Extracting and analyzing inorganic material synthesis procedures in the literature

Kohei Makino (bear.kohei@gmail.com)
Toyota Technological Institute: Toyota Kogyo Daigaku https://orcid.org/0000-0002-0427-6050
Fusataka Kuniyoshi
National Institute of Advanced Industrial Science and Technology: Kokuritsu Kenkyu Kaihatsu Hojin Sangyo Gijutsu Sogo Kenkyujo https://orcid.org/0000-0001-5914-009X
Jun Ozawa
National Institute of Advanced Industrial Science and Technology: Kokuritsu Kenkyu Kaihatsu Hojin Sangyo Gijutsu Sogo Kenkyujo https://orcid.org/0000-0003-4212-9008
Makoto Miwa
Toyota Technological Institute: Toyota Kogyo Daigaku https://orcid.org/0000-0002-2330-6972

Research article

Keywords: Artificial intelligence, data mining, information extraction, materials, materials science and technology, materials informatics, natural language processing, neural networks

Posted Date: September 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-636735/v2

Version of Record: A version of this preprint was published at IEEE Access on January 1st, 2022. See the published version at https://doi.org/10.1109/ACCESS.2022.3160201.

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Extracting and Analyzing Inorganic Material Synthesis Procedures in the Literature

KOHEI MAKINO[1,2], FUSATAKA KUNIYOSHI[1,3], JUN OZAWA[1,3], AND MAKOTO MIWA.[1,2]

[1]Panasonic-AIST Advanced AI Cooperative Research Laboratory, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[2]Toyota Technological Institute, Nagoya, Japan
[3]Panasonic Corporation, Osaka, Japan

Corresponding author: Kohei Makino (e-mail: sd21505@toyota-ti.ac.jp).

ABSTRACT Analyzing material synthesis procedures in the literature is required to collect structural information of material names and synthesis procedures for designing materials computationally. Since synthesis procedures are mostly written in natural language in paper or technical documents, they need to be extracted and structured into a format that can be handled by a computer through information extraction. Moreover, to represent a synthesis procedure, it is necessary to express information such as conditions and the order of operations in the procedure, but existing databases that compile structural information of material names and synthesis procedures of materials do not provide such information about procedures. It is, therefore, necessary to create a framework that extracts and organizes the information of synthesis procedures in text so that the information is enough for material development such as the order of operations and the links among materials, operations, and conditions. In this study, we construct a pipeline system that extracts synthesis procedures from a text in the form of a flow graph. The extraction system consists of preprocessing, deep learning-based entity extraction, rule-based relation extraction, and selection for paragraph-containing procedures. We applied the system to a large body of literature and extracted flow graphs (procedures) that include about 4 million entities and 3 million relations. We took several statistics on the extracted graphs and performed several analyses on the extracted graphs. We experimentally confirmed that some extracted operations were specific to the target material and the frequently extracted sub-graphs include reasonable operations.

INDEX TERMS Artificial intelligence, data mining, information extraction, materials, materials science and technology, materials informatics, natural language processing, neural networks

I. INTRODUCTION

In recent years, materials informatics has been attracting attention in the development of new materials by analyzing existing material properties and synthesis procedures with computers. Materials informatics contributes to the reduction of development costs by reducing the number of experimental procedures using real materials through experiments and analyses that are completed on computers. In order to take such an approach, it is necessary to statistically analyze a vast amount of information.

One of the major challenges in the computational experiment design is the collection of structural data of material names and material synthesis procedure [2]–[6]. However, it is very difficult to create a recipe database for material informatics from real experiments conducted in a laboratory because the know-how is often not documented. The development of actual procedures for materials requires a huge amount of time on the order of years and development by experts since the procedures are often developed by experts through repeated experiments based on experience and intuition.

Several studies have attempted to extract material synthesis procedures [7]–[12] from the literature by natural language processing. Mysore et al. [12] created a corpus to extract synthetic procedures in the literature in the field of inorganic materials. Kuniyoshi et al. [7] created a corpus for a specialized field of inorganic materials, focusing on the whole battery, and studied the extraction methods. However, although there have been active proposals of methods for extracting procedures from papers, there have been few papers

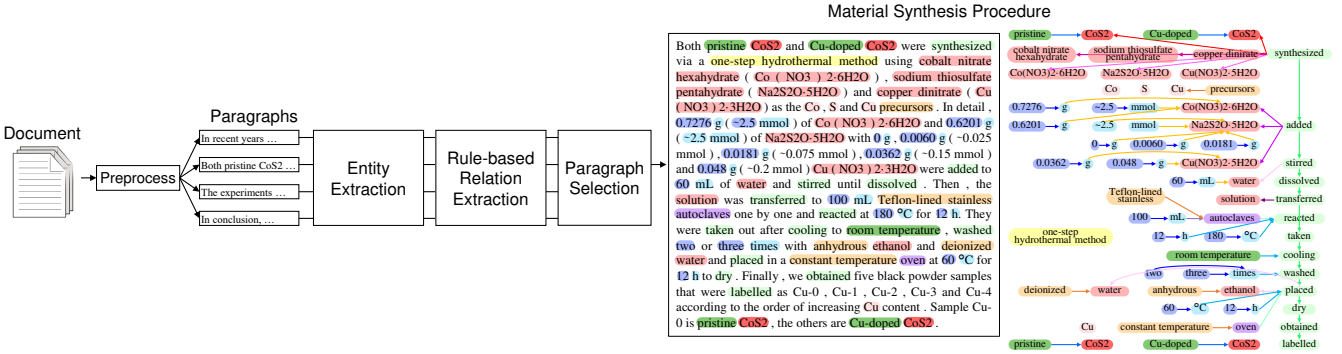

FIGURE 1. Overview of the extraction pipeline system and illustration of the material synthesis procedure adapted from Zhang et al. [1]

analyzing the procedures of materials extracted from actual papers.

In this study, we analyze the procedures of materials extracted from large-scale literature by a system that extracts procedures of materials from actual literature as flow graphs, and confirm the characteristics of the procedures extracted by the system. We build a pipeline extraction system based on the synthetic procedure extraction method of Kuniyoshi et al. [7], and apply this system to a large set of literature for analysis to examine usability of extracted procedures. We analyzed the procedures of the material as procedures rather than as single units of operation by collecting statistics as a flow graph.

II. METHODS

We construct a pipeline system that extracts synthesis procedures in a graph form from the raw data of the literature to extract the material synthesis procedures, as shown in Figure 1.

The overview of the extraction pipeline system is shown in Figure 1. First, the pipeline extracts paragraphs from the raw data collected by preprocessing them into a text format that can be inputted to subsequent models. Then, the system performs neural entity and rule-based relation extraction for all paragraphs. Finally, rules are applied to the extracted results to select the paragraphs in which the synthesis procedures are described.

A. MATERIAL SYNTHESIS PROCEDURE

We define synthesis procedures using a flow graph that is adapted from the previous studies of Kuniyoshi et al. [7] and Mysore et al. [12]. In the flow graph, edges connect each node, which are synthesis operations and materials, to represent relations between entities, including the order of operations and the input of material to the operation.

The flow graph is defined in the materials science procedural text corpus [12] to structure the flow of a procedure. In this corpus, the synthesis procedure is annotated for the inorganic materials literature, and mentions of entities and relations between them are directly annotated in the document. Moreover, the data published by Mysore et al. [12] consisted of 200 training and 15 development and evaluation datasets.

In addition, in the aforementioned corpus, 19 entity labels and 16 relation labels were defined. An entity is defined as an element in a procedure, which can be categorized into several types such as an operation or a material among others. Conversely, a relation is defined as a relationship between entities that describes the order of operations or the input of a material to an operation. The statistics and descriptions for each of these labels are shown in Table 1 and 2.

B. SYNTHESIS PROCEDURE EXTRACTION PIPELINE

We built a pipeline that enables the extraction of synthesis procedures from raw data. It can be divided into four parts. First, the paragraphs from the raw literature data were extracted for preprocessing. Second, entities from the paragraphs were also extracted using an entity extractor. A bidirectional long short-term memory [13] and conditional random field [14] (Bi-LSTM-CRF) model [15] was used to predict entities using the representation of the pre-training model Mat-ELMo [16], which is trained on the literature in the field of inorganic material science. Third, we extracted the synthesis procedure based on the entities using a rule-based relation extractor. The entity and relation extractors are based on the method of Kuniyoshi et al. [7], which extracts the synthesis procedure effectively from the literature on solid-state batteries. Finally, we classified paragraphs depending on whether they contained the synthesis procedure or not using the extracted entities and relations. We chose only the paragraphs that contained the synthesis procedure.

1) Preprocessing

In the preprocessing stage of the data, paragraphs in the literature were extracted. In this study, we used all paragraphs to extract procedures, in contrast to the method proposed by Mysore et al. [12] in which the paragraphs containing a procedure were identified first before annotating.

2) Entity extractor

Following Kuniyoshi et al. [7], entities were extracted by formulating the extraction task as a sequence labeling task with IOB (inside—outside—beginning) tags using a Bi-LSTM-CRF model [15] with Mat-ELMo [16] for the token embeddings. Mat-ELMo is an ELMo (embeddings from language

TABLE 1. Definitions and statistics of entity labels in the materials science procedural text corpus [12]

Entity label	Train	Dev	Test	Description
OPERATION	3,249	212	242	Operations in the procedure
MATERIAL	4,271	277	316	Materials in the procedure
NONRECIPE-MATERIAL	329	33	25	Materials not in the procedure
NUMBER	2,872	224	219	Number
PROPERTY-MISC	481	25	16	Properties of materials
PROPERTY-TYPE	124	10	4	Types for properties for materials
PROPERTY-UNIT	92	7	8	Units for properties for materials
AMOUNT-MISC	149	14	7	Properties for the amount of materials
AMOUNT-UNIT	1,193	96	98	Units for the amount of materials
CONDITION-MISC	468	32	20	Operation conditions
CONDITION-UNIT	1,363	101	87	Units for conditions for operations
CONDITION-TYPE	119	2	1	Types for conditions for operations
SYNTHESIS-APPARATUS	433	20	34	Apparatuses in the procedure
CHARACTERIZATION-APPARATUS	54	2	11	Apparatuses for the evaluation
APPARATUS-UNIT	89	6	16	Units for properties for apparatuses
APPARATUS-PROPERTY-TYPE	26	0	6	Types for properties for apparatuses
MATERIAL-DESCRIPTOR	1,214	67	89	Descriptor of materials
APPARATUS-DESCRIPTOR	165	10	9	Descriptor of apparatus
BRAND	291	30	27	A brand of materials or apparatuses
META	128	12	13	A name of synthesis method
REFERENCE	106	10	11	A citation ID
Overall	17,216	1,190	1,259	

TABLE 2. Relations in the materials science procedural text corpus

Relation class	Train	Dev	Test	Description
NEXT_OPERATION	2,898	184	202	Next operation from before operation
RECIPE_PRECURSOR	876	67	89	Input material of synthesis procedure
RECIPE_TARGET	363	31	22	Target material of synthesis procedure
PARTICIPANT_MATERIAL	1,723	113	124	Intermediate generated material
SOLVENT_MATERIAL	463	28	33	Material for solvent
ATMOSPHERIC_MATERIAL	183	11	14	Material for atmosphere
NUMBER_OF	2,805	219	209	Relation between number and its unit
PROPERTY_OF	586	35	21	Property of material
AMOUNT_OF	1,512	130	121	Condition of amount of material
CONDITION_OF	1,810	132	107	Conditioning for operation
APPARATUS_OF	455	20	36	Conditioning the apparatus for operation
APPARATUS_ATTR_OF	90	6	11	Numerical requirements for the apparatus
DESCRIPTOR_OF	1,495	91	102	Description of the subject
BRAND_OF	423	42	41	Brand of the subject
TYPE_OF	164	7	13	Conditioning on numerical conditions
COREF_OF	267	12	14	Coreference
Overall	16,113	1,128	1,159	

models) [17] model pretrained on the materials literature. We adopted this model although there are several other methods for entity extraction [18] because the survey conducted by Kuniyoshi et al. [7] confirmed that this method is effective for extracting synthesis procedures.

The input text was tokenized and embedded into a dense vector representation for each token using Mat-ELMo, as follows.

$$\boldsymbol{e} = \text{Mat-ELMo}(\boldsymbol{t}), \tag{1}$$

where $\boldsymbol{t} = [t_1, t_2, \cdots t_L]$ is a list of tokens with length L and $\boldsymbol{e} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots \boldsymbol{e}_L]$ is a list of embeddings for each token. The probabilities of classes for tokens are obtained from LSTM.

$$\boldsymbol{p} = \text{LSTM}(\boldsymbol{e}), \tag{2}$$

where $\boldsymbol{p} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots \boldsymbol{p}_L]$ is a list of probabilities of classes for the tokens. The CRF is applied to $\boldsymbol{p}$ to determine the class from the probability.

$$\boldsymbol{c} = \text{CRF}(\boldsymbol{p}), \tag{3}$$

where $\boldsymbol{c} = [c_1, c_2, \cdots, c_L]$ is a list of classes for the tokens in the input text. We maximized the log-likelihood of the prediction of sequences to train this model.

In addition to the entities in the corpus, we introduced a Target-Material class and extracted the entities using the entity extractor. Moreover, we automatically induced the Target-Material entities by considering the Material entities connected by the Recipe_Target relation as the Target-Material entities. This is necessary because it is difficult to develop a simple rule to distinguish the Recipe_Target relation from other relations in Section II-B3, and neural methods are suitable for distinguishing the relation. By introducing the Target-Material entities, we can extract the Recipe_Target relation using a rule as in Section II-B3.

TABLE 3. Dictionary for SOLVENT_MATERIAL

Regular expression
.*water.*
.*(n|alcoh|glyc)ol.*
.*NaOH.*
.*HCl.*
.*acetone.*
.*acid.*
.*H2O.*
.*chloroform.*
.*sodium hydroxide.*
.*DMF.*
.*THF.*
.*N,N-dimethylformamide.*
.*hexane.*
.*toluene.*
.*H2SO4.*
.*EtOH.*

TABLE 4. Dictionary for ATMOSPHERIC_MATERIAL

Regular expression
.*air.*
.*argon.*
.*Ar.*
.*N2.*
.*nitrogen.*
.*H2.*
.*oxygen.*
.*O2.*
.*hydrogen.*
.*CH4.*
.*H2S.*
.*He.*

TABLE 5. Dictionary for PARTICIPANT_MATERIAL

Regular expression	
.*solution.*	.*precipitates.*
.*mixture.*	.*carbon.*
.*product.*	.*silica.*
.*samples.*	.*HCl.*
.*precipitate.*	.*NaBH4.*
.*powder.*	.*slurry.*
.*suspension.*	.*Cu.*
.*precursor.*	.*H2O.*
.*sample.*	.*Samples.*
.*products.*	.*ZnO.*
.*chemicals.*	.*KOH.*
.*powders.*	.*compound.*
.*GO.*	.*filtrate.*
.*solid.*	.*NaOH.*
.*pellets.*	.*films.*
.*solvent.*	.*graphene.*
.*materials.*	.*polymer.*
.*particles.*	.*CTAB.*
.*material.*	.*zeolite.*
.*gel.*	.*SnO2.*
.*reagents.*	.*membrane.*
.*precursors.*	.*hydrochloric aci.*

3) Rule-based relation extractor

Our relation extractor is a rule-based model used to extract the relation between an entity pair. We adapted the rules of Kuniyoshi et al. [7] from the materials science procedural text corpus [12], which depends on the labels of the entities of an entity pair, the distance between an entity pair, and the order of occurrence of the entities. According to the combination of labels of the entities, our rules are divided into three types: Operation–Operation, Operation–Material, and other relations. In the following description of the rules, the starting point of a relation is called the head, the ending point is called the tail, and an edge is denoted as Head–Tail.

a: Operation–Operation

The relation Operation–Operation takes only a Next_Operation label, which indicates the progress of the operation.

Next_Operation: We assumed that Operation is described in the order of the operation progression. Therefore, Operation entities are connected in the order in which they appear.

b: Operation–Material

The Operation–Material relations mean Operation is performed using Material. Moreover, the relation classes are divided into five categories depending on the function of the Material: (1) Recipe_Precursor indicates the input of a material, (2) Recipe_Target indicates the generation of a product, (3) Participant_Material indicates the generation of an intermediate product, (4) Solvent_Material indicates the solvent material of an operation, and (5) Atmospheric_Material indicates the atmosphere of an operation.

We classified Solvent_Material, Atmospheric_Material, and Participant_Material labels based on dictionary matches because the words in Material are distinctive. A dictionary was prepared for each label. The relations connect a Material to the nearest Operation in the sentence if Material matches in the dictionary because these relations take specific Material entities. The dictionaries are listed in Tables 3, 4, and 5.

For a Recipe_Target label, as Material is the target to be extracted as Target-Material by the entity extractor in the previous section, the relation extractor connects the Target-Material to the nearest Operation.

Recipe_Precursor connects all Material except Target-Material that do not match the dictionary of Solvent_Material, Atmospheric_Material, and Participant_Material to the nearest Operation.

c: Remaining relations

The remaining nine relation labels are defined between the other pairs of entity labels: Property_Of, which indicates the condition of a material; Condition_Of, which indicates the condition of an operation; Number_Of, which indicates the relationship between a number and a unit; Amount_Of, which indicates a condition of a quantity; Type_Of, which indicates type of the numerical condition; Brand_Of, which indicates the brand of a material or equipment; Apparatus_Of, which indicates the equipment used in an operation; Apparatus_Attr_Of, which indicates the numerical condition of the equipment; and Descriptor_Of, which indicates other conditions. For these labels, the rules are defined based only on the labels of the head and tail entities and the distance between them. We explain the detailed rules in the remainder

of this section.

Property_Of: This relation takes Property-Unit or Property-Misc as the head and Material, including Target-Material or Nonrecipe-Material as the tail. When Property-Unit is a head, it is connected to the nearest Material in the sentence. When Property-Misc is a head, it is connected to the nearest Material or Nonrecipe-Material in the sentence.

Condition_Of: This relation takes Condition-Unit or Condition-Misc as the head and Operation as the tail. Condition-Unit and Condition-Misc are connected to the nearest Operation in the sentence using this relation.

Number_Of: Number is connected to the nearest Property-Unit, Condition-Unit, or Apparatus-Unit that appear after the Number in the sentence. We assume that, in most cases, the relation between
num and its unit matches a pattern of "(Number) (Unit)" such as "100 mL".

Amount_Of: This relation connects Amount-Unit and Amount-Unit to the nearest Material or Nonrecipe-Material in the sentence.

Descriptor_Of: When Material-Descriptor is a head, it is connected to the nearest Material or Nonrecipe-Material in the sentence. When Apparatus-Descriptor is a head, Synthesis-Apparatus or Characterization-Apparatus can be a tail, but only the nearest Synthesis-Apparatus in the sentence is connected because Characterization-Apparatus is an apparatus for measuring characteristics and detailed descriptions are rarely provided.

Apparatus_Of: This relation connects Synthesis-Apparatus and Characterization-Apparatus with the nearest Operation with the priority given to the Operation that appears before the Apparatus in the sentence.

Type_Of: Property-Type and Apparatus-Property-Type are connected to the nearest Property-Unit and Apparatus-Unit in the sentence with this relation, respectively. When Condition-Type is a head, it is connected to the nearest Condition-Unit that appears before the Condition-Type in the sentence.

Brand_Of: The relation connects Brand to the nearest entities that may have brands (i.e., Material, Nonrecipe-Material, Synthesis-Apparatus, and Characterization-Apparatus) in the sentence.

Apparatus_Attr_Of: Apparatus-Unit is connected to the nearest Synthesis-Apparatus or Characterization-Apparatus.

Coref_Of: The relation is not detected by the rules because it is difficult to describe rules.

4) Selecting paragraphs

Although extractors were applied to all paragraphs, only few paragraphs may contain procedures. We selected such paragraphs based on the extracted entities and relations.

Several deep learning methods have been proposed to select such paragraphs [19]. However, we employed a simple approach to select these paragraphs using the extracted procedures. As the target of this research is to extract procedural sequences, the extracted procedures require the inclusion of sequences wherein target materials are generated from other materials. Therefore, we selected the paragraphs that contain Recipe_Precursor, Recipe_Target, and Operation entities that intervene between them.

III. RESULTS

In our experiments, we first evaluated our pipeline to determine if it could be used in practice. In the evaluation of the pipeline, we evaluated separately each module, the entity extractor, the relation extractor, and the selection of the paragraphs, and finally evaluated the entire pipeline. Then, we applied the pipeline to extract synthesis procedures from a large set of documents that we collected and analyzed the extracted procedures.

A. EVALUATION SETTINGS FOR ENTITY AND RELATION EXTRACTORS

We utilized the publicly available materials science procedural text corpus [12] to train and develop the entity and relation extractors using the training data of this corpus and by observing these training data, respectively.

Flair [20], a library of machine learning for natural language processing, was used to implement the entity extractor. Meanwhile, the pretraining model Mat-ELMo [16] was fixed during training because of the small size of the corpus. Moreover, we employed a stochastic gradient descent method for training. The learning rate was halved from the initial value of 0.1 when there was no improvement in performance over three epochs, and the method was stopped when the learning rate became smaller than 0.0001. A model that showed the best performance on the development data was used.

The individual evaluation of the entity and relation extractors was performed using the test data of the corpus in which precision, recall, micro-F, which is an overall F-score, and macro-F, which is the averaged F-score of individual classes, were employed as the evaluation metrics.

We compared our entity extractor with BERT (Bidirectional Encoder Representations from Transformers)-base [21] and ELECTRA (efficiently learning an encoder that classifies token replacements accurately)-base [22], which have shown excellent performance in various natural language processing tasks, to verify the effectiveness of the setting for the corpus. BERT is a transformer-based [23] model that is pretrained by masked language modeling in a large number of studies. Conversely, although ELECTRA is similar to BERT, it has shown higher performance using an improved training method. To evaluate these transformer-based models, the embeddings of our entity extractor were replaced with these models, and the training was performed in the same way.

We evaluated the performance of the relation extractor separately from the entity extractor by assigning gold entities to verify the validity of the developed rules. This is the same setting as the rules that were developed. We also evaluated the performance of both extractors in the pipeline to determine the performance of the extraction system.

TABLE 6. Comparison of the entity extractors

	Precision	Recall	Micro-F	Macro-F
Ours	0.759	0.875	0.813	0.565
BERT-base	0.789	0.796	0.792	0.495
ELECTRA-base	0.759	0.866	0.809	0.516

TABLE 7. Detailed evaluation results for each entity class

Class	Precision	Recall	F-score
OPERATION	0.794	0.938	0.860
MATERIAL	0.768	0.881	0.820
TARGET-MATERIAL	0.458	0.524	0.489
NONRECIPE-MATERIAL	0.450	0.429	0.439
NUMBER	0.949	0.936	0.942
PROPERTY-MISC	0.300	0.400	0.343
PROPERTY-TYPE	0.250	0.250	0.250
PROPERTY-UNIT	0.500	0.500	0.500
AMOUNT-MISC	0.273	0.429	0.333
AMOUNT-UNIT	0.969	0.949	0.959
CONDITION-MISC	0.588	1.000	0.741
CONDITION-TYPE	0.067	1.000	0.125
CONDITION-UNIT	0.914	0.977	0.944
SYNTHESIS-APPARATUS	0.677	0.677	0.677
CHARACTERIZATION-APPARATUS	0.444	0.364	0.400
APPARATUS-UNIT	0.667	0.308	0.421
APPARATUS-PROPERTY-TYPE	0.000	0.000	0.000
MATERIAL-DESCRIPTOR	0.663	0.708	0.685
APPARATUS-DESCRIPTOR	0.533	0.889	0.667
BRAND	0.657	0.852	0.742
META	0.429	0.462	0.444
REFERENCE	0.571	0.727	0.640

B. EVALUATION OF THE ENTITY EXTRACTOR

The evaluation results of the entity extractors, including transformer-based models, are shown in Table 6. Our entity extractor succeeds in extracting entities with the highest performance, showing that it is important to use the model learned from the literature in the same domain as the target corpus (i.e., the materials science field). From this result, we concluded that Mat-ELMo is suitable for entity extraction in the field of materials science. As the transformer-based models use subword tokenization obtained from general text such as those found in Wikipedia, they might have difficulties in obtaining representations specific to the materials field.

Focusing on the detailed extraction results of our entity extractor for each class in Table 7, it was found that the prediction performance is not consistent among the classes. In some classes, none of the correct entities were extracted; there were also classes with low F-scores (i.e., >0.2). Moreover, a few incorrect predictions can be detrimental because a small number of positive examples are included in the test data.

C. EVALUATION OF THE RELATION EXTRACTOR

The results for the rule-based relation extractor in the pipeline setting and with gold entities are shown in Table 8. Our extractor with gold entities obtained a micro-F of more than 0.8 without using any sophisticated deep learning methods. The descriptions of the procedures are written in certain patterns (e.g., step-by-step explanation), and the simple rules were able to capture the patterns.

The extraction performance of the pipeline obtained a micro-F of 0.609, indicating that our rules can extract relations from the extracted entities as a part of the pipeline with a reasonable performance of approximately 0.81.

Moreover, the performance of each relation class for both the given entity and pipeline settings is shown in Table 8 to give a more detailed evaluation. When the correct entities are given, classes that are critically related to the procedures such as Next_Operation and Recipe_Target can be extracted with high accuracy. Conversely, the extraction performance of classes such as Participant_Material and Solvent_Material is low. Coref_Of was not extracted at all because we do not have the rule for this relation.

For the pipeline result, the class of Next_Operation, which is a key relation for the sequence of operations on the procedure, exceeds 0.6. This indicates that the main flow of the procedure can be extracted to some extent even if e error has propagated from the previous entity extractor. By contrast, some classes such as Property_Of, Brand_Of, and Apparatus_Attr_Of show extremely low performance.

D. LARGE-SCALE EXTRACTION

We collected articles from the Journal of Materials Chemistry A (JMCA), which focuses on areas such as batteries, fuel cells, sustainable materials, photovoltaics, supercapacitors, and water splitting, published by The Royal Society of Chemistry as the source of literature for large-scale extraction. We purchased all articles from 2015 to 2019 in XML format. The total number of articles was 14,310.

We applied the pipeline extractor to all articles. First, as a preprocessing step, we extracted all paragraphs from the articles in XML format. Then, entities and relations from the paragraphs were also extracted and the paragraphs that describe the procedures were selected. From 14,310 studies, we obtained 347,480 paragraphs in total; this was reduced to 89,578 after selection, in which one-third of the total paragraphs were determined to contain procedures. Surprisingly, each literature mentioned six procedures on average, indicating that there is plenty of information about material synthesis procedures in the text even though there are miscounts due to prediction errors. The statistics of the extracted entities and relations are shown in Table 9 and10, respectively.

IV. DISCUSSION

We analyzed the procedures extracted from a large body of literature to investigate their characteristics. In this investigation, we aimed to verify the reasonableness of the extracted procedures and determine whether they can be used to obtain useful information for material development or other tasks.

A. FREQUENT SUBGRAPHS

We checked the subgraphs that frequently appear using gSpan [24], which is an algorithm for mining frequent subgraphs in the extracted paragraphs, and whether they were reasonable.

TABLE 8. Performance of the relation extractor when the gold entities were given and were extracted by our extractor

	Entity given			Pipeline		
Class	P	R	F	P	R	F
NEXT_OPERATION	0.875	0.990	0.929	0.544	0.743	0.628
RECIPE_PRECURSOR	0.474	0.733	0.575	0.360	0.628	0.458
RECIPE_TARGET	0.952	0.952	0.952	0.500	0.571	0.533
PARTICIPANT_MATERIAL	0.794	0.435	0.562	0.551	0.374	0.446
SOLVENT_MATERIAL	0.438	0.656	0.525	0.340	0.531	0.415
ATMOSPHERIC_MATERIAL	0.375	1.000	0.545	0.270	0.833	0.408
NUMBER_OF	0.962	0.946	0.954	0.888	0.854	0.871
PROPERTY_OF	1.000	0.950	0.974	0.250	0.300	0.273
CONDITION_OF	0.990	0.980	0.985	0.750	0.891	0.814
AMOUNT_OF	0.937	0.781	0.852	0.722	0.614	0.664
APPARATUS_OF	0.846	0.971	0.904	0.526	0.588	0.556
APPARATUS_ATTR_OF	0.818	0.900	0.857	0.400	0.200	0.267
DESCRIPTOR_OF	0.975	0.929	0.952	0.560	0.600	0.580
BRAND_OF	0.917	0.564	0.698	0.364	0.308	0.333
TYPE_OF	1.000	0.875	0.933	0.077	0.125	0.095
COREF_OF	0.000	0.000	0.000	0.000	0.000	0.000
Micro	0.829	0.832	0.830	0.576	0.645	0.609
Macro	0.791	0.772	0.762	0.510	0.444	0.459

TABLE 9. Statistics of entities in the extracted procedures

OPERATION	688,311
MATERIAL	912,266
TARGET-MATERIAL	327,093
NONRECIPE-MATERIAL	231,290
NUMBER	441,297
PROPERTY-MISC	275,665
PROPERTY-TYPE	43,674
PROPERTY-UNIT	95,409
AMOUNT-MISC	5,232
AMOUNT-UNIT	108,585
CONDITION-MISC	36,707
CONDITION-TYPE	20,342
CONDITION-UNIT	171,513
SYNTHESIS-APPARATUS	47,766
CHARACTERIZATION-APPARATUS	87,368
APPARATUS-UNIT	5,326
APPARATUS-PROPERTY-TYPE	861
MATERIAL-DESCRIPTOR	293,905
APPARATUS-DESCRIPTOR	16,154
BRAND	19,920
META	31,007
REFERENCE	21,215
All entities	3,880,906

TABLE 10. Statistics of relations in the extracted procedure

NEXT_OPERATION	598,733
RECIPE_PRECURSOR	457,558
RECIPE_TARGET	231,931
PARTICIPANT_MATERIAL	143,200
SOLVENT_MATERIAL	59,213
ATMOSPHERIC_MATERIAL	55,253
NUMBER_OF	401,816
PROPERTY_OF	329,280
AMOUNT_OF	111,233
CONDITION_OF	166,740
DESCRIPTOR_OF	308,773
APPARATUS_OF	101,001
APPARATUS_ATTR_OF	4,645
BRAND_OF	21,757
TYPE_OF	33,009
All relations	3,024,142

The top 10 frequent subgraphs that contain Operation entities and appear more than 50 times are shown in Figure 2 for each number of nodes in the subgraphs. The figure shows an example of a general procedure of a frequent subgraph; for example, the first and sixth frequent 4-nodes subgraphs indicate that the compounds are heated and dried at 60 °C to remove ethanol, and the fourth 5-nodes graph shows that 60 °C is appropriate for drying ethanol.

B. TARGET-MATERIAL

We analyzed the Target-Material to test the feasibility of the extracted procedures. In our results, every procedure has Target-Material because we filtered out the procedures that do not include Target-Material. We found that it was difficult to directly analyze them because Target-Material is unique in most literature. Therefore, we analyzed the elements included in the Target-Material.

As a basic step, we counted the included elements and extracted the element names from left to right, ignoring valences and other numerical values. We filtered out the Target-Material that was determined to be composed of one element or none to avoid the effect of erroneous extraction.

The frequency of the elements is summarized in Figure 3. Moreover, the elements were normalized using the number of Target-Material entities and was scaled logarithmically. Elements with a single letter may be counted more than their actual occurrences owing to errors in parsing elements contained in Target-Material. For example, the tin (IV) oxide (SnO2) nanoparticles mentioned as "SnO2NP" are recognized as a combination of SnO2, nitrogen (N), and phosphorus (P). However, both N and P represent nanoparticles. The results show that elements O, C, N, P, and S, which are commonly contained in various compounds, and Ni and Ti, which are conventionally used in electrodes, appear frequently.

We calculated the importance of each operation for each element in Target-Material with the term frequency-inverse document frequency (TF-IDF)to analyze the correlations

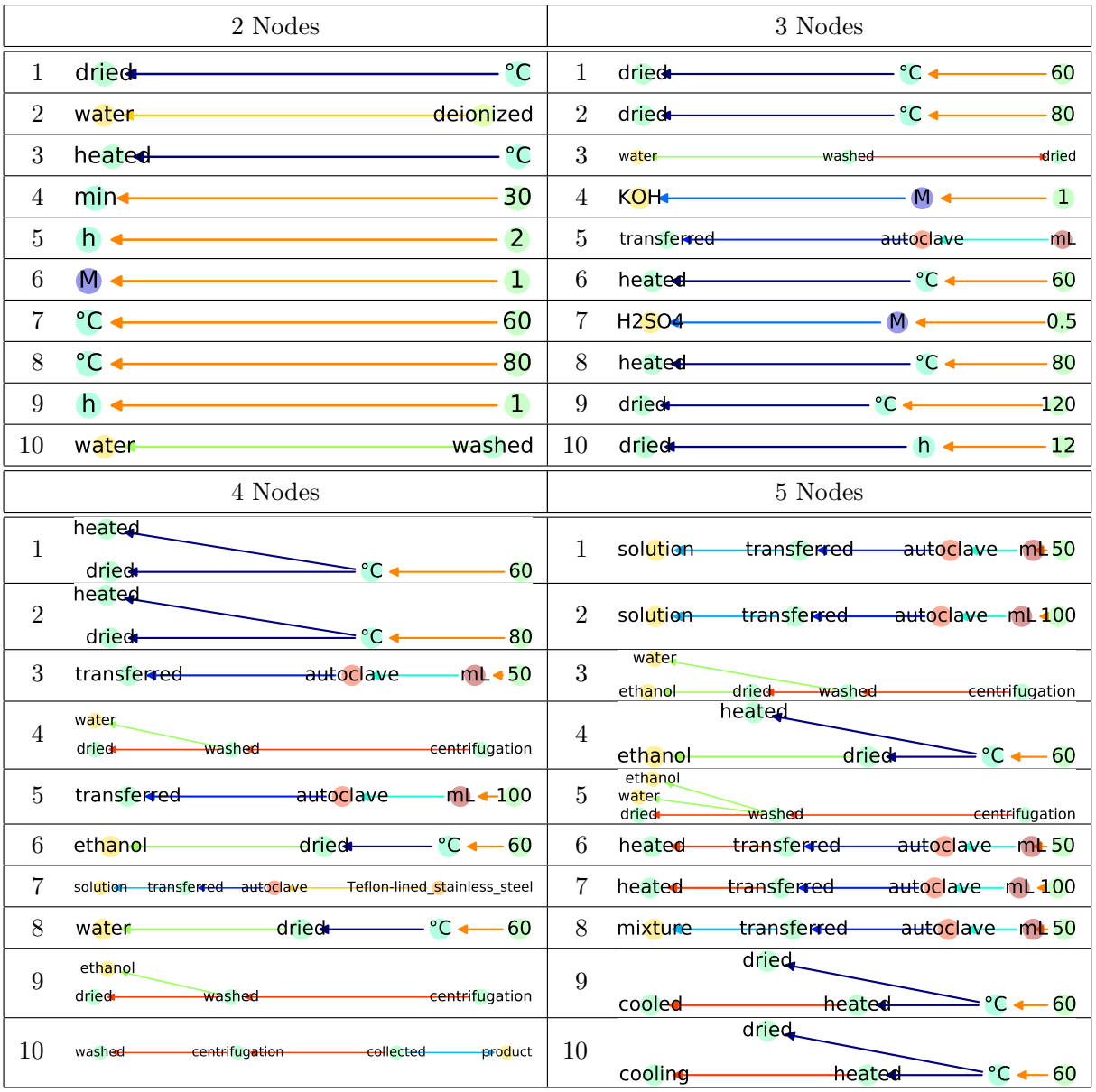

FIGURE 2. Frequent sub-graphs in the extracted procedures

among procedures. Here, the operations in the procedure are considered as terms in a document, and each element in the Target-Material is considered as a document. A document $\mathcal{D}_e$ corresponding to an element e is defined in which a set of all procedures is $\mathcal{P}$, a set of lemmatized Operations in a procedure p is $\mathcal{O}_p$, and a set of elements in Target-Material of a procedure p is $\mathcal{E}_p$.

$$\mathcal{D}_e = \bigcup_{p \in \mathcal{P}} \mathbb{U}[e \in \mathcal{E}_p] \cap \mathcal{O}_p \tag{4}$$

$\mathbb{U}[\cdot]$ is a function that returns a universal set U if the condition in the bracket is satisfied, and an empty set ϕ otherwise. The TF-IDF score $\mathrm{TFIDF}(d,t)$ is defined as a function of a term t in a document d, where $\mathrm{count}(d,t)$ is a function that returns a term t count in a document d, and $\mathbb{1}[\cdot]$ is a function that returns 1 if the condition in the bracket is satisfied and 0 otherwise.

$$\mathrm{TF}(d,t) = \frac{\mathrm{count}(d,t)}{\sum_{w \in d} \mathrm{count}(d,w)} \tag{5}$$

$$\mathrm{IDF}(t) = \log \frac{\#\mathcal{E}}{\sum_{e \in \mathcal{E}} \mathbb{1}[t \in \mathcal{D}_e]} \tag{6}$$

$$\mathrm{TFIDF}(d,t) = \mathrm{TF}(d,t) \cdot \mathrm{IDF}(t) \tag{7}$$

The top 20 Operation ranked by the TF-IDF score are shown in Table 11 for the frequent typical elements (O and C) and metallic elements (Co, Ni, and Ti) shown in Figure 3. We lemmatized Operation using scispaCy [25] before calculating the TF-IDF score. Moreover, we observed

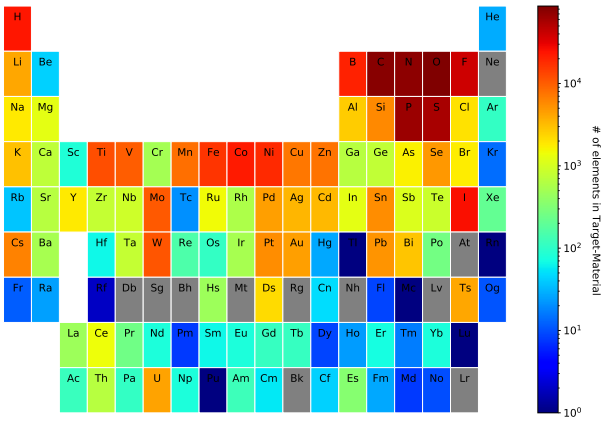

FIGURE 3. The number of individual elements included in TARGET-MATERIAL. The elements with gray did not appear in the extracted procedures.

TABLE 11. The top 20 OPERATIONS ranked by the TF-IDF score for each element in TARGET-MATERIAL.

	O	C	Co	Ni	Ti
1	obtain	prepare	prepare	obtain	obtain
2	form	form	synthesize	synthesize	synthesize
3	synthesize	synthesize	form	form	form
4	use	observe	use	use	use
5	observe	use	formation	observe	fabricate
6	fabricate	increase	observe	fabricate	observe
7	formation	formation	fabricate	formation	add
8	increase	fabricate	increase	increase	dry
9	add	add	dry	dry	increase
10	dry	dry	add	add	deposit
11	perform	perform	maintain	maintain	formation
12	wash	show	perform	grow†	anneal*
13	dissolve	decrease	wash	perform	wash
14	show	measure	grow†	calculate	achieve
15	achieve	calculate	dissolve	dissolve	heat
16	measure	achieve	calculate	wash	measure
17	decrease	maintain	heat	achieve	stir
18	heat	wash	achieve	heat	perform
19	deposit	remove	show	anneal*	dissolve
20	maintain	dissolve	anneal*	show	mix

the general Operation in the top 10 Operation. However, the 11th to 20th Operation are characteristic of each element. We used the Operation that occurs with the O element, which is one of the most common elements in the analysis. In addition, "grow†", which belongs to the columns of Co and Ni, is a characteristic Operation as they are used when sheets or crystals are created. "anneal*", which is seen in metallic elements, is a unique Operation for metals. These aspects are correct for the knowledge gained from the extracted procedures.

C. CASE STUDY OF EXTRACTED PROCEDURES

An example of an extraction procedure is shown in Figure 1. In this study, we sampled 10 documents and checked 64 procedures that were included in the sampled documents to test the feasibility of extracted procedures. The obtained procedures that were manually chosen from the sampled procedures provided information and conditions about each operation, as well as how the Target-Material was produced from the input material. However, there were several errors found, such as "0̃.075 mmol" in the middle of the paragraph that was not extracted as an entity and "water" and "ethanol" that were the Solvent_Material of "washed" were incorrectly connected to "placed" by the relation extractor. However, the general framework of the procedure was extracted, and the analysis of a large number of extracted procedures is expected to contribute to the development of materials.

V. RELATED WORK

Several studies have attempted to extract material names with their properties [26], [27] and material synthesis procedures [7]–[11] by natural language processing. Moreover, studies on the analysis of the extracted material information were also conducted. Saal et al. [28] developed a machine learning model to predict new compounds from the database. Raccuglia et al. [29] predicted the reaction outcomes for the crystallization of templated vanadium selenites from experimental notebooks.

Material synthesis procedures are defined in various ways [7], [9], [12], but the main elements to be extracted are common; target materials, ingredients of target materials, operations such as mixing and annealing, and conditions such as temperature and time. These elements are the basis of materials science. The analyses of the synthesis procedures in the literature have provided information about these materials for real world applications. For example, Mahbub et al. analyzed the experimental conditions to gain insight on solid-state battery materials from database constructed from information in the literature [30], Kim et al. predicted precursor materials of perovskite using text embedding [16], and Segler et al. designed synthesis procedures for organic molecules by predicting precursor from target material [31]. These studies demonstrated the utility of material synthesis procedures in the literature.

VI. CONCLUSIONS

In this study, we constructed a pipeline system for extracting synthesis procedures from the literature. We applied an entity extractor, which consisted of Mat-ELMo and Bi-LSTM-CRF models, and a rule-based relation extractor to extract and structure the synthesis procedures as a graph. Our pipeline system performed a large-scale extraction of procedures to analyze the synthesis procedures. We confirmed that the extracted procedures include reasonable operations such as "drying ethanol at 60 degrees," and the statistics show the characteristic operations of the elements (e.g., the "anneal" operation for metallic elements).

For further study, we will attempt to improve the performance of the extraction of procedures and provide more insights into the synthesis procedures, represented as procedures for discovering new materials. Since our extractor is able to extract only a limited number of classes with high performance, improving the extraction performance of the system is our next challenge to obtain more accurate procedures. In addition, we will continue to analyze the procedures

extracted by the system and develop materials based on the obtained knowledge.

REFERENCES

[1] J. Zhang, B. Xiao, X. Liu, P. Liu, P. Xi, W. Xiao, J. Ding, D. Gao, and D. Xue, "Copper dopants improved the hydrogen evolution activity of earth-abundant cobalt pyrite catalysts by activating the electrocatalytically inert sulfur sites," *J. Mater. Chem. A*, vol. 5, pp. 17 601–17 608, 2017. [Online]. Available: http://dx.doi.org/10.1039/C7TA05433E

[2] C. Draxl and M. Scheffler, "Nomad: The fair concept for big data-driven materials science," *Mrs Bulletin*, vol. 43, no. 9, pp. 676–682, 2018.

[3] G. R. Schleder, A. C. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From dft to machine learning: recent approaches to materials science–a review," *Journal of Physics: Materials*, vol. 2, no. 3, p. 032001, 2019.

[4] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018.

[5] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature materials*, vol. 12, no. 3, pp. 191–201, 2013.

[6] J. M. Rickman, T. Lookman, and S. V. Kalinin, "Materials informatics: From the atomic-level to the continuum," *Acta Materialia*, vol. 168, pp. 473–510, 2019.

[7] F. Kuniyoshi, K. Makino, J. Ozawa, and M. Miwa, "Annotating and extracting synthesis process of all-solid-state batteries from scientific literature," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1941–1950. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.239

[8] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL materials*, vol. 1, no. 1, p. 011002, 2013.

[9] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, and G. Ceder, "Text-mined dataset of inorganic materials synthesis recipes," *Scientific Data*, vol. 6, 12 2019.

[10] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," *npj Computational Materials*, vol. 5, no. 1, pp. 1–17, 2019.

[11] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," *Chemistry of Materials*, vol. 29, no. 21, pp. 9436–9444, 2017.

[12] S. Mysore, Z. Jensen, E. Kim, K. Huang, H.-S. Chang, E. Strubell, J. Flanigan, A. McCallum, and E. Olivetti, "The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures," in *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 56–64. [Online]. Available: https://www.aclweb.org/anthology/W19-4007

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289.

[15] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *CoRR*, vol. abs/1508.01991, 2015. [Online]. Available: http://arxiv.org/abs/1508.01991

[16] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, and E. Olivetti, "Inorganic materials synthesis planning with literature-trained neural networks," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1194–1201, 2020, pMID: 31909619. [Online]. Available: https://doi.org/10.1021/acs.jcim.9b00995

[17] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[18] S. Eltyeb and N. Salim, "Chemical named entities recognition: a review on approaches and applications," *Journal of cheminformatics*, vol. 6, no. 1, pp. 1–12, 2014.

[19] H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan, and G. Ceder, "Semi-supervised machine-learning classification of materials synthesis procedures," *npj Computational Materials*, vol. 5, no. 1, pp. 1–7, 2019.

[20] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *COLING 2018, 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[22] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[24] X. Yan and J. Han, "gspan: graph-based substructure pattern mining," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, pp. 721–724.

[25] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. [Online]. Available: https://www.aclweb.org/anthology/W19-5034

[26] M. C. Swain and J. M. Cole, "Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature," *Journal of chemical information and modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[27] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti, "Machine-learned and codified synthesis parameters of oxide materials," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[28] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd)," *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013.

[29] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, pp. 73–76, 2016.

[30] R. Mahbub, K. Huang, Z. Jensen, Z. D. Hood, J. L. Rupp, and E. A. Olivetti, "Text mining for processing conditions of solid-state battery electrolytes," *Electrochemistry Communications*, vol. 121, p. 106860, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388248120302113

[31] M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic ai," *Nature*, vol. 555, no. 7698, pp. 604–610, 2018.

KOHEI MAKINO received the B.E. and M.E. degrees from Toyota Technological Institute, Aichi, Nagoya, Japan, where he is currently pursuing the doctor's degree.

From 2019 to 2021, he was a Research Assistant at the National Institute of Advanced Industrial Science and Technology. His research interests include deep learning, natural language processing, and information extraction.

FUSATAKA KUNIYOSHI received the M.S. degree in computer science from Nara Institute of Science and Technology in 2017. He is currently a researcher at the National Institute of Advanced Industrial Science and Technology and Panasonic Corporation. His research interests include natural language processing and computer vision.

JUN OZAWA received the Ph.D. degree in system science from Tokyo Institute of Technology, Yokohama, Japan in 1998. From 1990 he was a researcher at Panasonic. He is currently a director of Panasonic-AIST (National Institute of Advanced Industrial Science and Technology) Advanced AI Research Laboratory. His research interests include machine learning and its industrial applications.

MAKOTO MIWA received the Ph.D. degree from the University of Tokyo in 2008. He is currently an associate professor at the Toyota Technological Institute and a visiting researcher at the National Institute of Advanced Industrial Science and Technology. His research interests include natural language processing, deep learning, and information extraction.