Extracting Problem Linkages to Improve Knowledge Exchange between Science and Technology Domains using an Attention-based Language Model

Hajime Sasaki
Institute for Future Initiatives
The University of Tokyo
Tokyo, Japan
sasaki@ifi.u-tokyo.ac.jp

Satoru Yamamoto
Data Artist Inc.
Tokyo, Japan
yamamoto@data-scientist.com

Amarsanaa Agchbayar
Data Artist Inc.
Tokyo, Japan
amar@data-artist.com

Nyamaa Enkhbayasgalan
Data Artist Inc.
Tokyo, Japan
enkhbayasgalan@mn.data-artist.com

***Abstract*—Science and technology activities can be considered problem-solving activities, and scientific papers and patent publications can be viewed as providing explicit knowledge gained from the problem-solving of academia and industry respectively. However, even in the same field, the approach to the same problem is not consistent between a paper and the patented technology. The creation of information silos in science and technology generates inefficiency in human intellectual production. Therefore, this study examines whether insights from technical problems can be shared with academics to solve scientific problems. We propose a concept to link the problems between these two domains using a linguistic approach for knowledge discovery that connects science and technology. We extracted scientific papers from the Association for Computational Linguistics dataset, and patent literature from the Derwent Innovation platform. From these, pairs of problem defining sentences were identified and extracted using an attention-based language model. For example, we were able to extract examples of issues that do not necessarily arise from scientific papers, such as annotation difficulties in the analysis of social network data, but can be hinted at by patented techniques prior to the paper. These results suggest that scientific problems and industrial solutions can provide mutual insight. This knowledge discovery approach is recommended not only for benefiting corporate activities but also for grasping research trends.**

Keywords-problem extraction; information matching; scientometrics; Literature-Based Discovery (LBD); attention-based language model

I. INTRODUCTION

Science progress and technology change have become important issues on innovation and economics studies [1-3]. The way science and technology interact is a long-standing question. The knowledge flow in some areas, such as pharmaceuticals, can be effectively explained by linear models through basic research, applied research, development, and diffusion (production) [4-6]. Linear models have been widely disseminated by academic institutions [7] lobbying for research funding, by economists [8] serving as expert advisors to policy makers and have been viewed as linear concepts of innovation by science and technology scholars [4]. On the other hand, recent research on innovation shows that such a linear model of innovation is insufficient to represent reality [5, 9-13]. The linear model does not consider the empirical evidence that technological change often results from experience and ingenuity rather than scientific theory and methods, the instrumental role of technological development in eliciting scientific explanation, and the importance of technology-based instruments for scientific research [14, 15]. It is sometimes pointed out that the linear model overlooks technology's influence on the setting of the scientific agenda [16, 17]. Innovation involves the transfer of knowledge between the scientific and industrial domains, as exemplified by the chain link model [11, 18, 19] and the network model [16]. Of course, the linear model has the ability to explain innovation, and thus it not the all ideas that consider innovation to be a linear model are wrong [20], however, when the complexity of science and technology interactions is understood, it is undeniable that science pushes technology and technology pushes science.

Information retrieval research involving academic articles and patents has a long history [21-23]. Information retrieval is defined as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [24]. Existing bibliometric methods regard research papers as representing scientific research and patents as representing innovation [25]. Non-patent literature is sometimes used as a method for directly measuring the

Corresponding author: Hajime Sasaki

relationship between science and technology [22, 26, 27]. The patent examiner searches for prior non-patent literature based on the patent specification field. The number of non-patent literature is an indication of how the relating technology has already been mentioned in the context of the science domain. As mentioned earlier, the relationship between science and technology is not always straightforward and simple. Traditionally, patent citations to papers have been comprehensively studied to understand the transfer of knowledge from science to technology [28]. However, the transmission from technology to science has not been well studied [29]. While patents contain detailed methodological information on successful innovations, references to patents are rarely found in applied science or science texts [29, 30, 31]. According to Glanzel and Meyer [29], the publications that have such reverse citations account for only 0.98% of all total publications between 1996 and 2000, of which 30% are in chemical-related fields. However, the absence of bibliographic references does not necessarily mean that technical and scientific knowledge are unrelated [16]. Further, the explosive increase in scientific and technical knowledge can be problematic. There are over three million articles written in English [32]. Regarding patents, there were 3.3 million patent applications in 2018, up 5.2% from 2017 for a ninth straight yearly increase [33]. In this context, it is becoming increasingly difficult for scientific papers and patents to fully reference each other. In every field, science and technology are fragmented into information silos, resulting in a condition in which one information system is unable to interoperate with other systems that are or should be associated with it. If debate proceeds only within the corresponding silos and information is not shared, even though the science and technology fields work on similar issues, the resources devoted to humankind's intellectual activities will be significantly wasted. In this study, we focus on the possibility of extracting common needs between science and technology that have not been fully addressed by existing articles.

There is an approach known as Literature-Based Discovery (LBD) [34-37]. One way to determine the common needs between two fields by using knowledge discovery methods and involving bibliographic information is Swanson's ABC model [38]. He succeeded in hypothesizing and verifying the unknown relationship between Raynaud's disease and fish oil based on bibliographic information. Although scientific papers and patent publications are usually used as knowledge sources for LBD, most studies using this approach focus on discovering hidden links in the same domain (scientific papers for science domain, patent articles for technology domain). For example, one paper discussed the semantic similarities between gerontology and robotics based on the clustering of direct citation networks in the scientific inner domain [39]. Meanwhile, other papers focus on knowledge discovery in a certain field in a cross-domain between scientific papers and patent publications. Authors in [40] identified the commercialization gap between fields amply discussed in the science domain but not nearly as well discussed in the technology domain, using the photovoltaics-related knowledge field as an example; they created clusters based on direct citation networks between scientific papers and patent publications and calculated the semantic similarities among these clusters [41]. Wang [42] also applied the same method for the micro biofuel field.

Considerable evidence indicates the importance of linking the same fields in the science and technology domains. Several studies have revealed how science pushes technological development [21, 27, 43, 44]. Thus, it is effective for industries to extract knowledge and contribute to the technology domain from the science domain [39]. Scientific articles provide readers with a problem-solving process in terms of an objective and reproducible knowledge. The Introduction, Methods, Results, and Discussion (IMRaD) format for scientific articles has been gradually adopted since the 1940s [45]. It allows the problem-solving process to be described explicitly and enables the reporting of scientific activities to follow a more standard construction. The establishment of such a document structure greatly influences the research of vocabulary patterns of written language. Authors in [46, 47] developed a way to analyze the structure of scientific and technological documents [45]. Their methods can be used as a possible approach for the purpose of automatic classification, information extraction, and automatic summarization for scientific articles. The extraction of sentences related to problem-solving can also be used for science and technology articles.

Research on information retrieval from patent publications has also attracted considerable attention. Many techniques have been proposed to classify text data of patent publications into problem and solution statements [48-50]. A Subject–Action–Object (SAO) structure can be recognized as a problem and solution pattern, which several patents have used [51-53]. Authors in [54] attempted to extract and analyze SAO structures to detect patent infringement. Authors in [55] focused on the identification of rapidly evolving technological trends, and authors in [56] proposed a method to recommend research and development candidates by extracting the SAO structure from problem–solution patterns of patent information. However, a few studies have described the relationships between problems and solutions extracted from papers. In this study, we will focus on this particular knowledge discovery issue. Regarding extracting information from technical documents, researchers have attempted to extract expressions that represent technical features from patent publications and scientific articles as subtasks of the patent mining task of NTCIR–8 (NII Testbeds and Community for Information access Research). This project aims at a large-scale evaluation for technologies that support the understanding and use of information, such as information retrieval, question answering, summarization, text mining, and machine translation, from a vast amount of information [57]. This extraction is also expected to be useful for the automatic creation of a technology trend map. However, as described above, because the terms used in patents are often more abstract or creative than those used in research papers to widen the scope of the claims, problem-extraction methods for patent publications are underdeveloped. On the other hand, based on the premise that the phrase "problem to be solved" in patent publications appropriately represents the technical problem, it is proposed that a more specific patent map can be created by paying attention to the sentence [58].

There are many approaches to information retrieval using scientific and patent texts, but there are still problems and uncertainties regarding the defining keywords. In this sense, information extraction that does not solely rely on keywords is required. Heffernan and Teufel [59] showed that word embeddings, a technique in which words are represented as vectors, can be used as features to extract sentences related to problems and solutions, using the Association for Computational Linguistics (ACL) anthology as a dataset. They claimed that the detection of the problem and solution statements from papers can enable the comparison of similar papers and lead to the automatic generation of review articles. However, they do not describe their method's application for cross-domain articles and also mention the linking of problem and solution statements as an area of future work.

The current study aims to answer the question "is it possible to extract sentences that refer to the same problem (i.e. needs) from both the science and technology domains and obtain information that contributes to knowledge discovery across domains?" referring to if knowledge from patents can provide insight into the scientific issues being investigated. In this paper, the concept of inter-domain links for knowledge discovery using a linguistic approach is proposed. This study makes a concrete contribution to the literature because it demonstrates the possibility of building a needs-focused portfolio that includes both science- and technology-related information by extracting appropriate sentences from scientific articles and patents. This study makes a concrete contribution to the literature because it shows it is possible to build a needs-focused portfolio that includes both science- and technology-related information by extracting appropriate sentences from scientific articles and patents. For example, research articles often mention potential future studies, and knowledge can be obtained from existing patent information for these future investigations. Thus, we show that not only does science support technology, but technology can support science. Another contribution of this research is a model that extracts problem statements (sentences) from the paper without preparing clue words in advance, and performs better than the existing method [59]. To achieve this, it is hypothesized that the application of a model of language understanding that enables context-sensitive processing, which has been evaluated in the field of natural language processing, would be effective.

II. METHOD

A. Overview

The methodology of this study is outlined in Figure 1. At first, data from scientific publications were taken from the ACL database, whereas data from patent publications were taken from the Derwent Innovation platform, as shown in Figure 1(1). Problem statements for patents were then identified by whether they begin with the phrase "problem to be solved," as shown in Figure 1(2). For scientific articles, the problem statements from the sampled data are extracted, as shown in Figure 1(3). Finally, we calculated the semantic similarity between the scientific and technical problem statements, as shown in Figure 1(4). These processes are described in more detail in the next section.

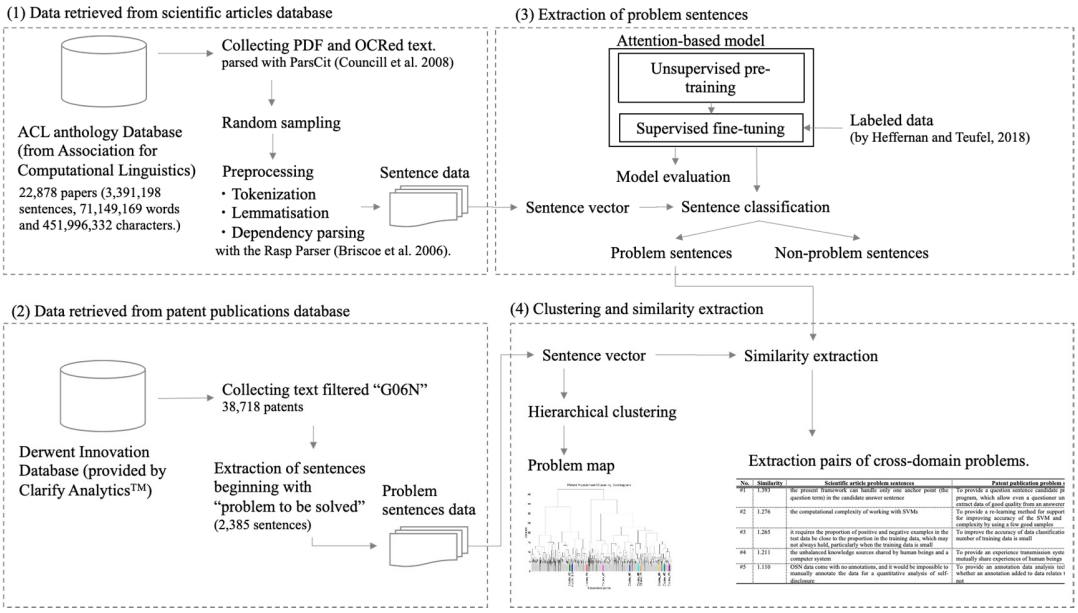

Fig. 1. Method overview.

B. Dataset

This section describes the data acquisition and preprocessing procedure shown in Figure 1(1). We considered that scientific articles contain scientific knowledge and patent articles industry knowledge. We limited the papers/patents to the field of computational linguistics. The computations linguistics corpus of scientific articles is a subset of the ACL anthology released in March 2016 and contains the full text of 22,878 articles. These data were parsed using ParsCit [60], and tokenization, sentencing, and dependency analysis were done with Rasp Parser [60]. We randomly sampled 2,500 articles

from this dataset, which is the same one used in [59] and is under the Creative Commons Attribution license (CC-BY). This allowed easy comparison of the classification performance with [59]. Patent data were extracted from the Derwent Innovation platform provided by Clarify Analytics. Computer science-related patent data classified as G06N were defined by the World Intellectual Property Organization as "a computer system based on a particular computational model" in the International Patent Classification. A total of 38,718 filtered patent publications were extracted from the database. The "problem-solving concept" is a statement describing the problem solved by the patent [61-63]. Patent gazettes often include important sentences that begin with the term "problem to be solved" [48]. Thus, we extracted statements containing "problem to be solved" from patents.

C. Extraction of Problem/Solution Sentences

The way sentences were classified as shown in Figure 1(2) is described in this section. We identified problem and solution sentences based on a previously established neural language approach by creating word embedding-based features [59]. Word embedding involves mapping words to a vector space in order to capture the meaning of the word or grammatical structure. It is based on the distribution hypothesis that words having similar meanings will appear in similar contexts, that is, the appearance distribution of surrounding words [64, 65]. Heffernan and Teufel [59] proposed a supervised learning model that classifies given sentences into problem or not-problem sentences. They indicated that embedding-based features were effective for classifying these sentences [59]. In this study, we used a neural network language model focused on "attention" that has become common [66-68]. "Attention" is a mechanism that allows machines to learn which vectors are important when there are multiple vectors. In other words, it informs the prediction model which part of the input data to focus on. We hypothesize that our method can extract problem sentences with higher accuracy by considering the entire context, whereas existing methods such as Word2Vec focus only on the area immediately before and after the clue word. In this study and by using this methodology, we constructed a model that determines whether a sentence is a problem statement based on whether it contains word with high probability to correspond to the problem. In this step, we conducted unsupervised and supervised learning.

1) Unsupervised Pre-Training

Given the token $\mathcal{U} = \{u_1, \dots, u_n\}$, the likelihood to be maximized in the standard stochastic language model is given by:

$$L_1(\mathcal{U}) = \sum_i logP(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

where k is the size of the window, and P is the neural network model with parameter Θ. Here, Θ is adjusted by stochastic gradient descent. A model using the attention mechanism is given by:

$$P(u) = \text{softmax}(h_n W_e^T) \quad (2)$$

where n is the number of layers in the neural network, and W_e is the token embedding matrix. In this study, we utilized a published learning model in which a multi-layer transformer decoder is implemented as a language model [67, 69, 70].

2) Supervised Fine-Tuning

Parameter adjustment was performed through supervised learning using the modeled function that has been learned in (1). We implemented tasks to classify problem and non-problem sentences as supervised learning tasks. Assuming a dataset $\mathcal{C}$ containing labels that contain a document consisting of a word string of input tokens $x^1, \dots, x^m$ and a label y in each instance. For example, suppose that there is a group of words that constitute a sentence as input tokens. If it is a problem sentence, 1 is assigned to the label y, and 0 is substituted if it is not. This input passes through the previously learned model with an output layer with parameters for predicting y:

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l{}^m W_y) \quad (3)$$

Below is a constraint that maximizes:

$$L_2(\mathcal{C}) = \sum_{(x,y)} logP(y|x^1, \dots, x^m) \quad (4)$$

We conducted a five-fold cross-validation and comprehensive evaluation with the average value of the following four evaluation indices.

- Precision is the percentage of positive data that is actually positive:

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)} \quad (5)$$

- Recall is the percentage of what was actually positive and was predicted to be positive:

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)} \quad (6)$$

- F-measure (F_1 score) is the harmonic mean of the precision and recall:

$$F_1 = \frac{2Precision \cdot Recall}{(Precision + Recall)} \quad (7)$$

- Accuracy is the percentage of data actually predicted to be positive or negative:

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \quad (8)$$

D. Clustering and Similarity Extraction

Here, the processing of clustering and extraction of similar problem pairs is described corresponding to Figure 1(4). We vectorized the documents against the obtained scientific paper problem sentences and patent problem sentences respectively, and performed clustering. For clustering, we used the Ward's [71] method, which is a kind of hierarchical clustering. Ward's method repeats the procedure of merging any two clusters with the smallest increment in the sum of squares in the cluster. This clustering method has shown high performance regarding hierarchical clustering. Additionally, Term Frequency-Inverse Cluster Frequency (TF-ICF) is calculated to extract the characteristic keywords for each cluster. The term frequency gives a measure of the importance of the term within the particular sentences. The inverse cluster frequency refers to a measure of the general importance of a term. The $TFICF$ of term i in cluster j is given as follows:

$$TFICF = tf_{i,j} \cdot icf_i = tf_{i,j} \cdot \log\left(\frac{N}{cf_i}\right) \quad (9)$$

where N is the total number of sentences. Each cluster was labeled based on the resulting characteristic keywords and sentences. Based on the dot products of embedded vectors of the obtained characteristic words, the similarities of problem sentences between the papers and patents were calculated. By focusing on problem-sentence pairs with high similarity, it is possible to manually confirm whether problem-solving information provided by the patents can be helpful to solve problems mentioned in scientific research papers.

III. RESULTS

A. Extracting Problem Sentences

Table II shows several results of scientific paper problem sentences classified with this model. The actual problem sentences were labeled as "1." Non-problem sentences were labeled as "0." The predicted result "1" means the sentence is predicted as a problem sentence. The predicted result "0" means the sentence is predicted as a non-problem sentence. Table I shows the classification performance index. Each number represents Precision/Recall/F_measure/Accuracy, in that order. Heffernan and Teufel's [59] results, whose study is the most similar to ours, are also shown for comparison. We were able to extract 2,385 sentences beginning with "problems to be solved" in patent publication abstracts. Table III represents an example of the actually extracted sentences.

B. Problem Clusters in Patents

Table IV shows a summary of the top ten clusters in the patents. Each cluster name was manually chosen after reviewing all featuring sentences. The first cluster was labeled "Information system" based on the problem sentences and keywords extracted by the TF-ICF, as the clusters related to information processing and input information technology. The second cluster was labeled "Memory efficiency and parameter optimization for neural networks," with many problem statements addressing efficiency and optimization. The third cluster was named "Data extraction and processing," with many problem statements including challenges in data extraction.

TABLE I. CLASSIFICATION PERFORMANCE

	Precision	**Recall**	**F_1**	**Accuracy**
[59]	0.82	0.83	0.82	0.82
Proposed	0.89	0.85	0.87	0.87

The fourth cluster, "Knowledge systems and humans" focuses on knowledge rather than data and therefore features several issues related to human behavior. The fifth cluster, which is a knowledge system as well, is named "User and knowledge systems", because many of the issue statements focus on the issues faced by users of the system. The sixth cluster was named "Data classification," with several problems focusing on classification, a machine learning task. The seventh cluster, "Image recognition," consists mainly of tasks that used images as data. The first bowl cluster was named "Circuit of a neuron model" because many of its issues focused on circuit design using the nervous system. Several task statements belonging to the ninth cluster focus on mathematical probabilistic tasks, and thus we named the cluster "Estimating parameters, probabilities, calculation methods, and so on." The tenth cluster is a concentration of issues in terms of control engineering and was named "Data processing for control engineering." The cluster name is representative of the set, and not all sentences matched the cluster name precisely. Figure 2 shows the result of the vectorization of each document against the problem sentence in patents followed by compression in two dimensions and clustering. The results are shown in different colors for the top ten clusters in order of cluster size. Other clusters are in gray.

TABLE II. SAMPLES OF EXTRACTED PROBLEM/NON-PROBLEM SENTENCES FROM SCIENTIFIC ARTICLES (N=10)

Sentences	**Label**	**Predicted**
"this reduces the efficiency of the dynamic programming"	1	1
"this is expensive"	1	1
"should probably be treated separately, as a preposition modifier"	0	0
"creating these rules requires much cost and that they are usually domain-dependent"	1	1
"it is not capable of modeling bilexical dependencies on the right hand side of the rules"	1	1
"Unsupervised constituency parsing is also an active research area"	0	0
"consuming very large parameter spaces"	1	1
"the time required to load and watch the videos"	1	0
"the need for large training data"	1	1
"the possible relationships that exist among the various factors"	0	1

TABLE III. SAMPLES OF PROBLEM-RELATED SENTENCES IN PATENTS (N=5)

Problem sentences (example)	**Application no.**
"Problem to be solved is the neuron action potential calculation speed is slow in large-scale computer simulation process, the method of the invention can greatly improve the calculation speed of the action potential, while maintaining a relatively high precision, and it is very suitable for simulation of large-scale brain nerve network."	CN106447032A
"PROBLEM TO BE SOLVED: To decrease the number of sensors in use without a significant loss of control precision by constituting a 2nd control system by using a 1st and a 2nd control signal."	JP2000187504A
"Problem to be solved is to easily register information for specifying the symptom not only by type but also by designating an individually managed subject."	WO2008007442A1
"PROBLEM TO BE SOLVED: To effectively recognize an object in a practical time, with practical accuracy and in a practical object range."	JP2001195381A
"PROBLEM TO BE SOLVED: To reduce the number of times of multiplication required for finding a covariant matrix for obtaining the coefficient of prediction for minimizing a square root error."	JP2001195586A

TABLE IV. SUMMARY OF TOP 10 CLUSTERS RELATED TO PATENT PROBLEM STATEMENTS

Cluster ID	No. of sentences	Cluster name	Keywords
Cluster #1:	21	Information system	identification, information, abnormality, assisting, creation, semi-supervised, artificial, image, input, system
Cluster #2:	17	Memory efficiency and parameter optimization	problem, solve, difficult, included, that, capability, sample, network, neural, conventional
Cluster #3:	17	Data extraction and processing	extracting, analyzing, annotation, correlation, expected, pattern, data, added, data, stored
Cluster #4:	16	Knowledge systems and humans	personality, artificial, intelligence, person, human, realize, answer, concepts, defined, divided
Cluster #5:	16	User and knowledge systems	knowledge, user, base, contents, concept, around, document, enormous, extracted, modeling
Cluster #6:	16	Data classification	classifying, classification, target, partial, support, enhancing, kind, source, high, unknown
Cluster #7:	15	Image recognition	monitoring, image, costs., evaluating, holding, interpretation, intention, generating, improving, attributes
Cluster #8:	14	Circuit of a neuron model	neuron, circuit, element., resistance, circuit, neural, output, network, element, bond
Cluster #9:	14	Estimating parameters, probabilities, calculation methods and so on	probability, calculation, similarity, arithmetic, unit, arbitrary, cluster, continuous, decision, independence
Cluster #10:	14	Data processing for control engineering	optimization, control, antenna, robust, ship, enhance, controller, service, efficiency

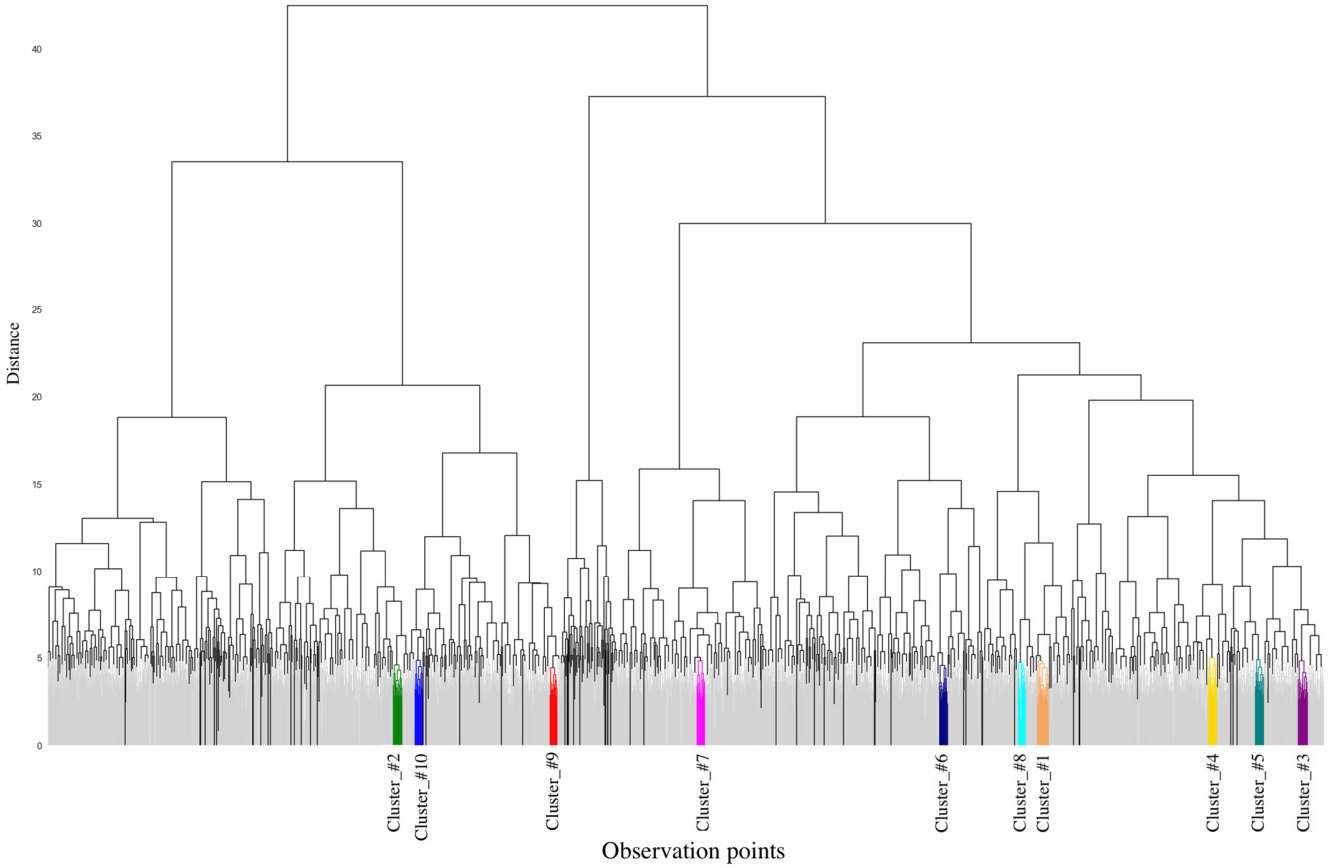

Fig. 2. Hierarchical clustering dendrogram in patents.

C. Semantic Similarity

For all the obtained scientific paper and patent problem sentences, feature words were extracted based on TF-ICF. The sentence similarity was calculated between the two different sources. Table V shows the five pairs with high similarity.

IV. DISCUSSION

At first, the proposed model's word selection and classification was described. In Table II, words and phrases such as "reduces the efficiency," "expensive," "it is not capable," and "consuming" intuitively imply the suggestion of a problem.

TABLE V. PROBLEM PAIRS WITH HIGH SEMANTIC SIMILARITY BETWEEN PATENT PUBLICATIONS AND SCIENTIFIC ARTICLES

No.	Similarity	Scientific article problem sentences	Patent publication problem sentences
#1	1.393	the present framework can handle only one anchor point (the question term) in the candidate answer sentence	To provide a question sentence candidate presentation device and a program, which allow even a questioner unfamiliar to interview to extract data of good quality from an answerer
#2	1.276	the computational complexity of working with SVMs	To provide a re-learning method for support vector machine (SVM) for improving accuracy of the SVM and reducing computational complexity by using a few good samples
#3	1.265	it requires the proportion of positive and negative examples in the test data be close to the proportion in the training data, which may not always hold, particularly when the training data is small	To improve the accuracy of data classification of test data even if the number of training data is small
#4	1.211	the unbalanced knowledge sources shared by human beings and a computer system	To provide an experience transmission system making it possible to mutually share experiences of human beings
#5	1.110	OSN data come with no annotations, and it would be impossible to manually annotate the data for a quantitative analysis of self-disclosure	To provide an annotation data analysis technology of determining whether an annotation added to data relates to content of the data or not

In the conventional sentence extraction from scientific articles and patents, such clue words must be manually collected in advance. However, this process is time-consuming and it is difficult to collect all clue words from thousands of documents. In addition, it can be difficult to determine whether a clue word indicates a problem. From Table I, we can confirm that the evaluation scale in the proposed problem/non-problem classifier exceeds the existing model [59]. The pre-learning attention mechanism classified documents in context to see if the clue words indeed correspond to a problem. This suggests that sentence classification in the proposed model is performed better than through the Word2Vec approach, which examines only the words before and after the clue words. In Table III five sample sentences are presented, all using the "problem to be solved" as a clue word from the abstract information in the patent data. It can be seen that each patent shows the objectives to be addressed, such as the slow computation speed (CN106447032A) and the desire to reduce the number of sensors without decreasing the precision (JP2000187504A). Table IV outlines the top 10 clusters resulting from these patent issue statements using Ward's method. Especially after the second cluster, the focus is on relatively specific tasks, indicating that the clustering of task sentences has been performed properly. In the dendrogram shown in Figure 2, the top 10 clusters are scattered, therefore, it is reasonable to understand that the top 10 clusters capture an overview of the issue awareness in the field. The third, fourth, and fifth clusters are close to each other. It can be seen from the keywords that both 4 and 5 are close in terms of knowledge and 3 and 5 are close in terms of extraction. From this dendrogram, it is possible to read the similarity of awareness of each other's issues.

Similar problem statements between article and patent texts (Table V) are discussed below.

Pair #1: The article phrase was extracted from the following complete sentence: "A more serious limitation is that the present framework can handle only one anchor point (the question term) in the candidate answer sentence," which comes from the section "Shortcoming and Extensions" of the paper "Learning Surface Text Patterns for a Question Answering System" [72]. This article examined an open-domain question answering system. The patent JP2011002872A has the most semantically similar problem sentences. This patent also refers to a question answering device and is related to interviewing problems stemming from human interviewers needing a sophisticated, adaptable interview technique. It is described as an effective way for handling multiple questions and extracting respondents' true intentions in an interview by flexibly changing question order and depth according to the flow of conversation with respondents. Based on the aforementioned problem, the invention proposes a mechanism for estimating topics of interest to respondents and presenting question candidates. In the scientific article problem sentence, the problem is that the machine making the question can only use a single viewpoint. Although viewpoints vary depending on whether the subject asking questions is a machine or human, the information in the patent publication could provide inspiration for solving the scientific problem.

Pair #2: The article phrase was extracted from "The real drawback is the computational complexity of working with SVMs [Support Vector Machines], thus the design of fast algorithm is an interesting future work," describing a limitation of the paper titled "Semantic Role Labeling via Tree Kernel Joint Inference" [73]. This sentence was described at the end of the conclusion as a future topic of research of the paper. The patent JP2011039831A contributes to reducing SVM computational complexity. The title of this patent is "Re-learning method for support vector machine," which provides a re-learning method for SVM that can improve the accuracy of SVM and reduce the computational complexity by using a small number of high-quality samples of SVM for re-learning. While the patent itself was published in the 2011 Public Gazette, the applicant had published a basic patent titled JP200421590A—"Re-learning method for support vector machine"—in 2004. This suggests that it is possible to solve the problems described as future work in a scientific paper published in 2006 through the industrial-technical level at the time. In other words, this finding indicates that science does not necessarily anticipate technology as represented by the linear model.

Pair #3: The article phrase was extracted from "One drawback of his algorithm is that it requires the proportion of positive and negative examples in the test data be close to the proportion in the training data, which may not always hold, particularly when the training data is small," from the paper "Semi-supervised learning for semantic parsing using support

vector machines" [74]. This sentence points out imbalanced data in the estimation algorithm proposed in [75]. In general, it is desirable that the data sizes of positive and negative examples are balanced in machine learning, especially for small datasets. The problem sentence of the corresponding patent, JP2002133389, proposes a method for improving the accuracy of data classification for test data even when the amount of training data is small. The paper was published in 1999 and the patent application in 2000. The imbalanced data problem was discussed in the artificial intelligence field around 2000. This pair is a good example of the information available at that time that could have been extracted by inter-domain knowledge-sharing and contributed to problem-solving.

Pair #4: The article phrase was extracted from "The bottleneck in Artificial Intelligence is the unbalanced knowledge sources shared by human beings and a computer system" in the paper "Latent Features in Automatic Tense Translation between Chinese and English" [76]. The paragraph with this sentence points out that the data that can be input into artificial intelligence mechanisms constitute only a small part of the data human beings can manage. The corresponding patent, JP20033233798A, provides a system for sharing human experience. At first glance, it seems the only common terms are "human being" and "system." However, the problem at the core of this patent is that feelings and experiences cannot be sufficiently conveyed through the Internet using only text-based document data. Although the context is different, common to each problem are abstract concepts, including insufficient data.

Pair #5: The article phrase was extracted from "The challenge with such analysis is that OSN [Online Social Network] data come with no annotations, and it would be impossible to manually annotate the data for quantitative analysis of self-disclosure" in a paper titled "Self-Disclosure Topic Model for Twitter Conversations" [77]. This problem sentence points out the difficulty of annotating self-disclosure information on OSNs. Data analysis of OSNs such as Twitter had just begun at the time the article was written. This problem sentence appeared in the abstract and turned out to express the essential problem highlighted by the paper. The problem indicated by the corresponding patent JP2010237864A is that the annotation in such social annotation services contains much information unrelated to the essence of the content, so it is necessary to remove it. As to why this pair was extracted, it is clear that "annotation" is a common word in each sentence. It is also interesting that the common issue of social service was unintentionally extracted. Even though the word "social" does not appear in the patent problem sentence, the common context can be extracted. Thus, a patent published in 2010 dealt with a technical problem that would have provided a clue to the essential problem highlighted by a paper published in 2014.

Although such an evaluation must be qualitative, we confirmed in several pairs that certain knowledge is likely to be obtained from a patent whose problem corresponds to one raised by scientific research.

V. CONCLUSIONS

Science and technology research involves the exploration and exploitation of knowledge. Scientific research strongly emphasizes "exploration" in the pursuit of new knowledge, while industrial technology has strongly emphasized the "exploitation" of the existing knowledge. However, it is natural in complex innovation processes that scientific knowledge involves new exploration through the exploitation of industrial technology. Gardner [78] defines four concepts regarding the relationship between science and technology: 1) the "demarcation view" when both are considered independent, 2) the "idea state view" when science development precedes technology development, 3) the "materialist view" when the technology development precedes science development, and 4) the "interaction model" when science and technology develop interactively.

In this study, we demonstrated the practicality of extracting problem sentences based on a language model and thus linking scientific and industrial knowledge. We collected data from scientific articles and patent publications related to information science (the natural language processing field). We proposed a model to extract problem-related phrases and confirmed that it shows higher performance than the existing models, especially for scientific articles. Clustering was performed on each extracted problem sentence for both scientific articles and patent publications to categorize and map these problems. By determining the similarity between the paper and patent problem sentences, we extracted pairs with the same problem consciousness. After examining some of the pairs with high similarity, we could understand not only the reason for the common words but also the essential background of the problem. This approach showed the possible insight to be gained that would be difficult to obtain with only a keyword search.

This research has several limitations. First, we did not fully consider the publication year of each of the papers and patents included in this study. For example, in Pair #3, the scientific article was published in 2006 and the patent in 2011. Thus, the patent presents information five years after the problem was described in 2006. In particular, knowledge is updated quickly in the information technology field and information becomes obsolete in about a year. We think this issue can be addressed by taking related information from documents published in the same year. For this, a sufficient dataset is required. We did not deal with this issue in this study, but it is essential for future work in this area. There is also room for improvement regarding the length mismatch of extracted problem sentences. Information from some articles is extracted as phrase units, while patents have relatively long sentences, which would not be appropriate to compare. We also want to improve the method of calculating similarity. Since a common word string is obtained for both sides, this gives a high degree of similarity, so it is arguable whether word-based extraction is necessarily appropriate. To capture a problem's essence, a good approach may be to consider the similarity of collocations with a series of functions represented by SAO as shown in Section I.

Although there are several points to be improved, this research makes an important contribution. It developed a

practical model for identifying problem sentences from scientific papers and a method of utilizing the perspective of technology management. We showed the possibility of solving problems in scientific research by finding issues common to both science and industrial technology. Identifying issues in science also contributes to the identification of important research topics, which can lead to insights into scientific trends. In addition, by clarifying the problems in industrial technology, it is possible to identify future targets for business.

ACKNOWLEDGMENT

We would like to thank Editage (www.editage.com) for the English language editing.

REFERENCES

[1] C. Freeman, "The economics of technical change," *Cambridge Journal of Economics*, vol. 18, no. 5, pp. 463–514, 1994.

[2] H. Grupp, *Foundations of the Economics of Innovation: Theory, Measurement and Practice*, Illustrated edition edition. Cheltenham:UK: Edward Elgar, 1998.

[3] G. Dosi, *Innovation, Organization and Economic Dynamics: Selected Essays*. Cheltenham:UK: Edward Elgar, 2000.

[4] B. Godin, "The Linear Model of Innovation: The Historical Construction of an Analytical Framework," *Science, Technology, & Human Values*, vol. 31, no. 6, pp. 639–667, Nov. 2006, doi: 10.1177/0162243906291865.

[5] D. Edgerton, "'The linear model' did not exist: Reflections on the history and historiography of science and research in industry in the twentieth century," in *The Science-Industry Nexus: History, Policy, Implications.*, 2004, pp. 1–36.

[6] D.A. Hounshell, "Industrial research: Commentary", in: *The Science-Industry Nexus. History, Policy, Implications, Science History Publications*, 2004, pp. 59-68

[7] National Science Foundation (U.S.), *Basic research; a national resource.* Washington, DC, USA, 1957.

[8] R. R. Nelson, "The Simple Economics of Basic Scientific Research," *Journal of Political Economy*, vol. 67, no. 3, pp. 297–306, Jun. 1959, doi: 10.1086/258177.

[9] W. J. Price and L. W. Bass, "Scientific Research and the Innovative Process," *Science*, vol. 164, no. 3881, pp. 802–806, 1969.

[10] S. J. Kline, "Innovation Is Not a Linear Process," *Research Management*, vol. 28, no. 4, pp. 36–45, Jul. 1985, doi: 10.1080/00345334.1985.11756910.

[11] R. Landau and Rosenberg, Eds., *The Positive Sum Strategy: Harnessing Technology for Economic Growth*. Washington, DC, USA: The National Academies Press, 1986.

[12] N. Rosenberg, *Exploring the Black Box: Technology, Economics, and History*. Cambridge, UK: Cambridge University Press, 1994.

[13] K. Grandin, N. Wormbs, and S. Widmalm, *The Science-industry Nexus: History, Policy, Implications : Nobel Symposium 123*. USA: Science History Publications, 2004.

[14] N. Rosenberg, *Inside the Black Box: Technology and Economics Paperback*. Cambridge, UK: Cambridge University Press, 1982.

[15] M. Gibbons, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Thousand Oaks, CL, USA: SAGE, 1994.

[16] A. Verbeek, K. Debackere, M. Luwel, P. Andries, E. Zimmermann, and F. Deleus, "Linking science to technology: Using bibliographic references in patents to build linkage schemes," *Scientometrics*, vol. 54, no. 3, pp. 399–420, 2002.

[17] W. E. Steinmueller, "Basic Research and Industrial Innovation," in *The Handbook of Industrial Innovation*, Cheltenham:UK: Edward Elgar, 1995.

[18] S. J. Kline, *Innovation Styles in Japan and the United States: Cultural Bases : Implications for Competitiveness : the 1989 Thurston Lecture.* Stanford University, Department of Mechanical Engineering, Thermosciences Division, 1990.

[19] M. B. Myers and R. S. Rosenbloom, *Rethinking the Role of Industrial Research*. Division of Research, Harvard Business School, 1994.

[20] M. Balconi, S. Brusoni, and L. Orsenigo, "In defence of the linear model: An essay," *Research Policy*, vol. 39, no. 1, pp. 1–13, Feb. 2010, doi: 10.1016/j.respol.2009.09.013.

[21] F. Narin and D. Olivastro, "Status report: Linkage between technology and science," *Research Policy*, vol. 21, no. 3, pp. 237–249, Jun. 1992, doi: 10.1016/0048-7333(92)90018-Y.

[22] F. Narin and D. Olivastro, "Linkage between patents and papers: An interim EPO/US comparison," *Scientometrics*, vol. 41, no. 1, pp. 51–59, Jan. 1998, doi: 10.1007/BF02457966.

[23] F. Narin, M. Rosen, and D. Olivastro, "Patent Citation Analysis: New Validation Studies and Linkages Statistics," *Science and Technology Indicators*, pp. 35–47, Jan. 1989.

[24] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2008.

[25] M. Meyer, "Tracing knowledge flows in innovation systems," *Scientometrics*, vol. 54, no. 2, pp. 193–212, Jun. 2002, doi: 10.1023/A:1016057727209.

[26] J. Callaert, B. Van Looy, A. Verbeek, K. Debackere, and B. Thijs, "Traces of Prior Art: An analysis of non-patent references found in patent documents," *Scientometrics*, vol. 69, no. 1, pp. 3–20, Oct. 2006, doi: 10.1007/s11192-006-0135-8.

[27] F. Narin, K. S. Hamilton, and D. Olivastro, "The increasing linkage between U.S. technology and public science," *Research Policy*, vol. 26, no. 3, pp. 317–330, 1997.

[28] M. P. Carpenter and F. Narin, "Validation study: Patent citations as indicators of science and foreign dependence," *World Patent Information*, vol. 5, no. 3, pp. 180–185, Jan. 1983, doi: 10.1016/0172-2190(83)90139-4.

[29] W. Glanzel and M. Meyer, "Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations," *Scientometrics*, vol. 58, no. 2, pp. 415–428, Oct. 2003, doi: 10.1023/A:1026248929668.

[30] F. Narin and E. Noma, "Is technology becoming science?," *Scientometrics*, vol. 7, no. 3, pp. 369–381, Mar. 1985, doi: 10.1007/BF02017155.

[31] T.-K. Hsiao and V. Torvik, "Knowledge transfer from technology to science: The longevity of paper-to-patent citations," *Proceedings of the Association for Information Science and Technology*, vol. 56, pp. 417–421, Jan. 2019, doi: 10.1002/pra2.41.

[32] R. Johnson, A. Watkinson, and A. Mabe, *The STM Report: An overview of scientific and scholarly publishing*, 5th ed. Hague, Netherlands: International Association of Scientific, Technical and Medical Publishers, 2018.

[33] *World Intellectual Property Indicators 2019*. Geneva, Switzerland: World Intellectual Property Organization, 2019.

[34] D. Swanson, N. Smalheiser, and V. Torvik, "Ranking indirect connections in literature-based discovery: The role of medical subject headings," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1427–1439, Sep. 2006, doi: 10.1002/asi.20438.

[35] M. Weeber, H. Klein, L. Berg, and R. Vos, "Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 7, pp. 548–557, May 2001, doi: 10.1002/asi.1104.abs.

[36] D. Hristovski, B. Peterlin, J. Mitchell, and S. Humphrey, "Using literature-based discovery to identify disease candidate genes," *International Journal of Medical Informatics*, vol. 74, pp. 289–298, Nov. 2004, doi: 10.1016/j.ijmedinf.2004.04.024.

[37] M. D. Gordon and R. K. Lindsay, "Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on

literature-based discovery of a connection between Raynaud's and fish oil," *Journal of the American Society for Information Science*, vol. 47, no. 2, pp. 116–128, Feb. 1996, doi: 10.1002/(SICI)1097-4571(199602)47:2%3C116::AID-ASI3%3E3.3.CO;2-P.

[38] D. R. Swanson, "Undiscovered public knowledge," *The Library Quarterly: Information, Community, Policy*, vol. 56, no. 2, pp. 103–118, Apr. 1986.

[39] V. Ittipanuvat, K. Fujita, Y. Kajikawa, J. Mori, and I. Sakata, "Finding linkage between technology and social issues: A literature based discovery approach," in *2012 Proceedings of PICMET '12: Technology Management for Emerging Technologies*, Vancouver, BC, Canada, Aug. 2012, pp. 2310–2321.

[40] N. Shibata, Y. Kajikawa, and I. Sakata, "Extracting the commercialization gap between science and technology — Case study of a solar cell," *Technological Forecasting and Social Change*, vol. 77, no. 7, Sep. 2010, doi: 10.1016/j.techfore.2010.03.008.

[41] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, no. 11, pp. 758–775, Nov. 2008, doi: 10.1016/j.technovation.2008.03.009.

[42] M.-Y. Wang, S.-C. Fang, and Y.-H. Chang, "Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels," *Technological Forecasting and Social Change*, vol. 92, Aug. 2014, doi: 10.1016/j.techfore.2014.07.008.

[43] M. Meyer, "Does science push technology? Patents citing scientific literature," *Research Policy*, vol. 29, no. 3, pp. 409–434, Mar. 2000, doi: 10.1016/S0048-7333(99)00040-2.

[44] M. Gittelman and B. Kogut, "Does Good Science Lead to Valuable Knowledge? Biotechnology Firms and the Evolutionary Logic of Citation Patterns," *Management Science*, vol. 49, no. 4, pp. 366–382, Apr. 2003, doi: 10.1287/mnsc.49.4.366.14420.

[45] L. Sollaci and M. Pereira, "The Introduction, Methods, Results, and Discussion (IMRAD) Structure: a fifty-year survey," *Journal of the Medical Library Association : JMLA*, vol. 92, no. 3, pp. 364–7, Aug. 2004.

[46] R. D. Huddleston, *Sentence and Clause in Scientific English*. Communication Research Centre, University College, 1968.

[47] M. Hoey, *Textual Interaction: An Introduction to Written Discourse Analysis*. Routledge, 2013.

[48] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, no. 5, pp. 1216–1247, Sep. 2007, doi: 10.1016/j.ipm.2006.11.011.

[49] H. Sakai, H. Nonaka, and S. Masuyama, "Extraction of Information on the Technical Effect from a Patent Document," *Transactions of The Japanese Society for Artificial Intelligence*, vol. 24, pp. 531–540, Jan. 2009, doi: 10.1527/tjsai.24.531.

[50] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama, "Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases," in *Proceedings of the Third NTCIR Workshop*, Tokyo,Japan, Oct. 2002.

[51] I. Bergmann, D. Butzke, L. Walter, J. P. Fuerste, M. G. Moehrle, and V. A. Erdmann, "Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips," *R&D Management*, vol. 38, no. 5, pp. 550–562, 2008, doi: 10.1111/j.1467-9310.2008.00533.x.

[52] G. Cascini, A. Fantechi, and E. Spinicci, "Natural Language Processing of Patents and Technical Documentation," in *Document Analysis Systems VI*, vol. 3163, 2004.

[53] G. Cascini and M. Zini, "Measuring patent similarity by comparing inventions functional trees," in *Computer-Aided Innovation (CAI)*, vol. 277, Springer, 2008.

[54] H. Park, J. Yoon, and K. Kim, "Identifying patent infringement using SAO based semantic technological similarities," *Scientometrics*, vol. 90, no. 2, pp. 515–529, Feb. 2012, doi: 10.1007/s11192-011-0522-7.

[55] J. Yoon and K. Kim, "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection," *Scientometrics*, vol. 90, no. 2, pp. 445–461, Feb. 2012, doi: 10.1007/s11192-011-0543-2.

[56] X. Wang *et al.*, "Identifying R&D partners through Subject-Action-Object semantic analysis in a problem & solution pattern," *Technology Analysis & Strategic Management*, vol. 29, no. 10, pp. 1167–1180, Nov. 2017, doi: 10.1080/09537325.2016.1277202.

[57] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto, "Overview of the Patent Mining Task at the NTCIR-8 Workshop," presented at the Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan, Jun. 2010, pp. 293–302.

[58] M. Iwayama, A. Fujii, and N. Kando, "Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task," presented at the Proceedings of NTCIR-5 Workshop Meeting, Tokyo, Japan, Dec. 2005, pp. 359–365.

[59] K. Heffernan and S. Teufel, "Identifying problems and solutions in scientific text," *Scientometrics*, vol. 116, no. 2, pp. 1367–1382, 2018, doi: 10.1007/s11192-018-2718-6.

[60] I. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: an Open-source CRF Reference String Parsing Package," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008, pp. 661–667.

[61] D. J. Phelps, "Automatic Concept Identification: Extracting Problem Solved Concepts From Patent Documents," presented at the IRFS 2007 Vienna Information Retrieval Facility Symposium, Vienna, Austria, 2007.

[62] S. Tiwana and E. Horowitz, "Extracting problem solved concepts from patent documents," in *Proceedings of the 2nd international workshop on Patent information retrieval*, Hong Kong, China, Nov. 2009, pp. 43–48, doi: 10.1145/1651343.1651356.

[63] C. Jeong and K. Kim, "Creating patents on the new technology using analogy-based patent mining," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3605–3614, Jun. 2014, doi: 10.1016/j.eswa.2013.11.045.

[64] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2–3, pp. 146–162, Aug. 1954, doi: 10.1080/00437956.1954.11659520.

[65] M. Sahlgren, "The distributional hypothesis," *Italian journal of linguistics*, vol. 20, no. 1, pp. 33–54, 2008.

[66] A. Vaswani *et al.*, "Attention Is All You Need," presented at the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[67] A. Radford, "Improving Language Understanding by Generative Pre-Training." 2018, Accessed: Jun. 13, 2020. [Online]. Available: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035, (preprint)

[68] S. Kobayashi, *soskek/chainer-openai-transformer-lm*. 2020.

[69] P. J. Liu *et al.*, "Generating Wikipedia by Summarizing Long Sequences," *arXiv:1801.10198 [cs]*, Jan. 2018, Accessed: Jun. 12, 2020. [Online]. Available: http://arxiv.org/abs/1801.10198.

[70] "Finetune Quickstart Guide — finetune 0.8.3 documentation." https://finetune.indico.io/ (accessed Jun. 13, 2020).

[71] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, Mar. 1963, doi: 10.1080/01621459.1963.10500845.

[72] D. Ravichandran and E. Hovy, "Learning Surface Text Patterns for a Question Answering System," presented at the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia,USA, Jul. 2002, pp. 41–47.

[73] A. Moschitti, D. Pighin, and R. Basili, "Semantic Role Labeling via Tree Kernel Joint Inference," in *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, New York City, Jun. 2006, pp. 61–68.

[74] R. J. Kate and R. J. Mooney, "Semi-Supervised Learning for Semantic Parsing using Support Vector Machines," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Short Papers (NAACL/HLT-2007)*, Rochester, NY, USA, Apr. 2007, pp. 81–84.

[75] T. Joachims, "Transductive Inference for Text Classification using Support Vector Machines," in *Proceedings of the Sixteenth International*

Conference on Machine Learning, San Francisco, CA, USA, Jun. 1999, pp. 200–209.

[76] Y. Ye, V. L. Fossum, and S. Abney, "Latent Features in Automatic Tense Translation between Chinese and English," in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia, Jul. 2006, pp. 48–55.

[77] J. Bak, C.-Y. Lin, and A. Oh, "Self-disclosure topic model for Twitter conversations," in *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, Baltimore, Maryland, USA, Jun. 2014, pp. 42–49.

[78] P. Gardner, "The representation of science-technology relationships in Canadian physics textbooks," *International Journal of Science Education*, vol. 21, no. 3, pp. 329–347, Mar. 1999, doi: 10.1080/095006999290732.

AUTHORS' PROFILES

Hajime Sasaki is an Associate Professor at the Institute for Future Initiatives, The University of Tokyo. He received his Ph.D. degree from the University of Tokyo, and his M.S. degree from the Tokyo Institute of Technology. His research interests include innovation management and data-driven decision-making.

Satoru Yamamoto is the President and CEO of Data Artist Inc., an AI solution company based in Tokyo. He majored in AI at the University of Tokyo. He aims to develop new and innovative AI solutions for various industries such as advertising, medicine, and finance.

Amarsanaa Agchbayar is the CTO of Data Artist Inc. He majored in data mining at the University of Tokyo. He is a bronze medalist in the International Mathematical Olympiad, and is presently leading an engineering team to develop new and innovative AI solutions for various industries.

Nyamaa Enkhbayasgalan is a Data Scientist at Data Artist Inc. He obtained his M.S. degree from Tokyo Institute of Technology by majoring Natural Language Processing. He has been involved in numerous projects utilizing AI technology.