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: Deep learning, Information extraction, Relation extraction, Materials informatics, Synthesis procedure

Posted Date: June 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-636735/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Extracting and analyzing inorganic material synthesis procedures in the literature

Kohei Makino^{1,2*}, Kuniyoshi Fusataka^{1,3}, Jun Ozawa^{1,3} and Makoto Miwa^{1,2}

*Correspondence:

sd21505@toyota-ti.ac.jp

¹Toyota Technological Institute,

468-8511 Nagoya, Aichi, Japan

Full list of author information is available at the end of the article

Abstract

Analyzing synthesis procedures from a considerable amount of literature is required to collect structural information of material names and synthesis procedures for designing materials computationally. There are databases comprising structural data of material names and material synthesis procedures. However, the types of material in these databases and the material property values included are limited and insufficient. Moreover, they are primarily described in the literature in the natural language of the researcher who proposed the procedure, and thus they cannot be understood universally. It is, therefore, necessary to create a framework that represents textual synthesis procedures in a flow graph that contains crucial information for material development such as the order of operations and the linkage between operations and conditions. This will facilitate obtaining material insights from the literature. However, there are no large-scale studies on the extraction of synthesis procedures in the form of a graph and analysis thereof. In this study, we propose a pipeline system that extracts synthesis procedures from a text in the form of a graph with a clear order of operations and objects of conditioning from the literature. The system consists of preprocessing, entity extraction, which is based on Mat-ELMo and Bi-LSTM-CRF models, rule-based relation extraction, and selection for paragraph-containing procedures. We applied the system to a large body of literature and extracted various synthesis procedures. We performed basic analyses of the extracted procedures to examine their usability. We experimentally confirmed that some extracted procedures were specific to the target material, and some of the obvious procedures were correctly extracted.

Keywords: Deep learning; Information extraction; Relation extraction; Materials informatics; Synthesis procedure

¹Introduction

²In recent years, inorganic materials science has started focusing on improving the²
³computational experiment design of materials. This allows for a rapid and efficient³
⁴development of design solutions for targeted materials that is necessary to satisfy⁴
⁵the increasing global demand for environmentally sustainable materials. However,⁵
⁶one of the major challenges in the implementation of the computational experiment⁶
⁷design is the collection of structural data of material names and material synthesis⁷
⁸procedures [1, 2, 3, 4, 5], for the active development of materials databases [6, 7].⁸
⁹The types of material in these databases and the material property values included⁹
¹⁰are limited and insufficient. Moreover, materials and material synthesis procedures¹⁰
¹¹are primarily described in the literature in the natural language of the researcher¹¹
¹²who proposed the procedure and thus they cannot be understood universally.¹²

¹³Several studies have attempted to extract material information from texts,¹³
¹⁴material names with their properties [8, 9], and material synthesis procedures¹⁴
¹⁵[10, 11, 12, 13, 14] by natural language processing to address this problem. Moreover,¹⁵
¹⁶studies on the analysis of the extracted material information were also conducted.¹⁶
¹⁷Saal et al. [15] developed a machine learning model to predict new compounds¹⁷
¹⁸from the database. Raccuglia et al. [16] predicted the reaction outcomes for the¹⁸
¹⁹crystallization of templated vanadium selenites from experimental notebooks.¹⁹

²⁰Material synthesis procedures are defined in various methods, but the main el-²⁰
²¹ements to be extracted are target materials, ingredients of target materials, op-²¹
²²erations such as mixing and annealing, and conditions such as temperature and²²
²³time. These elements are the basis of materials science. Real-world applications by²³
²⁴analyzing the synthesis procedures from academic documents have provided infor-²⁴
²⁵mation about these materials, for example, the analysis of experimental conditions²⁵
²⁶to gain insight on solid-state battery materials [17], prediction of precursor mate-²⁶
²⁷rials of perovskite [18], and design synthesis procedures for organic molecules [19].²⁷
²⁸These studies demonstrated the functionality of material synthesis procedures in²⁸
²⁹the literature.²⁹

³⁰In particular, it is necessary to create a framework that represents textual synthe-³⁰
³¹sis procedures in a flow graph [20, 10] that contains crucial information for material³¹
³²development such as the order of operations and the linkage between operations³²
³³and conditions; this will facilitate obtaining material insights from the literature.³³

However, there are no large-scale studies on the extraction and analysis of synthesis procedures. Moreover, it is necessary to extract the graph of synthesis procedures from a large amount of literature in an integrated manner and statistically analyze these information to determine if the graph can be utilized for material development. In this study, we aim to build a system that extracts synthesis procedures from text in graph form to obtain material insights from the synthesis procedures described in a large body of literature. First, we create a pipeline system that extracts synthesis procedures based on the method presented by Kuniyoshi et al. [10]. We also provide the results of basic analyses of the material synthesis procedures and discuss their feasibility. Furthermore, we explore 89,578 paragraphs, including the synthesis procedure extracted by our system in 14,300 unlabeled materials science literature to demonstrate the utility of our system and analyze the extracted material synthesis procedures.

Methods

We propose 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 proposed extraction pipeline system is shown in Figure 2. 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.

Material synthesis procedure

We define synthesis procedures using a flow graph that is adapted from the previous studies of Kuniyoshi et al. [10] and Mysore et al. [20]. 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 [20] 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

Table 1 Definitions and statistics of entity labels in the materials science procedural text corpus [20]

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	

between them are directly annotated in the document. Moreover, the data published by Mysore et al. [20] 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.

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 [21] and conditional random field [22] (Bi-LSTM-CRF) model [23] was used to predict

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	

entities using the representation of the pre-training model Mat-ELMo [18], 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. [10], 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.

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. [20] in which the paragraphs containing a procedure were identified first before annotating.

Entity extractor

Following Kuniyoshi et al. [10], entities were extracted by formulating the extraction task as a sequence labeling task with IOB (inside—outside—beginning) tags using

¹a Bi-LSTM-CRF model [23] with Mat-ELMo [18] for the token embeddings. Mat-¹
²ELMo is an ELMo (embeddings from language models) [24] model pretrained on²
³the materials literature. We adopted this model although there are several other³
⁴methods for entity extraction [25] because the survey conducted by Kuniyoshi et⁴
⁵al. [10] confirmed that this method is effective for extracting synthesis procedures.⁵
⁶ The input text was tokenized and embedded into a dense vector representation⁶
⁷for each token using Mat-ELMo, as follows. 7

$$\begin{array}{l} 8 \\ 9 \end{array} \quad \mathbf{e} = \text{Mat-ELMo}(\mathbf{t}), \quad (1) \quad \begin{array}{l} 8 \\ 9 \end{array}$$

¹⁰ 10
¹¹where $\mathbf{t} = [t_1, t_2, \dots, t_L]$ is a list of tokens with length L and $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_L]$ ¹¹
¹²is a list of embeddings for each token. The probabilities of classes for tokens are¹²
¹³obtained from LSTM. 13

$$\begin{array}{l} 14 \\ 15 \end{array} \quad \mathbf{p} = \text{LSTM}(\mathbf{e}), \quad (2) \quad \begin{array}{l} 14 \\ 15 \end{array}$$

¹⁶where $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_L]$ is a list of probabilities of classes for the tokens. The CRF¹⁶
¹⁷is applied to $\mathbf{p}$ to determine the class from the probability. 17

$$\begin{array}{l} 18 \\ 19 \end{array} \quad \mathbf{c} = \text{CRF}(\mathbf{p}), \quad (3) \quad \begin{array}{l} 18 \\ 19 \end{array}$$

²⁰ 20
²¹where $\mathbf{c} = [c_1, c_2, \dots, c_L]$ is a list of classes for the tokens in the input text. We²¹
²²maximized the log-likelihood of the prediction of sequences to train this model. 22

²³ In addition to the entities in the corpus, we introduced a TARGET-MATERIAL²³
²⁴class and extracted the entities using the entity extractor. Moreover, we automat-²⁴
²⁵ically induced the TARGET-MATERIAL entities by considering the MATERIAL enti-²⁵
²⁶ties connected by the RECIPE_TARGET relation as the TARGET-MATERIAL entities.²⁶
²⁷This is necessary because it is difficult to develop a simple rule to distinguish the²⁷
²⁸RECIPE_TARGET relation from other relations in Section , and neural methods are²⁸
²⁹suitable for distinguishing the relation. By introducing the TARGET-MATERIAL en-²⁹
³⁰tities, we can extract the RECIPE_TARGET relation using a rule as in Section . 30

³¹*Rule-based relation extractor* 31

³²Our relation extractor is a rule-based model used to extract the relation between an³²
³³entity pair. We adapted the rules of Kuniyoshi et al. [10] from the materials science³³

¹procedural text corpus [20], which depends on the labels of the entities of an entity¹
²pair, the distance between an entity pair, and the order of occurrence of the entities.²
³According to the combination of labels of the entities, our rules are divided into³
⁴three types: OPERATION–OPERATION, OPERATION–MATERIAL, and other relations.⁴
⁵In the following description of the rules, the starting point of a relation is called the⁵
⁶head, the ending point is called the tail, and an edge is denoted as HEAD–TAIL. ⁶
⁷ ⁷
⁸ OPERATION–OPERATION The relation OPERATION–OPERATION takes only a ⁸
⁹ NEXT_OPERATION label, which indicates the progress of the operation. ⁹
¹⁰ ¹⁰
¹¹ NEXT_OPERATION: We assumed that OPERATION is described in the order of the ¹¹
¹² operation progression. Therefore, OPERATION entities are connected in the order in ¹²
¹³ which they appear. ¹³
¹⁴ ¹⁴
¹⁵ OPERATION–MATERIAL The OPERATION–MATERIAL relations mean OPERATION¹⁵
¹⁶ is performed using MATERIAL. Moreover, the relation classes are divided into five¹⁶
¹⁷ categories depending on the function of the MATERIAL: (1) RECIPE_PRECURSOR¹⁷
¹⁸ indicates the input of a material, (2) RECIPE_TARGET indicates the generation of a¹⁸
¹⁹ product, (3) PARTICIPANT_MATERIAL indicates the generation of an intermediate¹⁹
²⁰ product, (4) SOLVENT_MATERIAL indicates the solvent material of an operation,²⁰
²¹ and (5) ATMOSPHERIC_MATERIAL indicates the atmosphere of an operation. ²¹
²² We classified SOLVENT_MATERIAL, ATMOSPHERIC_MATERIAL, and PARTICI-²²
²³ PANT_MATERIAL labels based on dictionary matches because the words in MATE-²³
²⁴ RIAL are distinctive. A dictionary was prepared for each label. The relations connect²⁴
²⁵ a MATERIAL to the nearest OPERATION in the sentence if MATERIAL matches in the²⁵
²⁶ dictionary because these relations take specific MATERIAL entities. The dictionaries²⁶
²⁷ are listed in Tables 3, 4, and 5. ²⁷
²⁸ For a RECIPE_TARGET label, as MATERIAL is the target to be extracted as²⁸
²⁹ TARGET-MATERIAL by the entity extractor in the previous section, the relation²⁹
³⁰ extractor connects the TARGET-MATERIAL to the nearest OPERATION. ³⁰
³¹ RECIPE_PRECURSOR connects all MATERIAL except TARGET-MATERIAL that do ³¹
³² not match the dictionary of SOLVENT_MATERIAL, ATMOSPHERIC_MATERIAL, and ³²
³³ PARTICIPANT_MATERIAL to the nearest OPERATION. ³³

Table 3 Dictionary for Solvent_Material

	Regular expression	
1		1
2	. *water.*	2
3	. *(n alcoh glyc)ol.*	3
4	. *NaOH.*	4
5	. *HCl.*	5
6	. *acetone.*	6
7	. *acid.*	7
8	. *H2O.*	8
9	. *chloroform.*	9
10	. *sodium hydroxide.*	10
11	. *DMF.*	11
12	. *THF.*	12
13	. *N,N-dimethylformamide.*	13
14	. *hexane.*	14
15	. *toluene.*	15
16	. *H2SO4.*	16
17	. *EtOH.*	17

Table 4 Dictionary for Atmospheric_Material

	Regular expression	
18	. *air.*	18
19	. *argon.*	19
20	. *Ar.*	20
21	. *N2.*	21
22	. *nitrogen.*	22
23	. *H2.*	23
24	. *oxygen.*	24
25	. *O2.*	25
26	. *hydrogen.*	26
27	. *CH4.*	27
28	. *H2S.*	28
29	. *He.*	29

23 *Remaining relations* The remaining nine relation labels are defined between the
 24 other pairs of entity labels: PROPERTY_OF, which indicates the condition of a mate-
 25 rial; CONDITION_OF, which indicates the condition of an operation; NUMBER_OF,
 26 which indicates the relationship between a number and a unit; AMOUNT_OF, which
 27 indicates a condition of a quantity; TYPE_OF, which indicates type of the numer-
 28 ical condition; BRAND_OF, which indicates the brand of a material or equipment;
 29 APPARATUS_OF, which indicates the equipment used in an operation; APPARA-
 30 TUS_ATTR_OF, which indicates the numerical condition of the equipment; and DE-
 31 SCRIPTOR_OF, which indicates other conditions. For these labels, the rules are de-
 32 fined based only on the labels of the head and tail entities and the distance between
 33 them. We explain the detailed rules in the remainder of this section.

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.*

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-

¹DESCRIPTOR is a head, SYNTHESIS-APPARATUS or CHARACTERIZATION-APPARATUS¹
²can be a tail, but only the nearest SYNTHESIS-APPARATUS in the sentence is con-²
³nected because CHARACTERIZATION-APPARATUS is an apparatus for measuring³
⁴characteristics and detailed descriptions are rarely provided. ⁴

⁵ APPARATUS_OF: This relation connects SYNTHESIS-APPARATUS and CHARACTERIZATION-⁵
⁶APPARATUS with the nearest OPERATION with the priority given to the OPERATION⁶
⁷that appears before the APPARATUS in the sentence. ⁷

⁸ TYPE_OF: PROPERTY-TYPE and APPARATUS-PROPERTY-TYPE are connected to⁸
⁹the nearest PROPERTY-UNIT and APPARATUS-UNIT in the sentence with this rela-⁹
¹⁰tion, respectively. When CONDITION-TYPE is a head, it is connected to the nearest¹⁰
¹¹CONDITION-UNIT that appears before the CONDITION-TYPE in the sentence. ¹¹

¹² BRAND_OF: The relation connects BRAND to the nearest entities that may¹²
¹³have brands (i.e., MATERIAL, NONRECIPE-MATERIAL, SYNTHESIS-APPARATUS,¹³
¹⁴and CHARACTERIZATION-APPARATUS) in the sentence. ¹⁴

¹⁵ APPARATUS_ATTR_OF: APPARATUS-UNIT is connected to the nearest SYNTHESIS-¹⁵
¹⁶APPARATUS or CHARACTERIZATION-APPARATUS. ¹⁶

¹⁷ COREF_OF: The relation is not detected by the rules because it is difficult to¹⁷
¹⁸describe rules. ¹⁸

¹⁹ ¹⁹
²⁰*Selecting paragraphs* ²⁰

²¹Although extractors were applied to all paragraphs, only few paragraphs may con-²¹
²²tain procedures. We selected such paragraphs based on the extracted entities and²²
²³relations. ²³

²⁴ Several deep learning methods have been proposed to select such paragraphs [26].²⁴
²⁵However, we employed a simple approach to select these paragraphs using the ex-²⁵
²⁶tracted procedures. As the target of this research is to extract procedural sequences,²⁶
²⁷the extracted procedures require the inclusion of sequences wherein target materi-²⁷
²⁸als are generated from other materials. Therefore, we selected the paragraphs that²⁸
²⁹contain RECIPE_PRECURSOR, RECIPE_TARGET, and OPERATION entities that in-²⁹
³⁰tervene between them. ³⁰

³¹Results ³¹

³²In our experiments, we first evaluated our pipeline to determine if it could be used³²
³³in practice. In the evaluation of the pipeline, we evaluated separately each module,³³

¹the entity extractor, the relation extractor, and the selection of the paragraphs,¹
²and finally evaluated the entire pipeline. Then, we applied the pipeline to extract²
³synthesis procedures from a large set of documents that we collected and analyzed³
⁴the extracted procedures.⁴

⁵

⁵

⁶Evaluation settings for entity and relation extractors⁶

⁷We utilized the publicly available materials science procedural text corpus [20] to⁷
⁸train and develop the entity and relation extractors using the training data of this⁸
⁹corpus and by observing these training data, respectively.⁹

¹⁰ Flair [27], a library of machine learning for natural language processing, was used¹⁰
¹¹to implement the entity extractor. Meanwhile, the pretraining model Mat-ELMo [18]¹¹
¹²was fixed during training because of the small size of the corpus. Moreover, we¹²
¹³employed a stochastic gradient descent method for training. The learning rate was¹³
¹⁴halved from the initial value of 0.1 when there was no improvement in performance¹⁴
¹⁵over three epochs, and the method was stopped when the learning rate became¹⁵
¹⁶smaller than 0.0001. A model that showed the best performance on the development¹⁶
¹⁷data was used.¹⁷

¹⁸ The individual evaluation of the entity and relation extractors was performed¹⁸
¹⁹using the test data of the corpus in which precision, recall, micro-F, which is an¹⁹
²⁰overall F-score, and macro-F, which is the averaged F-score of individual classes,²⁰
²¹were employed as the evaluation metrics.²¹

²² We compared our entity extractor with BERT (Bidirectional Encoder Representa-²²
²³tions from Transformers)-base [28] and ELECTRA (efficiently learning an encoder²³
²⁴that classifies token replacements accurately)-base [29], which have shown excellent²⁴
²⁵performance in various natural language processing tasks, to verify the effective-²⁵
²⁶ness of the setting for the corpus. BERT is a transformer-based [30] model that is²⁶
²⁷pretrained by masked language modeling in a large number of studies. Conversely,²⁷
²⁸although ELECTRA is similar to BERT, it has shown higher performance using²⁸
²⁹an improved training method. To evaluate these transformer-based models, the em-²⁹
³⁰beddings of our entity extractor were replaced with these models, and the training³⁰
³¹was performed in the same way.³¹

³² We evaluated the performance of the relation extractor separately from the entity³²
³³extractor by assigning gold entities to verify the validity of the developed rules.³³

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

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.

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.

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

Class	Entity given			Pipeline		
	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

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.

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

¹correct entities are given, classes that are critically related to the procedures such¹
²as NEXT_OPERATION and RECIPE_TARGET can be extracted with high accuracy.²
³Conversely, the extraction performance of classes such as PARTICIPANT_MATERIAL³
⁴and SOLVENT_MATERIAL is low. COREF_OF was not extracted at all because we⁴
⁵do not have the rule for this relation.⁵

⁶ For the pipeline result, the class of NEXT_OPERATION, which is a key relation⁶
⁷for the sequence of operations on the procedure, exceeds 0.6. This indicates that⁷
⁸the main flow of the procedure can be extracted to some extent even if e error has⁸
⁹propagated from the previous entity extractor. By contrast, some classes such as⁹
¹⁰PROPERTY_OF, BRAND_OF, and APPARATUS_ATTR_OF show extremely low per-¹⁰
¹¹formance.¹¹

¹² Large-scale extraction¹³

¹⁴We collected articles from the Journal of Materials Chemistry A (JMCA), which¹⁴
¹⁵focuses on areas such as batteries, fuel cells, sustainable materials, photovoltaics,¹⁵
¹⁶supercapacitors, and water splitting, published by The Royal Society of Chemistry,¹⁶
¹⁷as the source of literature for large-scale extraction. We purchased all articles from¹⁷
¹⁸2015 to 2019 in XML format. The total number of articles was 14,310.¹⁸

¹⁹ We applied the pipeline extractor to all articles. First, as a preprocessing step,¹⁹
²⁰we extracted all paragraphs from the articles in XML format. Then, entities and²⁰
²¹relations from the paragraphs were also extracted and the paragraphs that describe²¹
²²the procedures were selected. From 14,310 studies, we obtained 347,480 paragraphs²²
²³in total; this was reduced to 89,578 after selection, in which one-third of the total²³
²⁴paragraphs were determined to contain procedures. Surprisingly, each literature²⁴
²⁵mentioned six procedures on average, indicating that there is plenty of information²⁵
²⁶about material synthesis procedures in the text even though there are miscounts²⁶
²⁷due to prediction errors. The statistics of the extracted entities and relations are²⁷
²⁸shown in Table 9 and 10, respectively.²⁸

²⁹ Discussion³⁰

³⁰We analyzed the procedures extracted from a large body of literature to investigate³⁰
³¹their characteristics. In this investigation, we aimed to verify the reasonableness of³¹
³²the extracted procedures and determine whether they can be used to obtain useful³²
³³information for material development or other tasks.³³

Table 9 Statistics of entities in the extracted procedures

1			1
	OPERATION	688,311	
2	MATERIAL	912,266	2
3	TARGET-MATERIAL	327,093	3
	NONRECIPE-MATERIAL	231,290	
4	NUMBER	441,297	4
5	PROPERTY-MISC	275,665	5
	PROPERTY-TYPE	43,674	
6	PROPERTY-UNIT	95,409	6
7	AMOUNT-MISC	5,232	7
	AMOUNT-UNIT	108,585	
8	CONDITION-MISC	36,707	8
	CONDITION-TYPE	20,342	
9	CONDITION-UNIT	171,513	9
10	SYNTHESIS-APPARATUS	47,766	10
	CHARACTERIZATION-APPARATUS	87,368	
11	APPARATUS-UNIT	5,326	11
12	APPARATUS-PROPERTY-TYPE	861	12
	MATERIAL-DESCRIPTOR	293,905	
13	APPARATUS-DESCRIPTOR	16,154	13
14	BRAND	19,920	14
	META	31,007	
15	REFERENCE	21,215	15
16	All entities	3,880,906	16

Frequent subgraphs

We checked the subgraphs that frequently appear using gSpan [31], which is an algorithm for mining frequent subgraphs in the extracted paragraphs, and whether they were reasonable.

The top 10 frequent subgraphs that contain OPERATION entities and appear more than 50 times are shown in Figure 3 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.

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.

Table 10 Statistics of relations in the extracted procedure

1			1
	NEXT_OPERATION	598,733	
2	RECIPE_PRECURSOR	457,558	2
3	RECIPE_TARGET	231,931	3
	PARTICIPANT_MATERIAL	143,200	
4	SOLVENT_MATERIAL	59,213	4
5	ATMOSPHERIC_MATERIAL	55,253	5
	NUMBER_OF	401,816	
6	PROPERTY_OF	329,280	6
7	AMOUNT_OF	111,233	7
	CONDITION_OF	166,740	
8	DESCRIPTOR_OF	308,773	8
	APPARATUS_OF	101,001	
9	APPARATUS_ATTR_OF	4,645	9
10	BRAND_OF	21,757	10
	TYPE_OF	33,009	
11	All relations	3,024,142	11

14 As a basic step, we counted the included elements and extracted the element¹⁴
 15 names from left to right, ignoring valences and other numerical values. We filtered¹⁵
 16 out the TARGET-MATERIAL that was determined to be composed of one element¹⁶
 17 or none to avoid the effect of erroneous extraction. ¹⁷

18 ¹⁸
 19 The frequency of the elements is summarized in Figure 4. Moreover, the elements¹⁹
 20 were normalized using the number of TARGET-MATERIAL entities and was scaled²⁰
 21 logarithmically. Elements with a single letter may be counted more than their actual²¹
 22 occurrences owing to errors in parsing elements contained in TARGET-MATERIAL.²²
 23 For example, the tin (IV) oxide (SnO₂) nanoparticles mentioned as “SnO₂NP” are²³
 24 recognized as a combination of SnO₂, nitrogen (N), and phosphorus (P). However,²⁴
 25 both N and P represent nanoparticles. The results show that elements O, C, N, P,²⁵
 26 and S, which are commonly contained in various compounds, and Ni and Ti, which²⁶
 27 are conventionally used in electrodes, appear frequently. ²⁷

28 We calculated the importance of each operation for each element in TARGET-²⁸
 29 MATERIAL with the term frequency-inverse document frequency (TF-IDF) to ana-²⁹
 30 lyze the correlations among procedures. Here, the operations in the procedure are³⁰
 31 considered as terms in a document, and each element in the TARGET-MATERIAL³¹
 32 is considered as a document. A document $\mathcal{D}_e$ corresponding to an element e is de-³²
 33 fined in which a set of all procedures is $\mathcal{P}$, a set of lemmatized OPERATIONS in a ³³

Table 11 The top 20 Operations ranked by the TF-IDF score for each element in Target-Material.

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

procedure p is $\mathcal{O}_p$, and a set of elements in TARGET-MATERIAL of a procedure p is $\mathcal{E}_p$.

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

$\mathbb{U}[\cdot]$ 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 $\text{TFIDF}(d, t)$ is defined as a function of a term t in a document d , where $\text{count}(d, t)$ 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.

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

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

$$\text{TFIDF}(d, t) = \text{TF}(d, t) \cdot \text{IDF}(t) \quad (7)$$

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 4. We lemmatized OPERATION using scispaCy [32] before calculating the TF-IDF score. Moreover, we observed the general OPERATION in the top

10 OPERATION. However, the 11th to 20th OPERATION are characteristic of each¹
2element. We used the OPERATION that occurs with the O element, which is one of²
3the most common elements in the analysis. In addition, “grow†”, which belongs to³
4the columns of Co and Ni, is a characteristic OPERATION as they are used when⁴
5sheets or crystals are created. “anneal*”, which is seen in metallic elements, is a⁵
6unique OPERATION for metals. These aspects are correct for the knowledge gained⁶
7from the extracted procedures. 7

8 Case study of extracted procedures 8

9
10An example of an extraction procedure is shown in Figure 1. In this study, we sam-10
11pled 10 documents and checked 64 procedures that were included in the sampled11
12documents to test the feasibility of extracted procedures. The obtained procedures12
13that were manually chosen from the sampled procedures provided information and13
14conditions about each operation, as well as how the TARGET-MATERIAL was pro-14
15duced from the input material. However, there were several errors found, such as15
16“0.075 mmol” in the middle of the paragraph that was not extracted as an entity16
17and “water” and “ethanol” that were the SOLVENT_MATERIAL of “washed” were17
18incorrectly connected to “placed” by the relation extractor. However, the general18
19framework of the procedure was extracted, and the analysis of a large number of19
20extracted procedures is expected to contribute to the development of materials. 20

21 Conclusions 21

22In this study, we proposed a pipeline system for extracting synthesis procedures from22
23the literature. We applied an entity extractor, which consisted of Mat-ELMo and23
24Bi-LSTM-CRF models, and a rule-based relation extractor to extract and structure24
25the synthesis procedures as a graph. Our pipeline system performed a large-scale25
26extraction of procedures to analyze the synthesis procedures. We confirmed that26
27the extracted procedures include reasonable operations such as “drying ethanol at27
2860 degrees,” and the statistics show the characteristic operations of the elements28
29(e.g., the “anneal” operation for metallic elements). 29

30For further study, we will attempt to improve the performance of the extraction30
31of procedures and provide more insights into the synthesis procedures, represented31
32as procedures for discovering new materials. Since our extractor is able to extract32
33only a limited number of classes with high performance, improving the extraction33

¹ performance of the system is our next challenge to obtain more accurate procedures.	¹
² In addition, we will continue to analyze the procedures extracted by the system and	²
³ develop materials based on the obtained knowledge.	³
⁴	⁴
⁵ Abbreviations	⁵
⁶ LSTM Long Short-Term Memory	⁶
⁷ CRF Conditional Random Field	⁷
⁸ ELMo Embeddings from Language Models	⁸
⁹ IOB tag Inside-Outside-Begging tag	⁹
¹⁰ BERT Bidirectional Encoder Representations from Transformers	¹⁰
¹¹ ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately	¹¹
¹² P Precision	¹²
¹³ R Recall	¹³
¹⁴ F F-score	¹⁴
¹⁵ Declarations	¹⁵
Ethics approval and consent to participate	
¹⁶ Not applicable.	¹⁶
¹⁷	¹⁷
Consent for publication	
¹⁸ Not applicable.	¹⁸
¹⁹	¹⁹
Availability of data and materials	
²⁰ The data that support the findings of this study are available from the Royal Society of Chemistry but restrictions	²⁰
²¹ apply to the availability of these data, which were used under license for the current study, and so are not publicly	²¹
²² available. Data are however available from the authors upon reasonable request and with permission of the Royal	²²
²³ Society of Chemistry.	²³
²⁴	²⁴
Competing interests	
²⁵ The authors declare that they have no competing interests.	²⁵
²⁶	²⁶
Funding	
²⁷ Not applicable.	²⁷
²⁸	²⁸
Authors' contributions	
²⁹ All authors have contributed in the production of this manuscript. FK, JO, and MM supervised all steps of the work.	²⁹
³⁰ KM built the system and conducted the experiments. The paper was written by KM, FK, and MM, and was read	³⁰
³¹ and approved by all authors. All authors contributed to the production of the manuscript. FK, JO, and MM	³¹
³² supervised all steps of the work. KM built the system and carried out the experiments. The paper was written by	³²
³³ KM, FK, and MM, and read and approved by all authors.	³³
³⁴	³⁴
Acknowledgements	
³⁵ Not applicable.	³⁵
³⁶	³⁶
Author details	
³⁷ ¹ Toyota Technological Institute, 468-8511 Nagoya, Aichi, Japan. ² Panasonic-AIST Advanced AI Cooperative	³⁷
³⁸ Research Laboratory, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan. ³ Panasonic	³⁸
³⁹ Corporation, 571-8501 Kadoma-shi, Osaka, Japan.	³⁹

References

1. Draxl, C., Scheffler, M.: Nomad: The fair concept for big data-driven materials science. *Mrs Bulletin* **43**(9), 676–682 (2018)
2. Schleder, G.R., Padilha, A.C., Acosta, C.M., Costa, M., Fazzio, A.: From dft to machine learning: recent approaches to materials science—a review. *Journal of Physics: Materials* **2**(3), 032001 (2019)
3. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.: Machine learning for molecular and materials science. *Nature* **559**(7715), 547–555 (2018)
4. Curtarolo, S., Hart, G.L., Nardelli, M.B., Mingo, N., Sanvito, S., Levy, O.: The high-throughput highway to computational materials design. *Nature materials* **12**(3), 191–201 (2013)
5. Rickman, J.M., Lookman, T., Kalinin, S.V.: Materials informatics: From the atomic-level to the continuum. *Acta Materialia* **168**, 473–510 (2019)
6. Belsky, A., Hellenbrandt, M., Karen, V.L., Luksch, P.: New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science* **58**(3), 364–369 (2002)
7. Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.H., Nelson, L.J., Hart, G.L., Sanvito, S., Buongiorno-Nardelli, M., et al.: Aflowlib. org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **58**, 227–235 (2012)
8. Swain, M.C., Cole, J.M.: Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling* **56**(10), 1894–1904 (2016)
9. Kim, E., Huang, K., Tomala, A., Matthews, S., Strubell, E., Saunders, A., McCallum, A., Olivetti, E.: Machine-learned and codified synthesis parameters of oxide materials. *Scientific data* **4**(1), 1–9 (2017)
10. Kuniyoshi, F., Makino, K., Ozawa, J., Miwa, M.: Annotating and extracting synthesis process of all-solid-state batteries from scientific literature. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1941–1950. European Language Resources Association, Marseille, France (2020). <https://www.aclweb.org/anthology/2020.lrec-1.239>
11. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* **1**(1), 011002 (2013)
12. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., Ceder, G.: Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data* **6** (2019). doi:10.1038/s41597-019-0224-1
13. Lookman, T., Balachandran, P.V., Xue, D., Yuan, R.: Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **5**(1), 1–17 (2019)
14. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., Olivetti, E.: Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials* **29**(21), 9436–9444 (2017)
15. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., Wolverton, C.: Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom* **65**(11), 1501–1509 (2013)
16. Raccuglia, P., Elbert, K.C., Adler, P.D., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J., Norquist, A.J.: Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**(7601), 73–76 (2016)
17. Mahbub, R., Huang, K., Jensen, Z., Hood, Z.D., Rupp, J.L.M., Olivetti, E.A.: Text mining for processing conditions of solid-state battery electrolytes. *Electrochemistry Communications* **121**, 106860 (2020). doi:10.1016/j.elecom.2020.106860
18. Kim, E., Jensen, Z., van Grootel, A., Huang, K., Staib, M., Mysore, S., Chang, H.-S., Strubell, E., McCallum, A., Jegelka, S., Olivetti, E.: Inorganic materials synthesis planning with literature-trained neural networks. *Journal of Chemical Information and Modeling* **60**(3), 1194–1201 (2020). doi:10.1021/acs.jcim.9b00995. PMID: 31909619. <https://doi.org/10.1021/acs.jcim.9b00995>
19. Segler, M.H., Preuss, M., Waller, M.P.: Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* **555**(7698), 604–610 (2018)
20. Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H.-S., Strubell, E., Flanigan, J., McCallum, A., Olivetti, E.: The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic

- 1 structures. In: Proceedings of the 13th Linguistic Annotation Workshop, pp. 56–64. Association for
2 Computational Linguistics, Florence, Italy (2019). doi:10.18653/v1/W19-4007.
https://www.aclweb.org/anthology/W19-4007
- 3 21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997).
4 doi:10.1162/neco.1997.9.8.1735
- 5 22. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting
6 and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning.
7 ICML '01, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
- 8 23. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. *CoRR* **abs/1508.01991**
9 (2015). 1508.01991
- 10 24. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized
11 word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the
12 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp.
13 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018).
14 doi:10.18653/v1/N18-1202. https://www.aclweb.org/anthology/N18-1202
- 15 25. Eltyeb, S., Salim, N.: Chemical named entities recognition: a review on approaches and applications. *Journal of*
16 *cheminformatics* **6**(1), 1–12 (2014)
- 17 1226. Huo, H., Rong, Z., Kononova, O., Sun, W., Botari, T., He, T., Tshitoyan, V., Ceder, G.: Semi-supervised
18 machine-learning classification of materials synthesis procedures. *npj Computational Materials* **5**(1), 1–7 (2019)
- 19 27. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th
20 International Conference on Computational Linguistics, pp. 1638–1649 (2018)
- 21 28. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for
22 language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the
23 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),
24 pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019).
25 doi:10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
- 26 29. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators
27 rather than generators. In: ICLR (2020). https://openreview.net/pdf?id=r1xMH1BtvB
- 30 1930. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.:
Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.,
31 Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 30, pp. 5998–6008. Curran Associates,
32 Inc., ??? (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
- 33 31. Yan, X., Han, J.: gspan: graph-based substructure pattern mining. In: 2002 IEEE International Conference on
Data Mining, 2002. Proceedings., pp. 721–724 (2002). doi:10.1109/ICDM.2002.1184038
- 32 32. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural
33 Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327.
Association for Computational Linguistics, Florence, Italy (2019). doi:10.18653/v1/W19-5034.
arXiv:1902.07669. https://www.aclweb.org/anthology/W19-5034
- 34 33. Zhang, J., Xiao, B., Liu, X., Liu, P., Xi, P., Xiao, W., Ding, J., Gao, D., Xue, D.: Copper dopants improved the
hydrogen evolution activity of earth-abundant cobalt pyrite catalysts by activating the electrocatalytically inert
sulfur sites. *J. Mater. Chem. A* **5**, 17601–17608 (2017). doi:10.1039/C7TA05433E

Figures

example.pdf

Figure 1 Illustration of the material synthesis procedure adapted from Zhang et al. [33]

Figures

Figure 1

Illustration of the material synthesis procedure adapted from Zhang et al. [33]

Figure 2

Overview of the extraction pipeline system

Figure 3

Frequent sub-graphs in the extracted procedures