Machine Learning: Science and Technology

PAPER • OPEN ACCESS

ARTS: autonomous research topic selection system using word embeddings and network analysis

To cite this article: Eri Teruya *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 025005

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
10 November 2021

REVISED
15 February 2022

ACCEPTED FOR PUBLICATION
29 March 2022

PUBLISHED
21 April 2022

PAPER

ARTS: autonomous research topic selection system using word embeddings and network analysis

Eri Teruya[1,*], Tadashi Takeuchi[1], Hidekazu Morita[1], Takayuki Hayashi[1] and Kanta Ono[2]

[1] Hitachi, Ltd, 6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo 100-8280, Japan
[2] Department of Applied Physics, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
[*] Author to whom any correspondence should be addressed.

E-mail: eri.teruya.yz@hitachi.com

Keywords: natural language processing, text mining, word embedding, information extraction, relation extraction, materials informatics

Abstract

The materials science research process has become increasingly autonomous due to the remarkable progress in artificial intelligence. However, autonomous research topic selection (ARTS) has not yet been fully explored due to the difficulty of estimating its promise and the lack of previous research. This paper introduces an ARTS system that autonomously selects potential research topics that are likely to reveal new scientific facts yet have not been the subject of much previous research by analyzing vast numbers of articles. Potential research topics are selected by analyzing the difference between two research concept networks constructed from research information in articles: one that represents the promise of research topics and is constructed from word embeddings, and one that represents known facts and past research activities and is constructed from statistical information on the appearance patterns of research concepts. The ARTS system is also equipped with functions to search and visualize information about selected research topics to assist in the final determination of a research topic by a scientist. We developed the ARTS system using approximately 100 00 articles published in the Computational Materials Science journal. The results of our evaluation demonstrated that research topics studied after 2016 could be generated autonomously from an analysis of the articles published before 2015. This suggests that potential research topics can be effectively selected by using the ARTS system.

1. Introduction

Many things in our lives have become more autonomous due to the remarkable progress in technologies such as machine learning and artificial intelligence (AI). Unlike an 'automation' system, which executes operations based on procedures and rules defined by humans, a system that is 'autonomous' can identify procedures or rules, judge objects, execute operations, and further find and invent new things by itself.

Materials science research might become autonomous in the near future. In this scenario, a system will be able to autonomously reveal a new material property, autonomously synthesize materials that show an inquired property, and autonomously clarify the mechanism of the property. Numerous studies have been conducted in this vein, including works on autonomous experiments, autonomous research idea generation, and so on [1–25].

However, most of this prior research has focused on the steps from 'idea construction' to 'experiment & evaluation' in the process of materials science research, as shown in figure 1, and there have not been very many studies on the 'research topic selection' or 'report' steps. In particular, the autonomous 'research topic selection' (ARTS) step has not been studied because it is quite complex to formulate from a mechanical perspective and it is difficult to estimate whether a research topic is likely to succeed. However, this step is of crucial importance for the eventual realization of fully autonomous materials science research because it is the first step of the materials science research process, and selecting a good research topic that has high possibility to reveal new scientific facts can accelerate the progress of science efficiently.

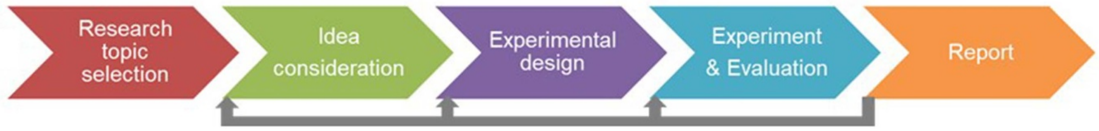


Figure 1. Typical materials science research process flow. This figure is drawn from [28–30] and our own research process.

Here, we review a few of the articles related to ARTS. Krenn and Zeilinger tried to predict research topic trends by analyzing a network constructed of research concepts from a large number of articles [23]. In their study, a concept network in the quantum physics field is constructed from nodes that correspond to chemical concepts and links that link two nodes when two concepts are concurrently studied in articles . They predict future research topics by link prediction so that new links can be formed between unconnected vertices of the network in the future given the current state of the network. However, their method only predicts future research from past research trends, and it is not clear whether it can select good research topics that will reveal scientific facts.

Another study was conducted by Tshitoyan *et al* [22], who showed that chemical insights to be used as hints for selecting research topics can be extracted from large numbers of articles by using word embedding, a method utilized to vectorize words by learning the co-occurrence of words from prior literature [26, 27]. They trained the abstracts of millions of scientific articles published up to a specific year (year X) to obtain word vectors and then extracted material names (e.g. 'thermoelectric') with small distances between word vectors. The results showed that some of the extracted materials were reported to have the 'thermoelectric' character after year X, and the reporting rate was higher than that of randomly extracted materials. Their work is essentially based on word pairs (e.g. pairs of the 'thermoelectric' character and material names) that show scientific facts and that are selected as research topics. However, since their method outputs many word pairs that have already been researched , scientists are required to investigate whether an extracted word pair has already been studied.

The two methods above assume that a system presents various research topics and that a scientist will ultimately determine which of them should actually be studied. Therefore, they require additional functions to help researchers search for topics and visualize information on them.

In the present work, we focus on the issue of ARTS to enable fully autonomous materials science research. As the first step, we propose the ARTS system, which we have designed on the basis of research topic selection processes by skilled scientists. The ARTS system enables the autonomous selection of potential research topics with high promise, which is likely to reveal scientific facts, and is one of the studies to combine the word embedding method proposed by Tshitoyan *et al* and the concept network proposed by Krenn and Zeilinger. The proposed ARTS system is also equipped with functions to search and visualize information about autonomously selected research topics to assist scientists in determining which topic to research.

In section 2 of this paper, we specify the definitions of research topic and research topic selection and introduce the basic design of the ARTS system. Section 3 discusses the processes of the ARTS system in detail. In section 4, we evaluate the ARTS system using materials science articles and present the results. We conclude in section 5 with a brief summary and mention of future work.

2. Basic design of ARTS system

We first define the research topic and ARTS methods used in this study. Second, we review the research topic selection process of a skilled scientist to clarify the processes of research topic selection. Finally, we present our design of the ARTS system based on the research topic selection process of skilled scientists.

2.1. Definition of a research topic and research topic selection

We define a research topic as a combination of research concepts. Here, a research concept is a keyword that represents research: for example, a material name (e.g. 'organic superconductor' or 'MgSn compound'), a chemical property name (e.g. 'thermoelectric'), or a model name (e.g. 'density functional theory'). We define ARTS as a state in which the system autonomously outputs a combination of research concepts.

To check the validity of the definition of research topics, we examined the paper titles, as titles typically state the research topic in a straightforward manner. As an example, take the title of a paper by Drs. Emmanuelle Charpentier and Jennifer Anne Doudna, who won the Nobel Prize in Chemistry in 2020, which is 'A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity' [31]. Their research topic can be understood as a combination of research concepts characterized by the enzyme name ‘programmable dual-RNA-guided DNA endonuclease’ and the property name ‘bacterial immunity’. As another example, the title of the current paper is ‘ARTS: Autonomous Research Topic Selection System using Word Embeddings and Network Analysis’, and our research topic can be understood as a combination of research concepts characterized by the system name ‘Autonomous Research Topic Selection System’ and the technical names ‘Word Embeddings’ and ‘Network Analysis’. In this way, a research topic is represented as a combination of multiple research concepts. We define ARTS as a system that outputs a good combination of research concepts. Note that our system is not designed to produce completely new research concepts that have never been used in past articles.

2.2. Research topic selection process of a skilled scientist

We reviewed the research topic selection process of a skilled scientist and clarified the flow as follows: the extraction of research concepts and research topic candidates, the calculation of information criteria for selection (the research topic index), and the selection of research topics based on the research topic index.

In the first step, the scientist selects research topic candidates based on a research concept as follows.

(a) The scientist searches for articles using an article search system utilizing the concepts in which he or she is interested as search keywords.
(b) The scientist reads the retrieved articles, and if additional interesting concepts are found, the article search is performed again using those concepts.
(c) The scientist repeats this process until interesting combinations of concepts he or she would like to study are found. These become the research topic candidates.

After the research topic candidates are selected, the scientist investigates or estimates various research topic indexes for each one, and if an index is deemed to be good, this one is selected as the research topic. Here, the research topic index is a criterion for research topic selection. Frequently considered research topic indexes include (1) the promise of the research, (2) the existence of previous research, (3) research trends, and (4) the importance of the research topics.

The promise of a research topic is an index of whether the research topic is likely to reveal new scientific facts. Obviously, scientists want to select a research topic that seems promising, but accurately estimating promise is generally impossible. This is why they generally estimate it based on their intuition and experience. The index of the existence of previous research is helpful in this regard because it shows whether a research topic has already been studied; if a research topic is not connected to any previous research, this is generally the one that is selected. Scientists investigate this index based on article retrieval or citation searches. Examples of trend indexes include the evolution of related research on a particular research topic, the progress made by researchers who have studied a research topic in the past who may be competitors, and the research trends in countries that focus on a particular research topic. Scientists generally investigate these trends by reading articles. The index of importance indicates whether a research topic is essential to scientists or society, i.e. the potential impact on the research community and society, practicality, and so on. It is estimated on the basis of the interests of scientists, the global situation, the needs of society, and the policies of research institutions. By considering each of these research topic indexes, scientists can finally determine their next research topic.

In summary, the following three processes are required to achieve ARTS:

(a) The extraction of concepts and research topic candidates from articles.
(b) The calculation of research topic indexes (e.g. promise, existence of prior research, trends, and importance).
(c) The selection of a research topic based on research topic indexes.

2.3. Autonomous research topic selection process in the ARTS system

We designed the basic process of the ARTS system with reference to the research topic selection process of skilled scientists derived in section 2.2. The basic design is shown in figure 2. Research concepts are first extracted from articles and then research topic candidates are created by combining multiple concepts. Second, the research topic index is calculated for each research topic candidate. We calculate three types of research topic index: (1) promise, (2) existence of previous research, and (3) trends. Third, the system outputs and visualizes the research topics based on the research topic index. Specifically, research topics that have a high promise index but a low index of existence in previous studies are output as potential research topics. In the case of trend indexes, the selection criteria are highly dependent on the scientist’s preferences or research strategy. For example, one scientist may select research topics that are recently popular, while another may intentionally select niche research topics. Therefore, the system simply visualizes the trend

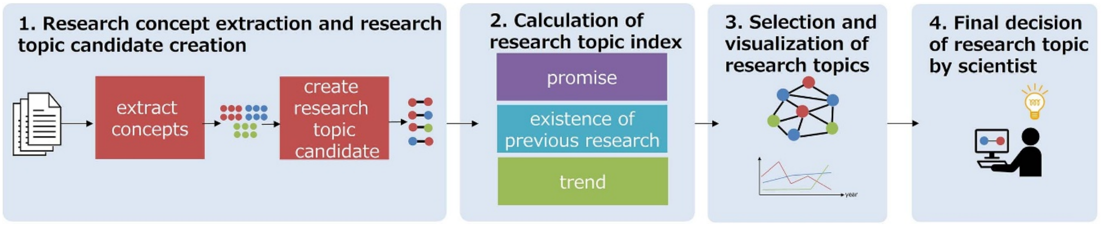

1. Research concept extraction and research topic candidate creation
extract concepts
create research topic candidate
2. Calculation of research topic index
promise
existence of previous research
trend
3. Selection and visualization of research topics
year
4. Final decision of research topic by scientist

Figure 2. Process of autonomous research topic selection in the ARTS system.

indexes so that scientists can choose. We also visualize additional information (e.g. article information) to assist with research topic extraction. A researcher makes the final decision on the research topic that he or she will conduct by selecting from the research topics output by the system.

3. Methodology

In this section, we describe how to implement the ARTS system described in section 2.3. An overview of the system is provided in figure 3. First, research concepts are extracted from a large number of articles using named entity recognition (NER) technology [32], and research topic candidates are created. Next, two research concept networks are constructed to calculate the research topic index. The first, which we call the embedded research concept network, is a network of promising research topics constructed from word vectors obtained by the word embedding of research concepts. The second, called the known research concept network, is a network that represents the known facts calculated from the statistics of the appearance patterns of research concepts. Finally, the system selects and visualizes the potential research topics. Research topics that have a high promise index and a low existence in previous research index are selected as potential research topics. The system also visualizes the transition of the trend index.

3.1. Research concept and research topic candidate extraction

Research concepts that represent research in materials science are extracted from articles using both the existing NER techniques and our newly developed NER algorithm. We extract concepts that belong to three categories: material names, chemical property names, and model names. The material names are extracted using ChemDataExtractor [33], and the other concepts are extracted using our NER algorithm.

Our NER algorithm takes the distant supervised learning approach [31] and is equipped with a user feedback function (described below). It can achieve a good concept extraction performance with little human cost. In a typical NER algorithm, the training data are sentences labeled with concepts in the inside-outside-beginning (IOB) format [34]. An annotator is required to read all the sentences and label the concepts, which is extremely time consuming. In contrast, our NER model utilizes distant supervised learning, which means the user only needs to prepare a list of concepts for each category as training data, which significantly reduces the preparation time. This approach has been utilized frequently in recent years; for example, one study applied it to the entity extraction of chemical substances [35]. The user feedback function is implemented to improve the concept extraction performance by utilizing the feedback from users about the extraction results. When the amount of training data is small, the extraction performance is often low, so our NER algorithm obtains feedback from users on the extraction results and then retrains the extraction model on the basis of this feedback, thus improving the extraction performance with a lower human cost.

Our NER algorithm consists of four steps: candidate selection, feature extraction, learning and inference, and feedback. In the candidate selection step, we extract concept candidates from articles based on part-of-speech sequences. A concept consists of one or more adjectives and/or nouns extracted as concept candidates. The part-of-speech sequences are obtained using Stanford CoreNLP [36], which is an NLP toolkit for English. In the feature extraction step, features that characterize each candidate are extracted. These features include the part-of-speech tag, the lemma of the word, the word(s) that make up the mention, the suffix, and the lemma around the concept on both the left and the right of the mention. In the learning and inference step, the factor graph method, which is a probabilistic graphical model described in the literature [37], is used. A factor graph is generated from the articles and the weights of the factor graph are learned from the training data. In the feedback step, our NER algorithm accepts feedback from the user and retrains. After the learning and inference step, the concept candidates and the probability of becoming a concept in each category are output. The concept candidates are presented to the user, who then inputs a

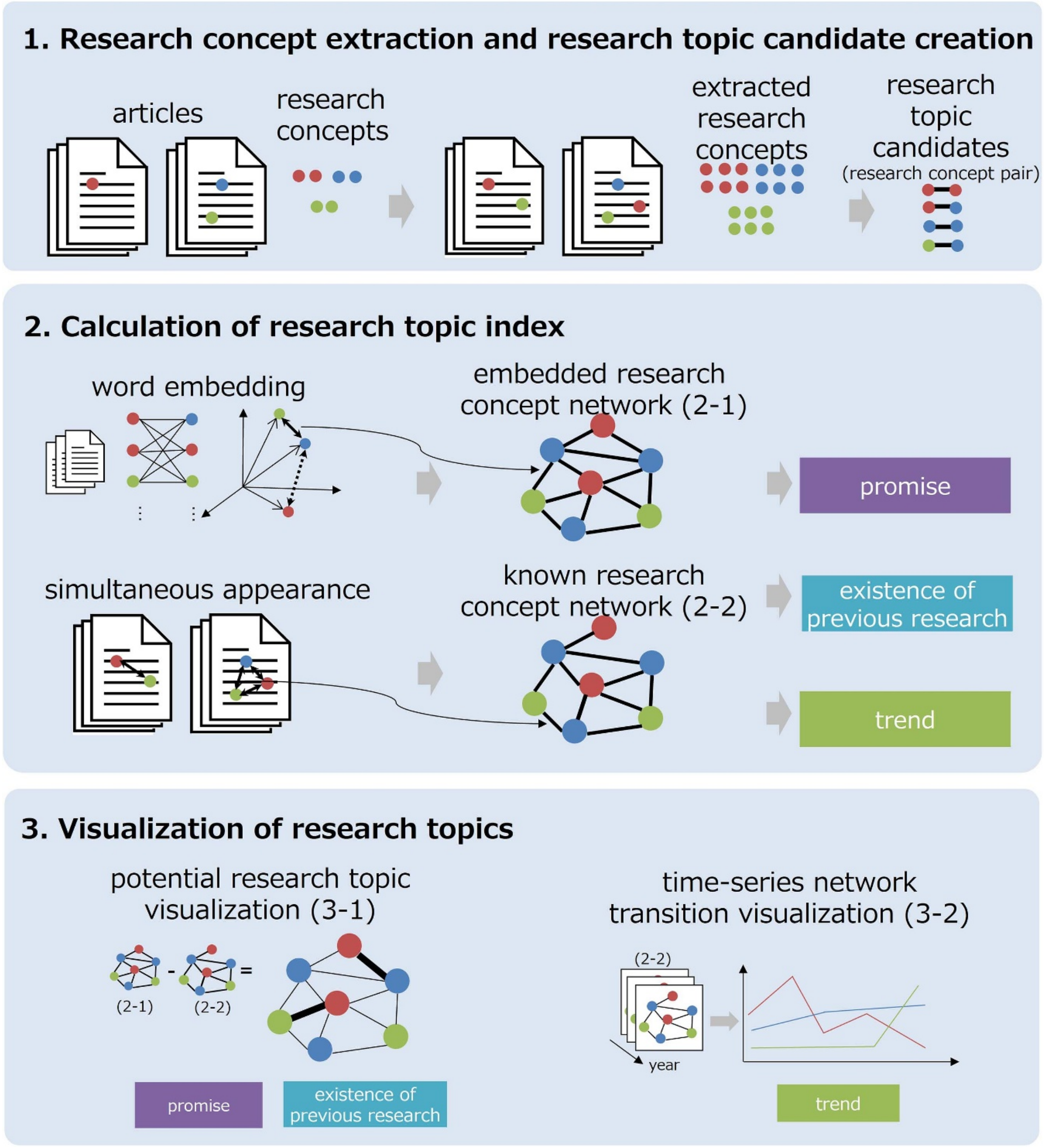

1. Research concept extraction and research topic candidate creation
articles
research concepts
extracted research concepts
research topic candidates
(research concept pair)
2. Calculation of research topic index
word embedding
embedded research concept network (2-1)
promise
simultaneous appearance
known research concept network (2-2)
existence of previous research
trend
3. Visualization of research topics
potential research topic visualization (3-1)
(2-1)
(2-2)
promise
existence of previous research
time-series network transition visualization (3-2)
(2-2)
year
trend

Figure 3. Overview of ARTS system.

judgment as to whether the concept candidate is valid or not. After obtaining the user's feedback, the NER model is retrained accordingly. The feedback step is repeated several times to improve the concept extraction performance.

Finally, research topic candidates are extracted. We define a research topic candidate as all combinations of concept pairs extracted in the concept extraction step.

3.2. Research concept networks and calculation of the research topic index

Three types of research topic index are calculated: metrics of promising research, existence of previous research, and trends. These are calculated using the embedded research concept network and the known research concept network.

3.2.1. Embedded research concept network

The embedded research concept network is a crucial part of the ARTS system. The nodes of the network are the concepts extracted in the research concept extraction step. These nodes are connected on the basis of the embedded research relevance calculated for each research topic candidate, namely, the distance between two word vectors obtained by word embedding [26, 27]. If the distance between the two vectors is small, the two nodes are connected on the embedded research concept network.

The distance between two concept vectors represents the promise of a research topic, and the embedded research concept network is considered a network that represents the promising research topics. As described in the Introduction, Tshitoyan *et al* showed that chemical insights could be extracted on the basis of the distance between two embedded word vectors [22], which suggests that promising research topics can be extracted from embedded words. Therefore, we define the index of a promising research topic as the distance between the vectors of two embedded words, and construct the embedded research concept network by connecting nodes with small distances in the vector space.

The concepts are embedded using word2vec [26, 27], which was also used in Tshitoyan's paper. The distance between two word vectors is calculated with the cosine distance, which is an inner product of two vectors. In other words, the embedded research relevance, $R_p(k_i \to k_j)$, is calculated as:

$$R_p(k_i \to k_j) = cosine(v_i \to v_j), \tag{1}$$

where v_i is the word-vector representation of concept k_i. If $R_p(k_i \to k_j) > Th_p$ and $R_p(k_i \to k_j)$ is in the top 100, k_i and k_j are connected through network links. Th_p is a threshold of the embedding research relevance, which we set as $Th_p = 0.5$ in this study so as to reduce the computational complexity.

3.2.2. Known research concept network

As discussed in section 3.2.1, the distance between two concept vectors is regarded as the promise of a research topic, and the embedded research concept network is a network that represents this promise. However, this network is likely to represent research topics that are already known, since they are trained from previous articles. Therefore, we need to remove these research topics.

This is done by using the known research concept network, which is constructed from concept nodes extracted in the concept extraction step and from links that are connected based on the known research relevance calculated for each research topic candidate. The known research relevance is calculated from the statistics of the appearance patterns of two research concepts: specifically, from the number of articles in which the two concepts are contained in a single article. If two concepts are contained in many articles, the known research relevance will be high, and the two concepts are connected on the known research concept network. Since known facts are mainly contained in articles, the known concept network is essentially a network that represents the known facts and the research activities to date; in other words, it shows research topics that have already been studied. In this study, we utilize the Apriori algorithm, for which multiple libraries exist, to calculate the known research relevance as follows.

All articles used in system D consist of D_t, where $D = \{D_1, D_2, \cdots, D_N\}$ and D_t is the set of articles published in a specific year t. D_t consists of each article d_{ti}, where $D_t = \{d_{t1}, d_{t2}, \cdots, d_{tM_t}\}$. Here, d_{ti} is the set of words $W_{ti} = \{k_{t1}, k_{t2}, \cdots, k_{tk}, \cdots, w_{t1}, w_{t2}, \cdots, w_{tl}\}$, where k_{tj} is a concept extracted by the concept extraction step in section 3.1, and w_{tj} is a word that is not extracted, i.e. a word that is not a material name, a chemical property name, or a model name. M_t is the total number of all articles examined.

The known research relevance between k_{it} and k_{jt} is calculated as:

$$R_{ot}(k_{it} \to k_{jt}) = \sigma(k_{it} \cup k_{jt})/M_t, \tag{2}$$

where $\sigma(X)$ is the number of d_{it} containing X, and $R_{ot}(k_{it} \to k_{jt})$ is calculated for all research topics. Known research concept networks are constructed for each year t. The nodes of the network are k_{it}, and if $R_{ot}(k_{it} \to k_{jt}) > Th_o$, k_{it} and k_{jt} are connected as network links. Th_o is a threshold of the known research relevance, which we set as $Th_o = 0.0$ in this study.

3.3. Calculation of research topic index

Research topic indexes are calculated from two concept networks. For the promise index of the research topic, the distance between the vectors of two concepts represents the promise. Therefore, $R_p(k_i \to k_j)$ is regarded as the promise of a research topic.

For the existence of previous research index, we assume that the frequency of two concepts in one paper is an index of prior research. We define $R_o(k_i \to k_j)$, which is an index of the existence of prior research, as follows:

$$R_o(k_i \to k_j) = \sum_t R_{ot}(k_{ti} \to k_{tj}). \tag{3}$$

We define the research trend index as a change in the known research relevance. The trend that a scientist wants to know is the chronological transition of information related to research topics. Therefore, we define the trend index as a transition concept that is related to a concept. The known research relevance

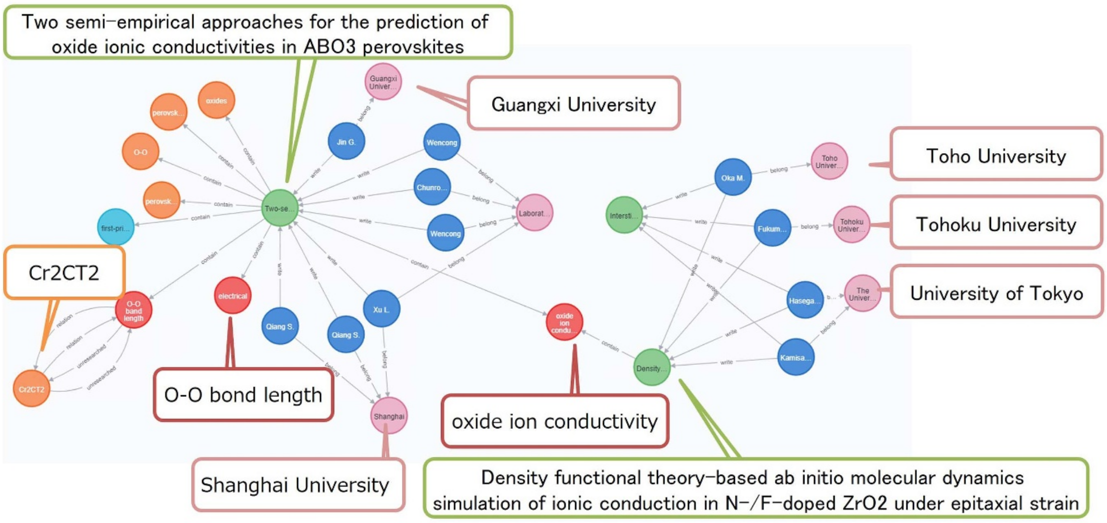


Figure 4. Image of potential research topic visualization. Nodes (circles) represent extracted concepts (orange, light blue, and red), titles of papers (green), authors (blue), and affiliations (pink). Arrows represent relations between nodes. Relation name is written on each arrow.

$R_{ot}(k_{it} \to k_{jt})$ is calculated as a function of the year. In addition to the known research relevance among concepts, we calculate the known research relevance among authors, countries, and funding of papers with the same metrics.

3.4. Visualization of research topics

One of the key points of the ARTS system is the visualization function, which visualizes potential research topics and the trend index. Some information about the research topics is also visualized to assist scientists with their final research topic determination.

3.5. Visualization of potential research topics

The potential research concepts are selected and visualized by the potential research concept visualization function. The embedded research concept network and some article information related to research topics are also visualized on one screen. Figure 4 shows an example of the visualization.

For simplicity, we first describe the visualization of the embedded research concept network, which is stored and visualized in Neo4j [38]. Neo4j is a graph database platform that provides data management for graph structure data, a search interface, and a visualization function. We utilize Neo4j because its visibility and search functions are superior. In figure 4, orange circles, light blue circles, and red circles represent material names, chemical property names, and model names, respectively. The nodes of the embedded research concept network are connected by *concept links*.

Next, we describe the visualization of article information related to research topics. The article title from which research concepts are extracted, the authors of the article, and the author affiliations are visualized as a node with various metadata, as respectively indicated by the green, blue, and pink circles in figure 4. The metadata of each node is listed in table 1. For example, each article title node is visualized with metadata such as doi, URL, journal, year of publication, number of citations, keywords attached by the author when the article was submitted, and funding for the articles. Nodes of concepts, article titles, authors, and affiliations are connected by links, which are listed in table 2. For example, article title nodes and concept nodes are connected by *contain links*. Scientists can determine the relation between information by observing these nodes and links.

Finally, we describe the visualization of potential research concepts. We regard concept pairs that satisfy the conditions $R_p(k_i \to k_j) > Th_p$ and $R_o(k_i \to k_j) > Th_o$ as a potential research topic and connect these pairs with *potential links*. Scientists can determine potential research topics by observing concept pairs that are connected to *potential links*.

3.6. Visualization of a time-series network transition

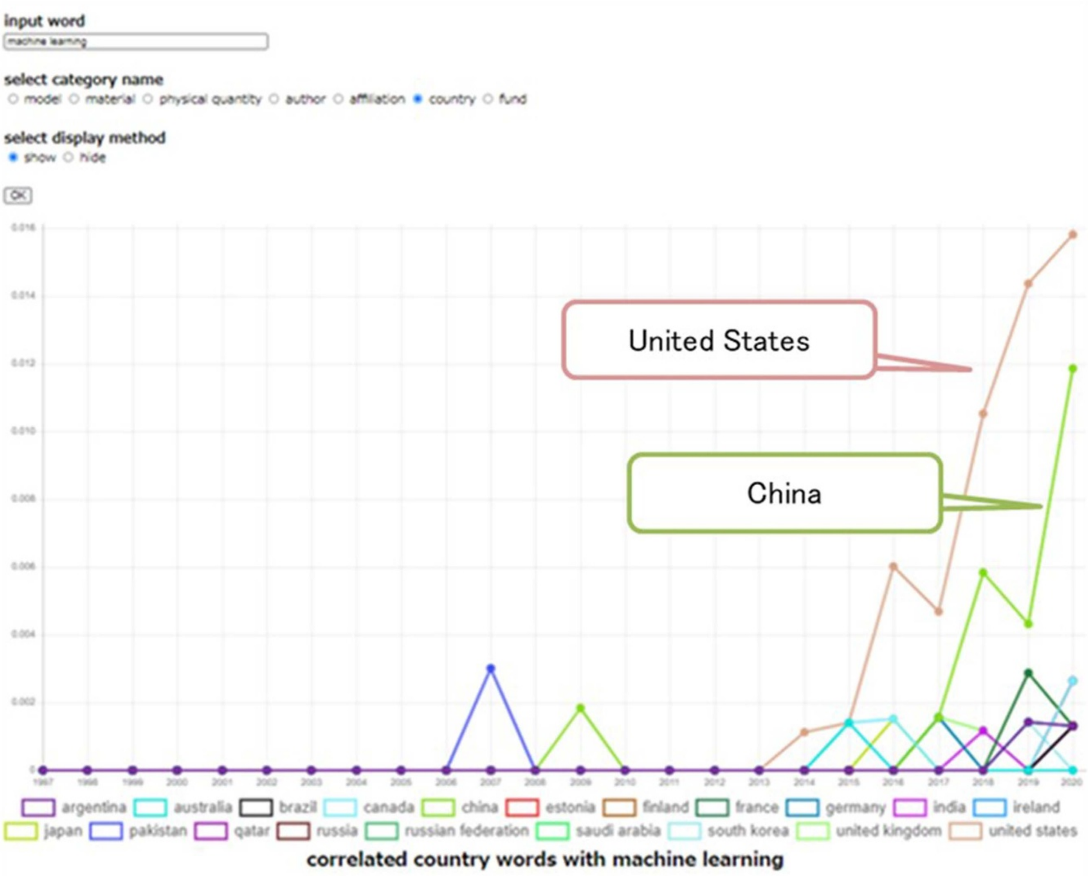
The trend index is visualized as a time-series network transition. We developed a web application to visualize it, with an image of the visualization shown in figure 5. If a scientist inputs a research concept (which

Table 1. Nodes and metadata visualized in the system.

No.	Node	Metadata
1	Material (concepts)	=
2	Model (concepts)	=
3	Chemical property (concepts)	=
4	Article title	doi, url, journal, published year, citation count, author attached concept, funding
5	Author	scopus author ID, url
6	Affiliation	country, city, ID, url

Table 2. Links visualized in the system.

No.	Link	Linked node
1	Concepts	Between concept nodes
2	Contain	Between concept nodes and paper nodes
3	Write	Between paper nodes and author nodes
4	Belong	Between author nodes and affiliation nodes
5	Potential	Between concept nodes



input word

machine learning

select category name

○ model ○ material ○ physical quantity ○ author ○ affiliation ● country ○ fund

select display method

● show ○ hide

OK

United States

China

Figure 5. Image of time-series network transition visualization. Countries associated with the research concept of 'machine learning' are shown.

corresponds to k_{it}) in the input word form on the web application, the known research relevance $R_{ot}(k_{it} \rightarrow k_{jt})$ is displayed as a function of the year for each k_{jt} on the screen. The displayed $R_{ot}(k_{it} \rightarrow k_{jt})$ can be filtered by category by selecting an option from the select category button.

Table 3. Nodes and metadata visualized in the system.

	Material	Model	Chemical property	Total
No. of independent concepts	5611	5390	5425	16 426
Total no. of concepts	23 965	13 125	12 450	49 540

Table 4. Number of links in embedding relation network. The total number of links is 1753.

No. of links	Material	Model	Chemical property
Material	959	41	70
Model	=	455	14
Chemical property	=	=	214

Table 5. Number of links in known research network. The total number of links is 35 664.

No. of links	Material	Model	Chemical property
Material	15 076	5519	4701
Model	=	3804	3062
Chemical property	=	=	3502

4. Results and discussion

We evaluated the proposed ARTS system using articles published in Computational Materials Science [39]. We collected titles, abstracts, and metadata of 10 205 articles from 1997 to 2021 and implemented them in the ARTS system using the method described in section 3.

4.1. Network features

First, we investigated the features of the two networks. The number of extracted independent research concepts and the total number of research concepts are shown in table 3. From the difference between the two, we found that, on average, 2.3 research concepts corresponded to a material name, 1.3 to a model name, and 1.2 to chemical property names that were contained in the title and the abstract of one article, which is consistent with our expectation.

The number of links in the embedding research network between categories is shown in table 4. The number of links between the same categories was large because research concepts with the same category were used in a similar way in sentences ; thus, the similarity between two word vectors was high. The total number of links in the known research concept network between the research concepts of each category is shown in table 5. We found that 1.5 links between materials were contained in one article, which is almost consistent with the fact that the number of research concepts corresponding to a material name contained in one article was 1.2. On the other hand, the numbers of links between other categories were small in comparison with the numbers of links between materials. This means that the same relations appeared in several articles.

We also investigated the distribution of the number of research concepts compared with the embedding research relevance and known research relevance, as shown in figure 6. We can see that the number of research concepts rapidly increased as $R_p(k_i \rightarrow k_j)$ decreased. The number of research concept candidates as a function of the known research relevance is shown in figure 7. Most research concepts were in a value range of 0–0.01, which means that most of them appeared in one or two articles.

4.2. Verification of ARTS system for functional materials

Next, we evaluated whether the potential research topics could be accurately selected with the ARTS system. For this evaluation, we constructed the ARTS system using articles published before 2015 and then checked whether research topics that were studied after 2016 were selected. Our focus here was on functional materials.

We selected steel materials, which support the foundations of modern society, as a representative functional material. Iron and steel are the most widely utilized materials in social infrastructure (e.g. bridges, railroads, automobiles, and other objects used in daily life), so it is beneficial not only in the academic field but also in the industrial field to look for promising and unstudied research topics related to these materials. Therefore, we searched for research topics using the ARTS system with ‘steel’ as the research concept.

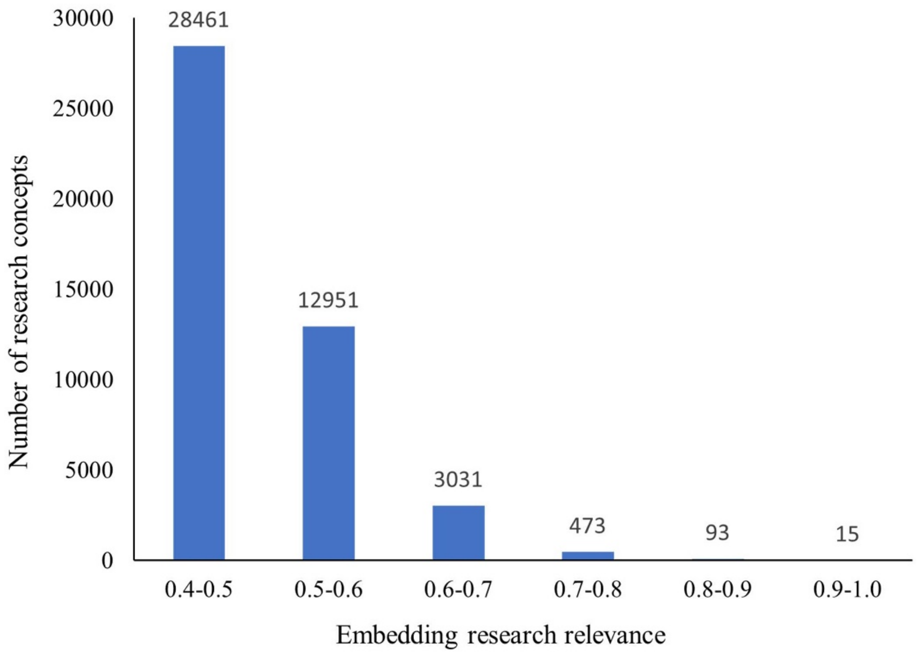


Figure 6. Number of research concept candidates as a function of the embedding research relevance.

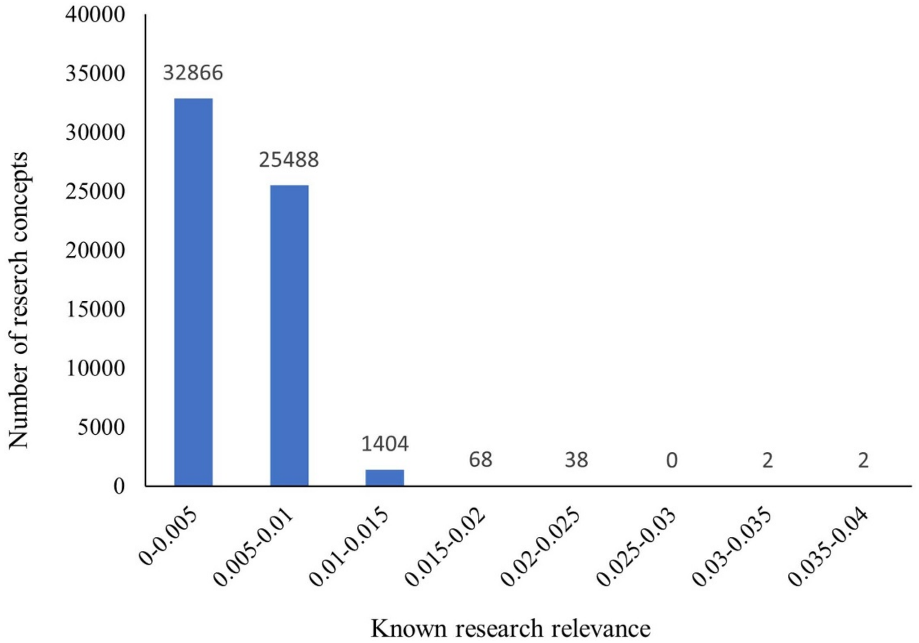


Figure 7. Number of research concept candidates as a function of the known research relevance.

The ARTS system visualized the results as a research topic graph of chromium-molybdenum steel (41xx steel or Chromoly steel). Chromium-molybdenum steel contains chromium and molybdenum and has an excellent strength-to-weight ratio, making it stronger and harder than standard steel. The following concept pair was selected as a potential research topic: **42CrMo steel** (material)–**rolling process** (chemical property).

42CrMo steel, which was the retrieved topic, is a typical medium carbon and low-alloy steel used for automobile crankshafts due to its excellent balance of strength, toughness, and wear resistance. We examined the results from the standpoint of materials science. In automotive applications, traditional spline shaft materials such as 35CrMo and 40Cr steels have been widely replaced by 42CrMo steel. Current active research is mainly focused on cold and hot forming. However, hot forming requires considerable heat energy, making it challenging to apply in actual production processes. On the other hand, the cold forming process of 42CrMo steel still has some weaknesses, such as a high forming load and low forming accuracy. Therefore, studying the rolling process at formation temperatures somewhere between cold and hot is meaningful.

A search for papers on the automatically obtained research topics revealed two papers published in 2017 [40, 41] , indicating that we can indeed extract topics researched after 2015. The following concept pair was also selected as a potential research topic: **42CrMo steel** (material)–**recrystallization kinetics** (model).

To use 42CrMo steel for automobile crankshafts, it needs to be formed by hot forging to achieve high performance. On the other hand, since recrystallization plays an essential role in hot forging, simulating recrystallization kinetics is essential to achieve high performance. We also searched for papers on this research topic and found that one paper was published in 2017 [42] , indicating that a promising research topic could be extracted.

These findings demonstrate that the ARTS system can select new and promising research topics in materials research and development that have not been studied before, and can thereby support the decision-making process of scientists.

4.3. Use case of visualization function

In the following, we present use cases of the two unique functions in the ARTS system. First, using the potential research topic visualization function (figure 4), scientists can obtain various information to assist their research topic determination by narrowing down information with the search queries provided by Neo4j. Scientists can search information using node names, link names, and metadata. For example, if a scientist is interested in the research concept 'O–O bond length', the search query can be used to visualize the research concepts and article information connected to 'O–O bond length'. The 'O–O bond length' node is displayed as an orange circle on the lower left of the screen. Since 'O–O bond length' and 'Cr2CT2' (red circle in the lower left corner) are connected by the *potential link*, this concept pair is a potential research topic. In this way, scientists can easily see potential research topics. If they are interested in a specific research topic, they can immediately see that the article containing 'O–O bond length' is 'Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO3 perovskites' (left green) [43] written by a research group from Shanghai University and Guangxi University in China.

A scientist can obtain even more information from this figure. The articles mentioned above also focus on 'oxide ion conductivity' (center red). The same research concept was found in an article written by a research group from the University of Tokyo, Tohoku University, and Toho University in Japan entitled 'Density functional theory-based *ab initio* molecular dynamics simulation of ionic conduction in N-/F-doped ZrO_2 under epitaxial strain' (right bottom green) [44]. Scientists can accordingly deduce that these two research groups may be competitors. They can also use the search query to look for articles published in a particular year or articles with a high number of citations.

Next, we describe a use case of the time-series network transition visualization function (figure 5). By looking at this visualization, scientists can identify trends such as changes in the material names and model names commonly studied in a particular country. For example, from the countries associated with the research concept of 'machine learning', we can observe that research on 'machine learning' has been active since approximately 2015; the U.S. led in the early research stage, but the amount of research in China has been growing rapidly since 2018.

4.4. Discussion

4.4.1. Calculation of promise index

In word2vec, words are indexed and then word vectors are calculated using the order information of the index as a feature. Therefore, the similarity of the strings is not considered (e.g. even words with similar strings, such as 42CrMo and 40CrMo, are treated as completely different). For this reason, promise is not learned from the similarity of strings. In other words, just because research topics that have similar strings, e.g. **40CrMo steel–rolling process**, have been researched in the past, it does not mean the research topic **42CrMo steel–rolling process** will be output.

Other word embedding techniques such as fastText [45] learn features of strings in addition to the co-occurrence of words. In cases where the similarity of strings is related to the similarity of objects, such as material names, it may be better to use a model that considers the string information.

4.4.2. Setting of thresholds

We set the thresholds of the embedding research relevance and the known research relevance for the ARTS system as described in section 3. We recommend that researchers freely set thresholds to help select a research topic that best suits their preferences.

Specifically, when visualizing research topics, researchers should increase the value of the embedding research relevance threshold if they want to select research topics that are high promise, and conversely, if they want to select many research topics even if the promise is somewhat low, they should decrease the value of that threshold. Similarly, the value of the known research relevance threshold should be decreased if researchers want to select research topics for which there is little existing past research, and that value should be increased if they want to look at many research topics regardless of the existence of past research.

4.4.3. Limitations

The first limitation of the ARTS system relates to how the research topic is defined. The system extracts research concepts from past articles, which means it is impossible to output completely new research concepts that have never before appeared in print. Additionally, a research topic is defined as a pair of two research concepts in this study, but in some cases it would be better represented by a combination of three or more concepts.

The second limitation relates to the research topic index. In section 2, we described the four research topic indexes that are most frequently considered, but we did not examine the fourth one (importance of the research topic) in this study.

The third limitation relates to the autonomy. The ARTS system requires the researcher to set the thresholds for presenting research topics and to make the final decision on the research topic. Therefore, it cannot be claimed that the system is fully autonomous. Krenn *et al* proposed a method for recommending research topics tailored to each researcher by using information on research topics that the researcher has actually engaged with in the past [23]. Their method may be a hint for adapting research topic selection to suit the researcher's preference.

The fourth limitation relates to the method of calculating promise. A researcher needs to interpret the meaning of the selected research topics because word2vec, which is used for the promise calculation, is unable to interpret context and meaning. For example, if there are a large number of sentences such as 'Polymers are generally less conductive', the promise between polymer and conductivity becomes high. However, it is not clear what the relationship between the polymer and conductivity is (e.g. does the polymer have a higher or lower conductivity?). Therefore, a researcher needs to interpret the meaning of the research topics.

Related to the above issue, the handling of polysemous words, i.e. words that can have different meanings depending on the research field or context, can be problematic because word2vec cannot take into account the differences in context. Therefore, polysemous words may become noise interfering with the promise calculation. Recent word embedding techniques such as BERT [46] and GPT [47–49] can calculate word vectors for each word in each context, so it might be possible to handle polysemous words using these techniques.

Additionally, knowledge of physics and chemistry, such as known equations and physical laws, are not explicitly included in the calculation of promise, which means that research topics that are impossible according to physical laws are sometimes selected. Selection of research topics that are more realistic may be achieved by incorporating knowledge of physics and chemistry into the calculation of promise. Efforts to improve prediction results by incorporating human knowledge into machine learning models, which are completely data-driven, have been studied [50–53].

The fifth limitation is related to visualization problems. It is difficult to visualize a network that contain abstract or broad research concepts because they are likely connected to many other research concepts. For example, the research concept 'AI' is connected to many model names, such as 'machine learning' and 'Convolutional Neural Network', which are equivalent to a hyponym of 'AI', as well as to the large number of material names and physical property names that have been studied using these models. We feel that concrete research concepts such as 'Convolutional Neural Network', 'thermoelectric', and so on are more useful in determining research topics compared to abstract words like 'AI'.

It should be possible to have the system output more meaningful research topics that are easier to visualize by hierarchizing the concepts into hypernym and hyponym and then narrowing them down to only the hyponyms for visualization. Databases that define hypernym and hyponym (e.g. WordNet [54]) could be utilized for this purpose. However, since many field-specific technical terms are not registered in existing databases, the hypernyms and hyponyms would need to be extracted, which remains a difficult task despite many years of research [55–58].

5. Conclusion and outlook

In this paper, we proposed the ARTS system, which autonomously selects potential research topics that have a high possibility to reveal new scientific fact and have appeared in few previous studies. The ARTS system is also equipped with functions to search and visualize information about autonomously selected research topics to assist scientists in their final determination of a research topic. We developed and evaluated the ARTS system using articles published in Computational Materials Science. The results showed that when we selected research topics using articles published after 2015, our system could select research topics that were researched after 2016. This suggests the potential of ARTS with the ARTS system.

To make our system more effective, it is necessary to use many more articles for its construction. In this work, we mainly used articles that focus on theoretical studies, but articles that focus on experimental studies should also be included to select genuinely valuable potential research topics. In future work, we will improve our system in this way, thus enabling researchers to use it as a tool for identifying the seeds of great discovery.

Data availability statement

The data used in this study were taken from articles published in the Computational Materials Science journal (https://www.journals.elsevier.com/computational-materials-science) and can be downloaded using the Elsevier API (https://dev.elsevier.com/). The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

This work was supported in part by the JST-Mirai Program under Grant No. JPMJMI21G2.

ORCID iDs

Eri Teruya https://orcid.org/0000-0002-7601-6987
Kanta Ono https://orcid.org/0000-0002-3285-9093

References

[1] Burger B *et al* 2020 A mobile robotic chemist *Nature* **583** 237–41

[2] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 Machine learning in materials informatics: recent applications and prospects *npj Comput. Mater.* **3** 1–13

[3] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 1–36

[4] Tanaka I 2018 *Nanoinformatics* (Singapore: Springer Nature)

[5] Vasudevan R, Pilania G and Balachandran P V 2021 Machine learning for materials design and discovery *J. Appl. Phys.* **129** 070401

[6] Morgan D and Jacobs R 2020 Opportunities and challenges for machine learning in materials science *Annu. Rev. Mater. Res.* **50** 71–103

[7] Batra R, Song L and Ramprasad R 2021 Emerging materials intelligence ecosystems propelled by machine learning *Nat. Rev. Mater.* **6** 655–78

[8] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55

[9] Greenhill S, Rana S, Gupta S, Vellanki P and Venkatesh S 2020 Bayesian optimization for adaptive experimental design: a review *IEEE Access* **8** 13937–48

[10] Frazier P I and Wang J 2016 Bayesian optimization for materials design *Information Science for Materials Discovery and Design* (Berlin: Springer) pp 45–75

[11] Shenghong J, Shiga T, Feng L, Hou Z, Tsuda K and Shiomi J 2017 Designing nanostructures for phonon transport via bayesian optimization *Phys. Rev.* X **7** 021024

[12] Ueno T, Rhone T D, Hou Z, Mizoguchi T and Tsuda K 2016 Combo: an efficient bayesian optimization library for materials science *Mater. Discovery* **4** 18–21

[13] Shields B J, Stevens J, Li J, Parasram M, Damani F, Alvarado J I M, Janey J M, Adams R P and Doyle A G 2021 Bayesian reaction optimization as a tool for chemical synthesis *Nature* **590** 89–96

[14] Shimizu R, Kobayashi S, Watanabe Y, Ando Y and Hitosugi T 2020 Autonomous materials synthesis by machine learning and robotics *APL Mater.* **8** 111110

[15] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm *npj Comput. Mater.* **6** 138

[16] Pilania G, Wang C, Jiang X, Rajasekaran S and Ramprasad R 2013 Accelerating materials property predictions using machine learning *Sci. Rep.* **3** 1–6

[17] Wan X, Feng W, Wang Y, Wang H, Zhang X, Deng C and Yang N 2019 Materials discovery and properties prediction in thermal transport via materials informatics: a mini review *Nano Lett.* **19** 3387–95

[18] Segler M H S, Preuss M and Waller M P 2018 Planning chemical syntheses with deep neural networks and symbolic AI *Nature* **555** 604–10

[19] Coley C W *et al* 2019 A robotic platform for flow synthesis of organic compounds informed by ai planning *Science* **365** 6453

[20] Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Nguyen Q L, Ho S, Sloane J, Wender P and Pande V 2017 Retrosynthetic reaction prediction using neural sequence-to-sequence models *ACS Cent. Sci.* **3** 1103–13

[21] Osakabe Y, Asahara A and Morita H 2020 Hitachi materials informatics analytics platform assisting rapid development *AAAI Symp.: Combining Machine Learning With Knowledge Engineering (1)*

[22] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–98

[23] Krenn M and Zeilinger A 2020 Predicting research trends with semantic and neural networks with an application in quantum physics *Proc. National Academy of Sciences* vol 117 pp 1910–6

[24] Brodiuk S, Palchykov V and Holovatch Y 2020 Embedding technique and network analysis of scientific innovations emergence in an arxiv-based concept network *2020 IEEE Third Int. Conf. on Data Stream Mining and Processing (DSMP)* (IEEE) pp 366–71

[25] Shetty P and Ramprasad R 2021 Automated knowledge extraction from polymer literature using natural language processing *Iscience* **24** 101922

[26] Mikolov T, Chen K, Corrado G and Dean J 2013 Efficient estimation of word representations in vector space (arXiv:1301.3781)

[27] Mikolov T, Sutskever I, Chen K, Corrado G S and Dean J 2013 Distributed representations of words and phrases and their compositionality *Advances in Neural Information Processing Systems* pp 3111–9

[28] Lovasz-Bukvova H 2009 Research as a process: a comparison between different research approaches *Sprouts: Work. Pap. Inf. Syst.* **9** 29

[29] Acs P 2015 *Data Analysis in Practice* (London: Chapman & Hall)

[30] Elliott K C, Cheruvelil K S, Montgomery G M and Soranno P A 2016 Conceptions of good science in our data-rich world *BioScience* **66** 880–9

[31] Mintz M, Bills S, Snow R and Jurafsky D 2009 Distant supervision for relation extraction without labeled data *Proc. Conf. 47th Annual Meeting of the ACL and the 4th Int. Conf. on Natural Language Processing of the AFNLP* pp 1003–11

[32] Nadeau D and Sekine S 2007 A survey of named entity recognition and classification *Lingvist. Investig.* **30** 3–26

[33] Swain M C and Cole J M 2016 Chemdata extractor: a toolkit for automated extraction of chemical information from the scientific literature *J. Chem. Inf. Model.* **56** 1894–904

[34] Ramshaw L A and Marcus M P 1999 Text chunking using transformation-based learning *Natural Language Processing Using Very Large Corpora* (Berlin: Springer) pp 157–76

[35] Onishi T, Kadohira T and Watanabe I 2018 Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity *Sci. Technol. Adv. Mater.* **19** 649–59

[36] Manning C D, Surdeanu M, Bauer J, Finkel J R, Bethard S and David M 2014 The stanford corenlp natural language processing toolkit *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* pp 55–60

[37] Shin J, Wu S, Wang F, De Sa C, Zhang C and Ré C 2015 Incremental knowledge base construction using deepdive *Proc. VLDB Endowment Int. Conf. on Very Large Data Bases* **vol 8** (NIH Public Access) p 1310

[38] Neo4j (available at: https://neo4j.com/)

[39] Susan Sinnott Computational materials science (available at: www.journals.elsevier.com/computational-materials-science)

[40] Cui M-C, Zhao S-D, Zhang D-W, Chen C, Fan S-Q and Li Y-Y 2017 Deformation mechanism and performance improvement of spline shaft with 42CrMo steel by axial-infeed incremental rolling process *Int. J. Adv. Manuf. Technol.* **88** 2621–30

[41] Cui M-C, Zhao S-D, Zhang D-W, Chen C and Li Y-Y 2017 Finite element analysis on axial-pushed incremental warm rolling process of spline shaft with 42crmo steel and relevant improvement *Int. J. Adv. Manuf. Technol.* **90** 2477–90

[42] Chen M-S, Yuan W-Q, Lin Y C, Li H-B and Zou Z-H 2017 Modeling and simulation of dynamic recrystallization behavior for 42CrMo steel by an extended cellular automaton method *Vacuum* **146** 142–51

[43] Xu L, Wencong L, Chunrong P, Qiang S and Jin G 2009 Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO_3 perovskites *Comput. Mater. Sci.* **46** 860–8

[44] Oka M, Kamisaka H, Fukumura T and Hasegawa T 2018 Density functional theory-based ab initio molecular dynamics simulation of ionic conduction in N-/F-doped ZrO_2 under epitaxial strain *Comput. Mater. Sci.* **154** 91–6

[45] Bojanowski P, Grave E, Joulin A and Mikolov T 2017 Enriching word vectors with subword information *Trans. Assoc. Comput. Linguist.* **5** 135–46

[46] Devlin J, Chang M-W, Lee K and Toutanova K 2018 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)

[47] Radford A, Narasimhan K, Salimans T and Sutskever I 2018 *Improving language understanding by generative pre-training* OpenAI

[48] Radford A *et al* 2019 Language models are unsupervised multitask learners *OpenAI blog* **1** 9

[49] Brown T B *et al* 2020 Language models are few-shot learners (arXiv:2005.14165)

[50] Greydanus S J, Dzumba M and Yosinski J 2019 Hamiltonian neural networks *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* **32** 15379–89

[51] Laura von R *et al* 2019 Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems (arXiv:1903.12394)

[52] Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S and Yang L 2021 Physics-informed machine learning *Nat. Rev. Phys.* **3** 422–40

[53] Willard J, Jia X, Shaoming X, Steinbach M and Kumar V 2021 Integrating scientific knowledge with machine learning for engineering and environmental systems (arXiv:2003.04919) pp 1–35

[54] Miller G A 1998 *Wordnet: An Electronic Lexical Database* (Cambridge, MA: MIT Press)

[55] Snow R, Jurafsky D and Ng A Y 2004 Learning syntactic patterns for automatic hypernym discovery *Proc. 17th Int. Conf. on Neural Information Processing Systems* pp 1297–304

[56] Erik T K S 2007 Extracting hypernym pairs from the web *Proc. 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proc. Demo and Poster Sessions* pp 165–8

[57] Wang S, Liang C, Zhaohui W, Williams K, Pursel B, Brautigam B, Saul S, Williams H, Bowen K and Lee Giles C 2015 Concept hierarchy extraction from textbooks *Proc. 2015 Symp. on Document Engineering, DocEng '15* (*New York, USA*: Association for Computing Machinery) pp 147–56

[58] Zhang C, Xie G, Liu N, Xiaojie H, Shen Y and Shen X 2021 Automatic hypernym-hyponym relation extraction with wordnet projection *2021 7th Int. Conf. on Systems and Informatics (ICSAI)* pp 1–6