Analyzing Research Trends
in Inorganic Materials Literature Using NLP

Fusataka Kuniyoshi[1,3][0000−0001−5914−009X] ✉,
Jun Ozawa[1,3][0000−0003−4212−9008], and Makoto Miwa[2,3][0000−0002−2330−6972]

[1] Panasonic Corporation, 1006, Oaza Kadoma, Kadoma-shi, Osaka 571-8501, Japan
[2] Toyota Technological Institute, 2–12–1, Hisakata, Tempaku-ku, Nagoya, 468–8511,
Japan
[3] National Institute of Advanced Industrial Science and Technology (AIST), 2–3–26,
Aomi, Koto-ku, Tokyo, 135–0064, Japan
kuniyoshi.fusataka@gmail.com, {ozawa.jun, makoto.miwa}@aist.go.jp

Abstract. In the field of inorganic materials science, there is a growing demand to extract knowledge such as physical properties and synthesis processes of materials by machine-reading a large number of papers. This is because materials researchers refer to produce promising terms of experiments for material synthesis. However, there are only a few systems that can extract material names and their properties. This study proposes a large-scale natural language processing (NLP) pipeline for extracting material names and properties from materials science literature to enable the search and retrieval of results in materials science. Therefore, we propose a label definition for extracting material names and properties and accordingly build a corpus containing 836 annotated paragraphs extracted from 301 papers for training a named entity recognition (NER) model. Experimental results demonstrate the utility of this NER model; it achieves successful extraction with a micro-F1 score of 78.1%. To demonstrate the efficacy of our approach, we present a thorough evaluation on a real-world automatically annotated corpus by applying our trained NER model to 12,895 materials science papers. We analyze the trend in materials science by visualizing the outputs of the NLP pipeline. For example, the country-by-year analysis indicates that in recent years, the number of papers on "MoS_2," a material used in perovskite solar cells, has been increasing rapidly in China but decreasing in the United States. Further, according to the conditions-by-year analysis, the processing temperature of the catalyst material "PEDOT:PSS" is shifting below 200 °C, and the number of reports with a processing time exceeding 5 h is increasing slightly.

Keywords: Natural Language Processing · Text Mining · Materials Informatics

1 Introduction

Materials science literature includes considerable information such as material names and their properties described in natural language. Therefore, the auto-

matic extraction of the details necessary to reproduce and validate materials synthesis processes in a materials science laboratory remains difficult and requires extensive human intervention. The automatic compilation of such literature into a structured form could enable realizing a data-driven materials discovery system that does not require human intervention; such a system could become a key enabler in the design and discovery of novel materials. In this regard, named entity recognition (NER) is helpful, as it seeks to locate spans and classify named entities in unstructured text into predefined categories such as material names.

NER has already found many applications in materials science. For example, material names have been linked to their properties, such as characteristic values or their structures, through a combination of database lookup and the parsing of systematic nomenclature to create reader-friendly semantically enhanced literature [21, 30]. Further, NER has been linked to material information retrieval techniques to search for materials similar to a query material from corpora [23–25, 6] or to predict the characteristic values of a query material [22, 15].

A technique that can extract natural language characteristic values and link a material name to a machine-readable representation will find importance in many practical applications. We believe that current research in this area is hampered by the lack of available annotated corpora.

In this study, we propose a natural language processing (NLP)-based approach to analyze trends in materials for developments in materials science (see Figure 1). Toward this end, we propose a pipeline that integrates an NER model and a numeric normalization module. To evaluate this pipeline, we annotate 836 paragraphs extracted from 301 papers to extract material terminology and conduct initial analyses to extract material data from 12,895 unlabeled full-text literature. Through this evaluation, we demonstrate the reliability of our NLP framework by presenting the detailed NER model training process and by showing the detailed evaluation of the trained NER model. Our NER model can extract material names and several important properties such as temperature, time, conductivity, and activation energy. To demonstrate the utility of our annotated corpus and analyze the research trends, we explore the extracted outputs of our NLP framework from 12,895 unlabeled materials science papers. This study makes the following contributions:

– We propose a manually annotated corpus and an NLP framework for extracting material names with properties using an NER tagger and apply a numeric normalization module to NER outputs.
– We evaluate the reliability of the NLP system using by showing the detailed process of training the NER model with sufficient evaluation.
– We demonstrate the analysis for observing the research trends using our NLP outputs for material knowledge discovery from scientific literature.
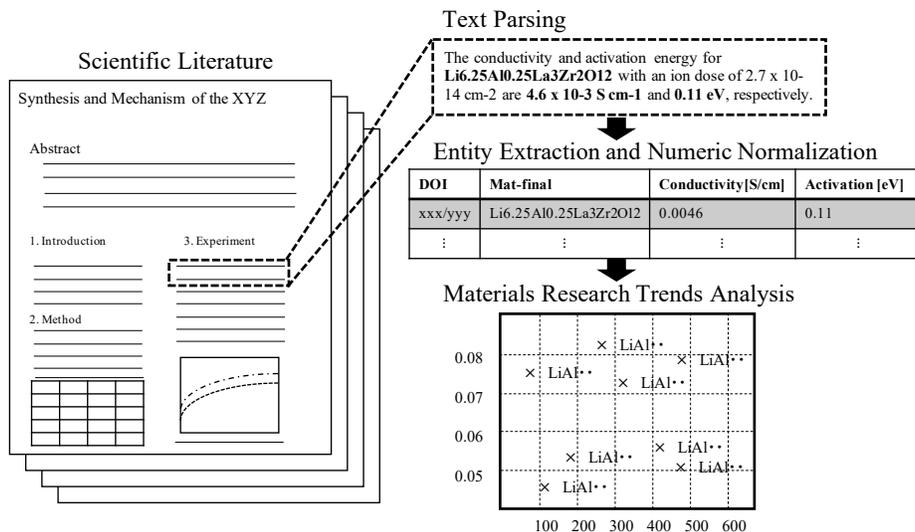
Text Parsing

The conductivity and activation energy for **Li6.25Al0.25La3Zr2O12** with an ion dose of 2.7 x 10-14 cm-2 are **4.6 x 10-3 S cm-1** and **0.11 eV**, respectively.

Scientific Literature

Synthesis and Mechanism of the XYZ

Abstract

1. Introduction   3. Experiment

2. Method

Entity Extraction and Numeric Normalization

DOI	Mat-final	Conductivity[S/cm]	Activation [eV]
xxx/yyy	Li6.25Al0.25La3Zr2O12	0.0046	0.11
⋮	⋮	⋮	⋮

Materials Research Trends Analysis

Fig. 1: Overview of pipeline for extraction.

2   Related Work

Many NLP systems and language resources are available for extracting different types of information from scientific literature, such as identifying drug names [13], discovering drugs [28], examining the side-effects of drugs [27], extracting biomedical terminology [19, 29, 32] or events [20, 34, 35], and extracting wet-lab protocols [7] are some of these examples.

In inorganic materials science, text mining is mainly used to search for a domain-specific material name or for classifying materials by their type, such as inorganic materials in general [23, 25, 10, 30, 33, 36], oxides [24], superconductors [26, 15], zeolites [22], and battery materials [6, 2, 31].

However, in inorganic materials science, few practical systems have been proposed to extract material names from a large number of papers by associating them with their property values. In this study, we propose an NLP system for extracting material names and properties.

3   Corpus Preparation

3.1   Definition of Types

Our proposed annotation scheme is based on Kuniyoshi's annotation scheme for materials synthesis processes [2]. We used 12 labels that were defined to annotate spans of text; these represent the materials, operations, and properties. In the list, we segmented the roles of materials (MAT), operations (OPE), properties (PROP), and characteristics (CHARA).

Mat-Final represents the final material (or product) of the material synthesis process; for example, "A solid solution of the lithium superionic conductor Li$_{10+}\delta$Ge$_{1+}\delta$P$_{2-}\delta$S$_{12}$ ($0 \leq \delta \leq 0.35$) was synthesized ..."

Mat-Solvent is a liquid that is used to dissolve substances and create solutions; for example, "Ga$_2$O$_3$ (99.999%) were ground and homogenized in ethanol."

Mat-Start is a raw material used to synthesize the final material; for example, "Precursor powders (10 g) containing a stoichiometric mixture of La$_2$O$_3$ (99.998%)."

Ope represents an individual action performed by the experimenters. It is often represented by verbs; for example, "Carbon black was dried at 80°C."

Prop-Equip represents equipment for analyzing a material; for example, "... spectrum analysis of the films was carried out on a UV-Vis spectrophotometer."

Prop-Maker represents a manufacturer of equipment or material powder; for example, "m-Cresol was obtained from Sigma-Aldrich."

Prop-Method represents a method to analyze a material sample; for example, "The surface morphologies of the relevant membranes were studied by using a high-resolution field-emission scanning electron microscopy."

Prop-Temp represents a temperature condition associated with an operation; for example, "... finally dried at 80°C in vacuum for 5 h."

Prop-Time represents a time condition associated with an operation; for example, "... finally dried at 80°C in vacuum for 5 h."

Chara-Name represents a characteristic name to classify characteristic values; for example, "... the glass ceramic has a room-temperature ionic conductivity as high as 3×10^{-5} S cm^{-1}."

Chara-Act represents a characteristic value of activation energy. For example, the unit of **activation energy** is eV; then, "The activation energy as a function of the vacancy concentration exhibits a minimum of 0.7 eV ..."

Chara-Cond represents a characteristic value of conductivity. For example, the unit of **conductivity** is S/cm; then, "The ionic conductivity of the prepared pellets is 1.03×10^{-3} S/cm."

3.2   Collecting Literature

Our corpus was constructed from papers published in the Journal of Material Chemistry A (JMCA; Royal Society of Chemistry (RSC))[4] from 2015 to 2019. JMCA focuses on energy and sustainability, and it publishes papers discussing materials such as solar cells, thermoelectric conversion materials, liquid lithium ion batteries (LIBs), and all-solid-state batteries. The RSC provides papers in XML format, in which contents have a hierarchical structure within different nested tags. For example, the <section> tag contains information such as the section title and paragraph. To extract plain text from such XML files, we created an extraction tool that exploits RSC's semantic markup features to extract information such as the title, abstract, and main contents. Then, we stripped

[4] https://www.rsc.org/journals-books-databases/about-journals/journal-of-materials-chemistry-a/

Fig. 2: Example of annotation of experimental section. Text labeling interface to annotate material names and their properties. Example of annotation of experimental sections. This text is referred from the study by Yuan [9].

out the embedded markups to produce the plain text and to create a linear stream of elements containing all data in the papers. These text data were then transferred into a document object comprising subobjects such as title, heading, and paragraph. Further, we automatically extracted paragraphs with their section and subsection titles by using regular expressions for target titles such as "Abstract," "Introduction," "Experimental," and "Conclusion."

3.3 Annotation

One Master's degree staff in the materials science department annotated labels on the 836 paragraphs extracted from 301 papers. Figure 2 illustrates annotations made to the text in the experimental section by using the brat annotation toolkit [8]. The annotated data are converted into the Inside, Outside, Beginning (IOB) scheme, where a token is labeled as I-$*$ if it is inside a named entity of type $*$, O if it is outside of named entities, and B-$*$ if it is at the beginning of an $*$ entity. Therefore, the model is trained to classify each word in a sequence into 25 different labels consisting of one O label and B-$*$ and I-$*$ labels for each of the 12 entity labels. Our corpus is shared at a github repository[5].

4 Approach

This section explains our framework for extracting material data, such as names and property values, from a large number of papers. Our framework (see Figure 3) consists of an NER-based sequence labeling tool and a module that converts a natural language phrase to numeric values.

4.1 Sequence Labeling Architecture

First, we briefly describe bidirectional long short-term memory (BiLSTM), a type of recurrent neural network, and a subsequent conditional random field

---

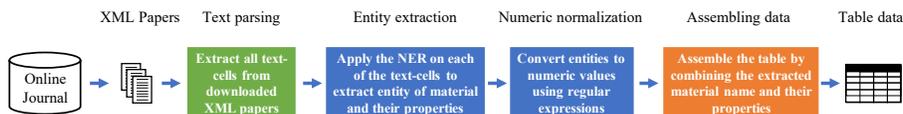[5] https://github.com/BananaTonic/Material_Synthesis_Corpus.git

Fig. 3: Overview of analysis pipeline.

(CRF). Then, we explain the hybrid labeling architecture that is based on a previous study [17, 18].

We extract the output hidden state after and before the word's token in a sentence from the corresponding forward and backward LSTMs to capture semantic-syntactic information from the beginning and ending of the sentence to the token, respectively. Both output hidden states are concatenated to form the final embedding and to capture the semantic-syntactic information of the word itself as well as its surrounding context. Figure 4 shows our proposed NER architecture.

Fig. 4: Proposed NER architecture.

Let the individual tokens in a sentence be $t_0, t_1, ..., t_n$. We define the contextual string embeddings of these tokens as $h_0, h_1, ..., h_n$, where h_t represents the output hidden state of a token t. The final word embeddings are passed to a BiLSTM-CRF sequence labeling module to address downstream sequence labeling tasks.

Calling the inputs to the BiLSTM gives

$$\mathbf{r}_i \simeq [\mathbf{r}_i^f; \mathbf{r}_i^b],$$

where $\mathbf{r}_i^f$ and $\mathbf{r}_i^b$ are forward and backward output states of the BiLSTM, respectively. The final sequence probability is then given by a CRF over the possible sequence labels $\mathbf{y}$:

$$\hat{P}(\mathbf{y}|\mathbf{r}) \propto \prod_{i=1}^{n} \phi(\mathbf{y}, \mathbf{r}),$$

where $\phi(\cdot)$ is a variation Markov Random Field of all clique potentials. Finally, the prediction of the label is given by

$$P(\mathbf{y}_i = j|\mathbf{r}_i) = softmax(\mathbf{r}_i)[j]$$

4.2   Numeric Normalization

We normalize the numeric values in a post-processing step. Although the phrases of the characteristic values extracted by the aforementioned entity extractor represent numerical values, they are annotated in various ways. Therefore, they
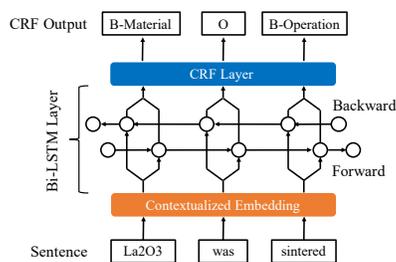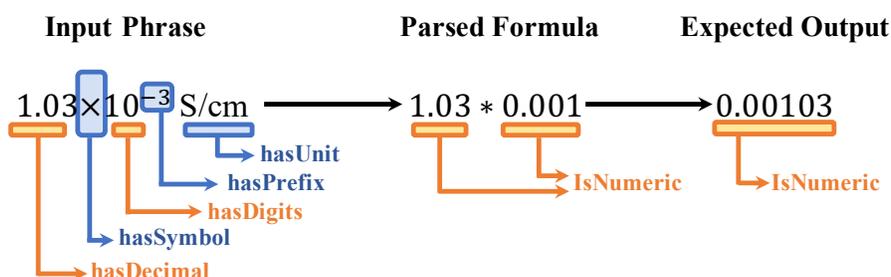
Fig. 5: Example of numeric normalization using pattern matching.

were normalized into a unified format that allows for comparisons and statistical processing. For example, when "1.03×10^{-3} S/cm" was extracted, it was normalized to a value of 0.00103 [S/cm]. In the text of the RSC paper, "10^{-3}" was described as "10⁻³." However, in this study, the XML tag was removed beforehand for simplicity, and the text was converted to the plain text "10 -3." In the case of this string, the string "0.1" was extracted as a value, "×" was extracted as a multiplication sign, and "10 -3" was extracted as the 3rd negative exponent of 10. Then, the extracted numbers were multiplied and normalized into the value "0.00103." Figure 5 shows an example of our numeric normalization, and Table 1 shows the string patterns that can be normalized by this system and their normalization results. The practical workflow of our numeric normalization is as follows: first, when we find measurement patterns in texts, we separate specific expressions into the numeric and unit parts. For example, when "14 °C – room temperature" was found, we replaced it with "14 °C – 22 °C." Next, we split extracted phrases into specific units such as "S/cm" and "°C." For example, when "irradiation times of 3 s to 8 min" was extracted, we split this as "["irradiation times of 3", "to 8"]" and "["s", "min"]". Finally, we extracted each value from split phrases when a unit had numeric patterns before it. For example, when we extracted "["0.53-0.58"]" and "["eV"]" through the previous operation, we extracted the values as the numeric values 0.53 and 0.58 and the unit "eV." In addition, there are variations in the expressions of characteristic values and units in each paper, such as the use of the ± symbol to indicate conductivity. Therefore, we used regular expressions to write down patterns of values, multiplication symbols, and powers and normalized phrases matching the written patterns into values.

5   Results

5.1   Inter-Annotator Agreement

The inter-annotator agreement (IAA) was evaluated to assess the reliability of the corpus. The IAA is calculated based on the matching of the spans of labels between two annotators who have Master's degrees in materials science.

Type	String pattern	Normalization results with unit
Temp.	room temperature or RT	22 °C
Temp.	500 K	227 °C
Time	overnight	8 hours
Time	half an hour or half a day	0.5 hours or 12 hours
Time	two hours or 2 h	2 hours
Cond.	1.66×10^{-4} S/cm	0.000166 S/cm
Cond.	4.2 mS/cm	0.0042 S/cm
Cond.	$4.28 \pm 0.41 \times 10^{-2}$	3.87 S/cm, 4.69 S/cm
AE	0.93-1.04	0.93 eV, 1.04 eV
AE	2.00(5) eV	1.95 eV, 2.05 eV
AE	$0.44 <$ Ea(eV) < 0.46	0.44 eV, 0.46 eV

Table 1: Example of normalization results for temperature (Temp.), time, conductivity (Cond.), and activation energy (AE). The type indicates the type of phrase, the string pattern indicates the text extracted by the NER extractor, and the normalization result indicates the value normalized to numerical data using regular expressions.

The agreement score was calculated by considering the labels identified by one annotator as the gold label and those identified by the other annotator as the prediction. To evaluate the extraction performance, we performed a binary evaluation that classified all entities into either positive or negative. The precision was defined as the fraction of entities predicted as positive that are in fact positive, and recall is defined as the fraction of positive entities that are correctly predicted as positive. More precisely, for true-positive (TP), false-positive (FP), and false-negative (FN) entities, based on the entities extracted by the model, we define precision = TP/(TP + FP), recall = TP/(TP + FN), and F1-score = 2 × precision × recall/(precision + recall). These validations were used to evaluate the machine extraction performance when worker A's labeling was considered the correct answer. To verify that the definitions of the types extracted are consistent among the annotators, we used the recall as an evaluation metric. Table 6a and Table 6b show the calculated IAA scores and confusion matrix of each label, respectively, by using 60 paragraphs from 10 papers in our corpus. The result showed that the overall recall of IAA was 0.736, indicating good agreement between the two annotators. The confusion matrix showed that there were no type errors; however, there were many discrepancies owing to misses.

5.2 Comparing Language Models

Our NER Tagger trained on the corpus in this study was compared with four different language models: ELMo for materials synthesis (MatELMo) [10][6], BERT [4][7],

---

[6] https://github.com/olivettigroup/materials-synthesis-generative-models
[7] https://huggingface.co/bert-base-cased

Label	Recall
MAT-START	0.720
MAT-SOLVENT	0.529
MAT-FINAL	0.697
OPE	0.709
PROP-EQUIP	0.537
PROP-MAKER	0.870
PROP-METHOD	0.569
PROP-TEMP	0.722
PROP-TIME	0.792
CHARA-NAME	0.867
CHARA-ACT	0.846
CHARA-COND	0.975
ALL	0.736

(a) Recall of each label. ALL is the overall macro-recall score.
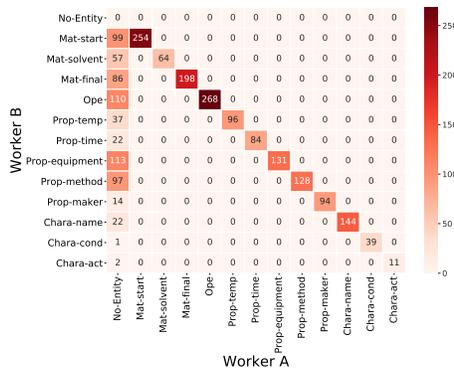
(b) Confusion matrix of IAA.

Fig. 6: IAA results.

SciBERT [5][8], and PubmedBERT [11][9] for token embedding. For SciBERT and BERT, we used transformers [12] to obtain embeddings and to connect to the NER Tagger extractor created in Flair [1], a framework for using state-of-the-art NLP models. For evaluations, the dataset was divided in a ratio of 6:2:2 for training, development, and testing, respectively. Table 2 shows the obtained results. The Mat-ELMo language model had the highest overall micro-F1 score of 0.778. SciBERT and BERT had higher F1-scores for the extraction of conductivity and activation energy, respectively. Although the present study aims to extract material names and property values, Mat-ELMo was selected for further analysis considering raw materials, temperature, and time conditions as it had the highest overall micro-F1 score.

5.3 Tuning Hyperparameters

We performed hyperparameter tuning for the NER model employing Mat-ELMo, which showed the best extraction performance as described in Section 5.2. We used optuna [14], a sophisticated optimization tool, for exploring the parameter space. Table 7b summarizes the evaluated hyperparameter space and the best parameters used for the final evaluation on the test dataset. After obtaining the optimal hyperparameter values, the model was trained again to evaluate its final performance. From the results shown in Table 7a, after tuning the hyperparameters, the micro-F1 score was improved from 77.8% to 78.1%.

---

[8] https://huggingface.co/allenai/scibert_scivocab_cased

[9] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

Model	MatELMo	PubmedBERT	SciBERT	BERT
MAT-FINAL	**0.613**	0.572	0.595	0.543
MAT-SOLVENT	**0.757**	**0.757**	0.724	0.705
MAT-START	**0.754**	0.726	0.688	0.651
OPE	0.825	**0.835**	0.832	0.826
PROP-EQUIP	**0.819**	0.801	0.798	0.795
PROP-MAKER	**0.869**	0.827	0.794	0.816
PROP-METHOD	0.798	**0.802**	0.785	0.793
PROP-TEMP	0.851	0.855	**0.875**	0.867
PROP-TIME	0.867	**0.892**	0.883	0.887
CHARA-NAME	**0.918**	0.912	0.914	0.914
CHARA-ACT	0.593	0.571	**0.654**	0.509
CHARA-COND	0.605	0.630	0.649	**0.685**
ALL	**0.778**	0.767	0.761	0.749

Table 2: F1 scores of sequence-labeling models with different base representations on development dataset. Micro-F1 scores were calculated using all labels (ALL). The highest value is indicated in bold.

5.4   Evaluation of Extracted NE Result

To verify the extraction performance of the tuned NER model, 100 paragraphs from 10 papers were labeled by NER. These paragraphs cover all sections that can be extracted from a paper. The labels were checked by the same annotator who labeled the corpus to correct any extraction errors or omissions. These validations were performed to evaluate the machine extraction performance when the machine's labeling was considered the correct answer, and the recall metric was used to evaluate the extraction performance of correctly extracted entities among the entities labeled by the NER model. Table 8a shows the agreement between our NER model and the human annotators for each label, and Table 8b shows the confusion matrix.

6   Research Trends Analysis

In this section, we analyze the NLP outputs to understand the trend of materials by year. Table 9a summarizes several key statistics of the NLP outputs, such as the number of papers, entities, and distribution of converted values. Further, Table 9b shows a country-by-country tabulation. Only abstracts extracted from papers were used for this tabulation, and the first author's country was counted.

   Next, we aggregated the extracted final materials by year for a quick analysis of the trends by year, as shown in Table 3. Only the paragraph section name "Abstract" was used in the extracted papers to prevent double-counting of papers. This result shows that the trend of frequently used materials differs by

Label	F1-score
Mat-Final	0.625
Mat-Solvent	0.771
Mat-Start	0.771
Ope	0.827
Prop-Equip	0.827
Prop-Maker	0.870
Prop-Method	0.813
Prop-Temp	0.857
Prop-Time	0.869
Chara-Name	0.917
Chara-Act	0.593
Chara-Cond	0.637
ALL	0.781

(a) Final result. ALL is the micro-F1 score.

Parameter	Range	Best
Learning rate	[0.05, 0.3]	0.15
Dropout	[0.3, 0.6]	0.3
Locked dropout	[0.3, 0.6]	0.4
Word dropout	[0.05, 0.15]	0.1
Hidden size	[32, 256]	256
RNN layers	[1, 3]	2
Weight decay	[0.0001, 0.0005]	0.0005

(b) Hyperparameters

Fig. 7: Micro-F1 score after hyperparameter tuning.

Label	Recall
Mat-Final	0.873
Mat-Solvent	0.956
Mat-Start	0.751
Ope	0.997
Prop-Equip	0.990
Prop-Maker	1.000
Prop-Method	0.987
Prop-Temp	0.924
Prop-Time	0.964
Chara-Name	0.994
Chara-Act	1.000
Chara-Cond	0.926
ALL	0.920

(a) Recall of each label. ALL is the macro recall score.
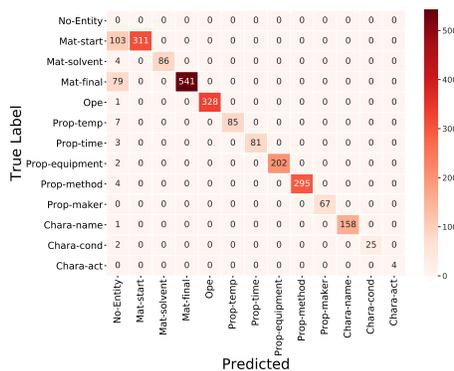
(b) Post-evaluation results

Fig. 8: Post-evaluation

Item	Count
Papers	12,895
Paragraphs	57,783
MAT-FINAL	919,645
MAT-SOLVENT	63,437
MAT-START	406,387
OPE	277,418
PROP-EQUIP	175,299
PROP-MAKER	55,018
PROP-METHOD	454,945
PROP-TEMP	75,004
PROP-TIME	77,889
CHARA-NAME	67,530
CHARA-ACT	14,596
CHARA-COND	31,005

Country	Count
China	7593
United States	947
Korea	884
Japan	384
India	365
UK	342
Germany	305
Australia	289
Spain	207
Singapore	204
Taiwan	150
France	133
Canada	113
Sweden	99

(a) Extracted data statistics.   (b) Top 14 countries.

Fig. 9: Base statistics

year. For example, "TiO_2" is ranked fourth in 2016–2017; however, it is ranked first in 2018–2019.

For the following analysis, we manually selected the following final materials from Table 3 for a query search using the word2vec model [37] to efficiently screen the many other materials obtained from extracted outputs: "$CH_3NH_3PbI_3$," "PEDOT:PSS," "TiO_2," "graphene," "ZnO," "MoS_2," "MOF," and "CNT." This model was trained using the same 12,895 papers that were input to the NLP pipeline. We then applied the trained word2vec model to the extracted final material and adopted the query with the highest cosine similarity as the type. The final material classified by type is used for trend analysis by country, as shown in Figure 10. This figure shows the features of developed materials by country and the change in the number of reported materials by year. Consider the comparison of China Figure 10a and United States Figure 10b: in China, the number of papers on "MoS_2" has been increasing in recent years, whereas in the United States, the number of papers has been decreasing. Further, the most frequently reported material in China in 2015 was "TiO_2," whereas that in 2019 was "MoS_2," indicating the shifting trend in materials science research.

We also visualized the condition-by-year for temperature and time, as shown in Figure 11. If multiple properties are extracted from a single paragraph, we select the property with the highest characteristic value. These results show that the trends of temperature and time when synthesizing "PEDOT:PSS" and "TiO_2" vary by year. In particular, the processing temperature of "PEDOT:PSS" shifted below 200°C. Further, the processing times of "PEDOT:PSS" and "TiO_2" differed in 2015, and the processing temperatures were similar in 2019. This indicates that there are similarities in the synthesis methods of

2016 – 2017	2018 – 2019
reduced graphene oxide(21)	TiO_2(16)
$CH_3NH_3PbI_3$(20)	reduced graphene oxide(15)
graphene(18)	graphene(13)
TiO_2(17)	$CH_3NH_3PbI_3$(10)
carbon(17)	SnO_2(10)
graphene oxide(10)	carbon(9)
ZnO(9)	$MAPbI_3$(9)
PEDOT:PSS(8)	MoS_2(8)
activated carbon(7)	covalent organic frameworks(8)
$BiVO_4$(6)	MOFs(8)

Table 3: Reported material final aggregation by year. Number in the brackets means number of extracted phrases.

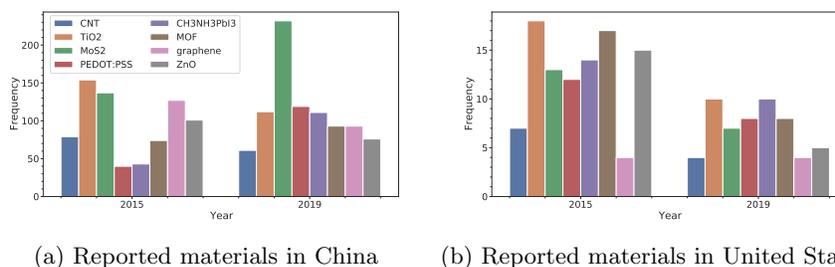(a) Reported materials in China   (b) Reported materials in United States

Fig. 10: The year of transition of material by country and by year.

"PEDOT:PSS" and "TiO_2." We have received comments from one material researcher that the results reported in this study are useful when investigating competitors and when designing material synthesis processes outside the laboratory.

7 Conclusion

This study proposed an NLP-based approach to analyze the trend in materials research for developments in materials science. We developed an NLP system with the BiLSTM-CRF model that was trained using manually labeled literature for extracting material properties from scientific literature. We conducted experiments to verify the effectiveness of the proposed NER method in the field of materials science.

The present study has two limitations: (1) linking multiple materials and property values when they are written in a single document, and (2) extracting characteristic values from nontextual components such as charts, diagrams, and

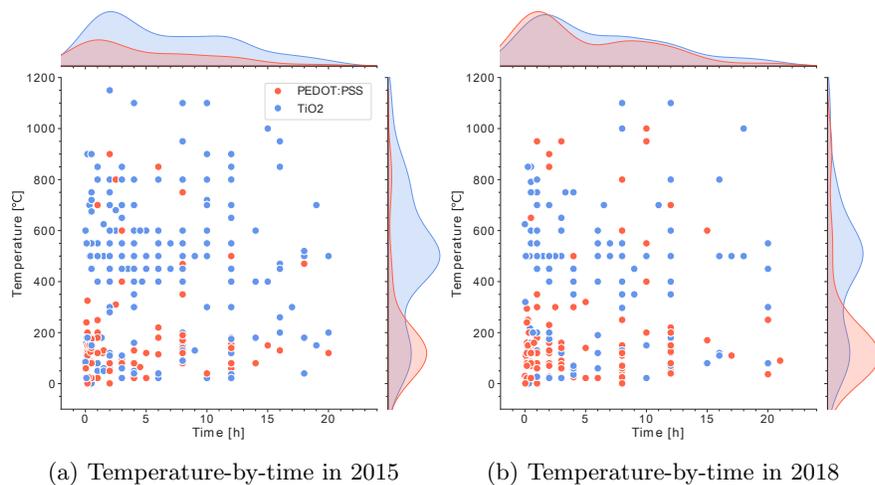(a) Temperature-by-time in 2015   (b) Temperature-by-time in 2018

Fig. 11: Condition-by-year.

tables that provide key information in many scientific documents. We aim to overcome these limitations through our ongoing studies.

In future work, we will aim to predict characteristic values such as conductivity and materials research trends from previous scientific literature to achieve computational materials synthesis.

References

1. A. Akbik and T. Bergmann and Duncan Blythe and K. Rasul and Stefan Schweter and Roland Vollgraf: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-HLT) 2019.
2. F. Kuniyoshi and K. Makino and J. Ozawa and M. Miwa: Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature. In: International Conference on Language Resources and Evaluation(LREC), 2020.
3. V. Tshitoyan and J. Dagdelen and L. Weston and A. Dunn and Ziqin Rong and O. Kononova and K. Persson and G. Ceder and A. Jain: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature **571**, 95–98, 2019.
4. J. Devlin and M. Chang and K. Lee and K. Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-HLT), 2019.
5. I. Beltagy and K. Lo and A. Cohan: SciBERT: A Pretrained Language Model for Scientific Text In: International Joint Conference on Natural Language Processing(IJCNLP), 2019.
6. S. Huang and J. Cole: A database of battery materials auto-generated using ChemDataExtractor. Scientific Data **7**, 260, 2020.

7. K. Chaitanya and X. Wei and R. Alan and M. Raghu: An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-HLT), 2018.
8. P. Stenetorp and S. Pyysalo and G. Topic and T. Ohta and S. Ananiadou and J. Tsujii: brat: a Web-based Tool for NLP-Assisted Text Annotation. In: European Chapter of the Association for Computational Linguistics(EACL), 2012.
9. Q. Yuan and P. Liu and G. Baker: Sulfonated polyimide and PVDF based blend proton exchange membranes for fuel cell applications. Journal of Materials Chemistry **3**, 3847–3853, 2015.
10. E. Kim and Z. and A. Grootel and K. Huang and M. Staib and S. Mysore and H. Chang and E. Strubell and A. McCallum and S. Jegelka and E. Olivetti: Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. Journal of chemical information and modeling **60**, 3, 1194–1201, 2020.
11. Y. Gu and R. Tinn and H. Cheng and M. Lucas and N. Usuyama and X. Liu and T. Naumann and J. Gao and H. Poon: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. In: ArXiv, abs/2007.15779, 2020.
12. T. Wolf and L. Debut and V. Sanh and J. Chaumond and C. Delangue and A. Moi and P. Cistac and T. Rault and R. Louf and M. Funtowicz and J. Brew: Transformers: State-of-the-Art Natural Language Processing. In: Empirical Methods in Natural Language Processing(EMNLP), 2020.
13. M. Krallinger and O. Rabal and F. Leitner and M. Vazquez and D. Salgado and others: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics **7**, S2–S2, 2015.
14. A. Takuya and S. Shotaro and Y. Toshihiko and O. Takeru and K. Masanori: Optuna: A Next-generation Hyperparameter Optimization Framework. In: International Conference on Knowledge Discovery and Data Mining(SIGKDD), 2019.
15. C. J. Court and J. Cole: Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. npj Computational Materials **6**, 1–9, 2020.
16. O. Kononova and T. He and H. Huo and A. Trewartha and E. Olivetti and G. Ceder: Opportunities and challenges of text mining in aterials research. iScience **24**, 3, 2021.
17. A. Akbik and D. Blythe and R. Vollgraf: Contextual String Embeddings for Sequence Labeling. In: International Conference on Computational Linguistics(COLING), 2018.
18. Z. Huang and W. Xu and K. Yu: Bidirectional LSTM-CRF Models for Sequence Tagging. In: ArXiv, abs/1508.01991, 2015.
19. M. Bada and M. Eckert and D. Evans and K. Garcia and K. Shipley and D. Sitnikov and W. Baumgartner and K. Cohen and K. M. Verspoor and J. Blake and L. Hunter: Concept annotation in the CRAFT corpus. BMC Bioinformatics **13**, 161–161, 2011.
20. J. D. Kim and T. Ohta and Y. Tateisi and J. Tsujii: GENIA corpus – a semantically annotated corpus for bio-textmining. BMC Bioinformatics **19 Suppl 1**, i180–2, 2003.
21. V. Tshitoyan and J. Dagdelen and L. Weston and A. Dunn and Z. Rong and O. Kononova and K. Persson and G. Ceder and A. Jain: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature **571**, 95–98, 2019.
22. Z. Jensen and E. Kim and S. H. Kwon and T. Z. H. Gani and Y. Román-Leshkov and M. Moliner and A. Corma and E. Olivetti: A Machine Learning Approach to

Zeolite Synthesis Enabled by Automatic Literature Data Extraction. ACS Central Science **5**, 892–899, 2019.

23. E. Kim and K. Huang and A. Saunders and A. McCallum and G. Ceder and E. Olivetti: Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. Chemistry of Materials **29**, 9436–9444, 2017.

24. E. Kim and K. Huang and A. Tomala and S. Matthews and E. Strubell and A. Saunders and A. McCallum and E. Olivetti: Machine–learned and codified synthesis parameters of oxide materials. Scientific Data **4**, 170127, 2017.

25. S. R. Young and A. Maksov and M. Ziatdinov and Y. Cao and M. J. Burch and J. Balachandran and L. Li and S. Somnath and R. M. Patton and S. V. Kalinin and R. Vasudevan: Data Mining for better material synthesis: the case of pulsed laser deposition of complex oxides. Journal of Applied Physics **123**, 115303, 2018.

26. K. Yamaguchi and R. Asahi and Y. Sasaki: SC-CoMIcs: A Superconductivity Corpus for Materials Informatics. In: International Conference on Language Resources and Evaluation(LREC), 2020.

27. Y. K. Jeong and Q. Xie and E. Yan and M. Song: Examining drug and side effect relation using author-entity pair bipartite networks. Journal of Informetrics **14**, 100999, 2020.

28. L. Hansson and R. B. Hansen and S. Pletscher-Frankild and R. Berzins and D. Hansen and D. Madsen and S. B. Christensen and M. R. Christiansen and U. Boulund and X. A. Wolf and S. Kim Kjærulff and M. van de Bunt and S. Tulin and T. S. Jensen and R. Wernersson and J. Jensen: Semantic text mining in early drug discovery for type 2 diabetes. PLoS ONE **15**(6), 2020.

29. D. Rebholz-Schuhmann and A. Oellrich and R. Hoehndorf: Text-mining solutions for biomedical research: enabling integrative biology. Nature Reviews Genetics **13**, 829–839, 2012.

30. O. Kononova and H. Huo and T. He and Z. Rong and T. Botari and W. Sun and V. Tshitoyan and G. Ceder: Text-mined dataset of inorganic materials synthesis recipes. Scientific Data **6**, 203, 2019.

31. R. Mahbub and K. Huang and Z. Jensen and Z. D. Hood and J. Rupp and E. Olivetti: Text mining for processing conditions of solid-state battery electrolyte. Electrochemistry Communications **121**, 106860, 2020.

32. L. Weber and M. Sänger and J. Munchmeyer and M. Habibi and U. Leser: Hun-Flair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. BMC Bioinfomatics, 2021.

33. H. Huo and Z. Rong and O. Kononova and W. Sun and T. Botari and T. He and V. Tshitoyan and G. Ceder: Semi-supervised machine-learning classification of materials synthesis procedures. npj Computational Materials **5**, 1–7, 2019.

34. M. Miwa and P. Thompson and S. Ananiadou: Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. BMC Bioinformatics **28**, 1759–1765, 2012.

35. J. Björne and T. Salakoski: Biomedical Event Extraction Using Convolutional Neural Networks and Dependency Parsing. In: Workshop on Biomedical Natural Language Processing (BioNLP), 2018.

36. S. Mysore and Z. Jensen and E. Kim and K. Huang and H. Chang and E. Strubell and J. Flanigan and A. McCallum and E. Olivetti: The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. In: Linguistic Annotation Workshop (LAW), 2019.

37. T. Mikolov and I. Sutskever and K. Chen and G. S. Corrado and J. Dean: Distributed Representations of Words and Phrases and their Compositionality. In: Neural Information Processing Systems (NeurIPS), 2013.