1193	---	334	1216	---	36
感度	1930	---	---	---	478	---	50
保存安定性	464	1066	49	237	30	---	4
現像性	1061	313	11	68	286	---	74
解像性	454	762	24	171	292	8	---
パターン形状	---	881	---	564	64	---	---
解像度	---	762	---	6	742	---	---
透明性	---	375	---	---	248	100	12

Figure 17: Analysis of characteristic terms by companies

Figure 6 shows the frequency of characteristic terms extracted by the proposed method from the 2,562 patent documents on “resist materials” for each company (applicant). For example, the word “感度” (*kando*: sensitivity) was often extracted from patents of “FUJIFILM” but not from “JSR.” In addition, “保存安定性” (*hozonn annteisei*: storage stability) was often extracted from patents of “JSR” but not from many from patents of other companies. Thus, we consider that it is possible to analyze the tendency of technology that companies are good at by analyzing the number of characteristic terms extracted by the proposed method for each company.

5 Conclusion

In this paper, we have proposed a method to automatically extract characteristic terms as key phrases suitable for technology trend analysis from patent documents belonging to a certain technical field. The proposed method comprises three steps. In Step 1, the proposed method extracts characteristic sentences from patent documents for each applicant using a deep learning model. In Step 2, a word2vec model is trained from the input patent documents. The proposed method inputs terms with a high probability of being characteristic terms into the trained word2vec model, and extracted terms are output from the model as characteristic term candidates. Finally, in Step 3, the proposed method extracts characteristic terms for each applicant using the characteristic sentences extracted in Step 1 and characteristic term candidates extracted in Step 2.

We evaluated the proposed using F terms, and the proposed method achieved an average precision $P'(v)$ of 0.767, average recall $R'(v)$ of 0.809, and average F value $F'(v)$ of 0.787 when threshold v was varied from 1 to 100. In addition, the validity of the extracted characteristic terms was demonstrated by manually creating a dictionary to convert characteristic terms to F terms and verifying it using Spearman's rank correlation coefficient and coverage.

Acknowledgement

The authors would like to thank Enago (www.enago.jp) for the English language review.

References

[1] H. Nonaka, A. Kobayashi, H. Sakaji, Y. Suzuki, H. Sakai, and S. Masuyama, "Extraction of the Effect and the Technology Terms from a Patent Document," Journal of Japan Industrial Management Associastion, vol.63, no.2E, 2012, pp.105-111.

[2] H. Sakaji, H. Nonaka, H. Sakai, and S. Masuyama, "Cross-Bootstrapping: An Automatic Extraction Method of Solution-Effect Expressions from Patent Documents," The IEICE transactions on information and systems, vol.J93-D, no.6, 2010, pp.742-755. (in Japanese)

[3] H. Sakai, H. Nonaka and S. Masuyama, "Extraction of information on the technical effect from a patent document," The Japanese Society for Artificial Intelligence, vol.24, no.6, 2009, pp.531-540. (in Japanese)

[4] H. Li, F. Xu, and H. Uszkoreit, "TechWatchTool: Innovation and Trend Monitoring," Recent Advances in Natural Language Processing, 2011, pp. 660-665.

[5] M. Okamoto, Z. Shan, and R. Orihara, "Applying Information Extraction for Patent Structure Analysis," the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 989-992.

[6] S. Suzuki and H. Takatsuka, "Extraction of Keywords of Novelties From Patent Claims," Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1192-1200.

[7] M. Vazquez, M. Krallinger, F. Leitner, and A, Valencia, "Text mining for drugs and chemical compounds: methods, tools and applications," Molecular Informatics, Vol. 30, No. 6-7, 2011, pp. 506–519.

[8] A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," International Journal of Electrical, Computer, and Systems Engineering, Vol. 4, No. 2, 2010, pp. 155-170.

[9] L. Luo, Z. Yang, P. Yang, Z. Yin, L. Wang, H. Lin, and J. Wang, "An attention-based bilstm-crf approach to document-level chemical named entity recognition," Bioinformatics, Vol. 34, No. 8, 2018, pp. 1381–1388.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.

[11] S. Kitamori, H. Sakai and H. Sakaji, “Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning,” IEEE Symposium on Computational Intelligence for Financial Engineering & Economics, 2017, pp.67-73.