PIEKM: ML-based Procedural Information Extraction and Knowledge Management System for Materials Science Literature

Huichen Yang
Kansas State University
Manhattan, Kansas, USA

Carlos Aguirre
Johns Hopkins University
Baltimore, Maryland, USA

William Hsu
Kansas State University
Manhattan, Kansas, USA

Abstract

The published materials science literature contains abundant description information about synthesis procedures that can help discover new material areas, deepen the study of materials synthesis, and accelerate its automated planning. Nevertheless, this information is expressed in unstructured text, and manually processing and assimilating useful information is expensive and time-consuming for researchers. To address this challenge, we develop a Machine Learning-based procedural information extraction and knowledge management system (PIEKM) that extracts procedural information (*recipe steps*), figures, and tables from materials science articles, and provides information retrieval capability and the statistics visualization functionality. Our system aims to help researchers to gain insights and quickly understand the connections among massive data. Moreover, we demonstrate that the machine learning-based system performs well in low-resource scenarios (i.e., limited annotated data) for domain adaption.

1 Introduction

The procedural information in materials science literature aims to help researchers reproduce experiments and gain insights to speed up the process of new materials synthesis development (Vaucher et al., 2020; Kononova et al., 2019). It takes the form of *recipes* (e.g., Figure 4) and is normally defined as a series of actions and their corresponding conditions and results. Such information contains imperatives, action verbs, steps of operations, and constructions (Yang et al., 2019). This information can be commonly found in method sections of materials science research literature. However, a great amount of scientific literature is published every year by the growing materials science research community. These well-established works provide a foundation to enlighten researchers and explore new materials development simultaneously. Acquiring valuable information from the year-over-year increasing scientific literature efficiently and effectively remains one of the great challenges (Kononova et al., 2021). The existing scholarly literature search engines, such as Google Scholar and Semantic Scholar, provide a good service to discover the relevant publications, but they cannot directly deliver the recipe steps of the experiments that are included in the literature. Therefore, an intelligent system that provides the functions of procedural information searching and viewing, visualization, and analysis is highly demanded.

Information Extraction (IE) which is a sub-area of Natural Language Processing (NLP) provides an efficient way to automatically extract structured information from large unstructured text data. Likewise, in the materials science domain, IE has been applied to similar tasks, such as experimental steps classification with unsupervised approaches of probabilistic methods for inorganic materials (Huo et al., 2019) and named entity recognition in materials science domain (Kim et al., 2017; Yang and Hsu, 2021). One of the biggest challenges of extracting information in materials science articles is that the annotated datasets are insufficient (Olivetti et al., 2020), which can be overcome by machine learning, particularly transfer learning, with pre-trained models obtained from other large training datasets (Zhang et al., 2021). Transfer learning can help with domain adaptation in materials science. We use transfer learning through fine-tuning a pre-trained language model with datasets in the materials science domain for chemical entity extraction. The trained model is integrated into the PIEKM system and performs well. This solves real cases where the number of training data in the materials science domain is very small for information extraction.

This paper presents PIEKM, a prototype of a machine learning-based procedural information extraction and knowledge management system based

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations, pages 57–62
November 20–23, 2022. ©2022 Association for Computational Linguistics

on materials scientific literature. The goal of PIEKM is to demonstrate procedural information extraction and information retrieval capabilities. Three crucial contributions of PIEKM are summarized as follows:

- This system helps researchers to obtain materials science-related procedural information efficiently and effectively from massive publications.
- The system utilizes transfer learning approaches, such as chemical entity extraction, which can solve the issues with the small size of training dataset.
- The system is flexible and can be easily deployed in other domains.

2 System Architecture

In this section, we describe the detailed architecture of PIEKM system. The proposed system consists of three modules: (A) Information Processing, (B) User Interface, and (C) Query Processing and Information Storage. Figure 1 shows the architecture of PIEKM system. Figure 2 shows the home page of PIEKM system. The details of each module are introduced as follows.

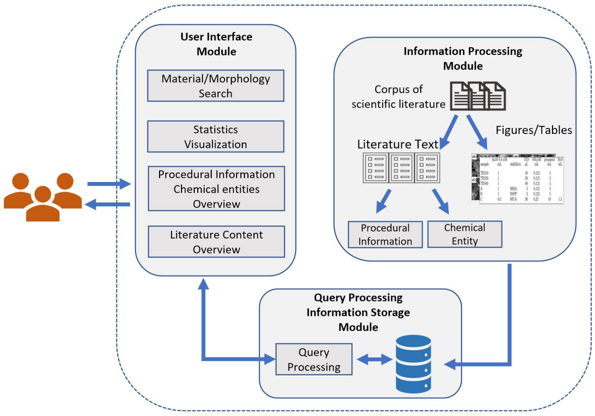
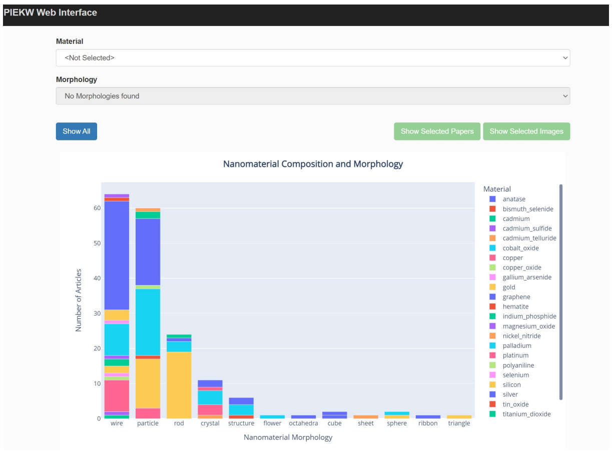

User Interface Module
Material/Morphology Search
Statistics Visualization
Procedural Information Chemical entities Overview
Literature Content Overview
Information Processing Module
Corpus of scientific literature
Literature Text
Figures/Tables
Procedural Information
Chemical Entity
Query Processing Information Storage Module
Query Processing

Figure 1: Architecture of PIEKM system

(A) Information Processing Module: This module processes the information from the digital scientific literature. We focus on Portable Document Format (PDF) digital scientific literature in the PIEKM system. The input corpus of digital scientific literature has been segmented into text and non-text (figures, tables) parts (section 3.1). Then the procedural information (section 3.2) and name entities (section 3.3) are extracted from these texts.


PIEKW Web Interface

Material
<Not Selected>

Morphology
No Morphologies found

Figure 2: Home page of PIEKM system

The extracted figures and tables are stored in the corresponding folders. The rest of extracted text information is stored in as a semi-structured format in the database for quick query response.

(B) User Interface Module: This module is in charge of responding to user queries and showing the result corresponding to each query, providing the preview of figures and tables of articles available in the system, and presenting the details of every single article which includes procedural information and chemical entities.

(C) Query Processing and Information Storage Module: This module is responsible for query processing and information storage. The queries are sent by users, then the answers would be acquired from the information storage database and returned back to the user interface module for display. The module supports different material compositions and morphology searches.

The PIEKM system is deployed by Flask[1] framework and written in Python. We use MongoDB[2] to store the information data and respond to queries. The Plotly[3] and Dash[4] are used for interactive visualizations.

3 Information processing

In this section, we present the implementation details of the information processing module in PIEKM system. This module serves as a pipeline including free-text extraction, procedural information extraction, and chemical entity recognition. The details of each stage are described as follows.

[1] https://flask.palletsprojects.com/
[2] https://www.mongodb.com/
[3] https://plotly.com/javascript/
[4] https://plotly.com/dash/

3.1 Free-Text Extraction

The first step is to extract the text from the digital scientific literature, such as Portable Document Format (PDF) files, which is not in readable format and cannot be processed by computer directly, before processing it further. However, the very diverse formats and structures of scientific literature corpora could make text extraction and section classification difficult. The existing tools, such as PDFMiner (Shinyama, 2007) and PDFReader (Polshcha, 2020), may not fully extract the sections (e.g., methodology, experiment, results), and leave them in the wrong order. This could significantly affect recipe extraction if the recipe steps are not in sequential order. We, therefore, make use of heuristic rule-based Metadata-Analytic Text and Section Extractor (MATESC) (Maria et al., 2018) system to solve the issues of different formats. MATESC is a heuristic rule-based pattern analysis tool that is used for extracting text and classifying sections from the scientific literature. The key purpose of this tool is to accelerate the extraction of information and semantic knowledge among variant formats of scientific literature across different domains. We can extract text spans and utilize metadata features (i.e., spatial layout location, font type, and size) via MATESC. By doing so, we are able to create grouped blocks of text which can be then classified into groups and subgroups depending on characterized paper sections.

MATESC extracts text including the metadata features of all characters (i.e., font type and size, spatial layout location) from PDF scientific articles which are considered as input. It is worth noting that the irrelevant text left in the margins of every page of these documents can be automatically removed from the extracted text based on the corresponding spatial layout location. Following that, words will be placed into the appropriate line, and then the different fonts and locations of characters will be considered to differentiate between section titles and section content. Lastly, the lines created will be merged into paragraphs which will then be ordered sequentially by the computation of the bounding box of paragraphs.

MATASC has been evaluated with 300 scientific articles, including 150 articles that are related to the materials science domain, and the others are randomly selected from online resources. All sections of these 300 articles are extracted as ground truth for MATASC performance evaluation. We choose GROBID (Lopez, 2009) which is a prevailing tool for metadata extraction from scholarly articles. The Longest Common Subsequence (LCS) that compares the longest common subsequence between ground truth and automatic extracts serves as the evaluation metric. Table 1 reports the performance evaluation results.

Table 1: Evaluation results comparison between MATESC and GROBID

Article	Name	Accuracy	F1
Random	MATESC	**0.85**	**0.57**
Random	GROBID	0.82	0.44
Relevant	MATESC	**0.88**	**0.72**
Relevant	GROBID	0.76	0.40

3.2 Procedural Information Extraction

The procedural information takes the form of the recipe in our PIEKM system. It describes the main synthesis steps of experiments in materials science literature. We use two approaches to ensure the quality of extracted procedural information: relevant synthesis sentence classification and checking if a relevant sentence contains recipe entities.

Relevant sentences classification: We applied the binary Naïve Bayes (NB) classifier to relevant sentence classification in the experiment section which was output by the free-text extraction. The sentences can be considered relevant if they contain the *recipe* elements (e.g., named compounds, chemical entities, unit operations or sub-procedures). We annotated 2600+ sentences from 98 relevant literature for training the classification model. Particularly, two domain experts annotated 120 sentences from 5 relevant literature and the rest of the annotation work was done by three trained annotators. To better predict the class attribute of the input sentence, we train the NB classifier with word term frequency as count features to achieve a leaned function with 80% accuracy of prediction.

Relevant sentences entities checking: We use ChemicalTagger (Hawizy et al., 2011), an open-source tool for semantic text-mining in the chemistry domain, for recipe sentence checking. The ChemicalTagger uses regex expression to tag different entities, such as conditions, molecules, actions, and phrases, from sentences. In our PIEKM system, the procedural information or recipe should include at least one action which can be represented by a verb word (e.g., dry, distill, dissolve). The system will ignore the sentence even if it has been classified as a relevant sentence.

3.3 Chemical Entity Extraction

Chemical entity extraction is considered as named entity recognition (NER) which is used to recognize and classify the concepts in texts to identify the objects of semantic value. We are able to broadly pre-define the name entities, such as material names, material properties, and sample deceptions, in materials science contexts, based on the task requirements (Mysore et al., 2019). Large, annotated corpora are required to be able to train a machine learning model for NER tasks, which brings a challenge in the materials science domain due to the insufficient annotated dataset. Additionally, manually labeling based on an enormous number of articles would be very time-consuming and expensive for domain exports. We address this issue with transfer learning that is based on the combination of attention-based pre-trained language model SciBERT (Beltagy et al., 2019), Bidirectional Long Short-term Memory (BiLSTM) (Huang et al., 2015), and Conditional Random Fields (CRF), or SciBERT-BiLSTM-CRF for short. Specifically, the pre-trained SciBERT model serves as the embedding layer which takes raw sentences as input and outputs the contextual embedding vectors for each word to the BiLSTM layer. BiLSTM layer takes these inputs for syntactic and semantic feature representation learning and outputs the predicted scores of each label which then will be fed into the CRF layer. Finally, the CRF layer will select the label sequence with the highest predicts score as output.

Considering the insufficient annotated datasets that are available for chemical entity extraction, we merged two annotated corpora in the materials science domain to train the model. One of them is materials synthesis procedural text corpus (**MSP**) (Mysore et al., 2019), which has 230 experiment paragraphs regarding synthesis procedure in the materials science domain and 21 different pre-defined named entities. The other corpus is in the field of solid oxide fuel cell (**SOFC**) (Friedrich et al., 2020), including 45 open-access scholarly articles and 5 different pre-defined named entities. In addition, the BIO format is used to annotate both of the corpora mentioned above, where B indicates the word beginning entity, I represents the words inside the entity, and O is the outside of the entity.

Table 2: Evaluation results comparison

Model	Precision	Recall	F1
SciBERT-BiLSTM-CRF	**0.93**	**0.91**	**0.92**
ChemDataExtractor	0.88	0.83	0.85

We keep only the material name as the predefined named entity in the corpus to train the SciBERT-BiLSTM-CRF model since PIEKM system only focuses on the chemical entity extraction rather than the extraction of other named entities. We compared our model with ChemDataExtractor (Swain and Cole, 2016), a tool for the automated extraction of chemical information from the scientific literature. Table 2 shows the comparison of evaluation results between SciBERT-BiLSTM-BRF and ChemDataExtractor. Note that the ChemDataExtractor is not trained on this merged corpora but only for evaluation comparison.

4 Query processing and information storage

We use the model that is fine-tuned from section **3.3** to extract the key information from the title of the paper. The key information in our system could be considered into two types: **material** and **morphology**. For example, *copper* and *gold* are the type of material; *nanocube* and *nanowire* are the type of morphology. We integrate all this information and store it into the database as a query feature for users.

5 Demonstration

The demonstration covers all of the features of PIEKM system. Figure 2 shows the home page of PIEKM system. It visualizes the overview of the association between the number of articles within the database of nanomaterial composition and the corresponding morphology. The user can click the material name to see the number of relevant articles across different morphology. The relevant literature can be searched by material or morphology name, and the search result page will show a preview figure browser that offers all different options of material or morphology names to select and provides all figures included in the relevant articles (Figure 3). In addition, the chemical entities, recipe, and the full content of extracted literature can be displayed after clicking the title on the top of each figure on the browser (Figure 4).

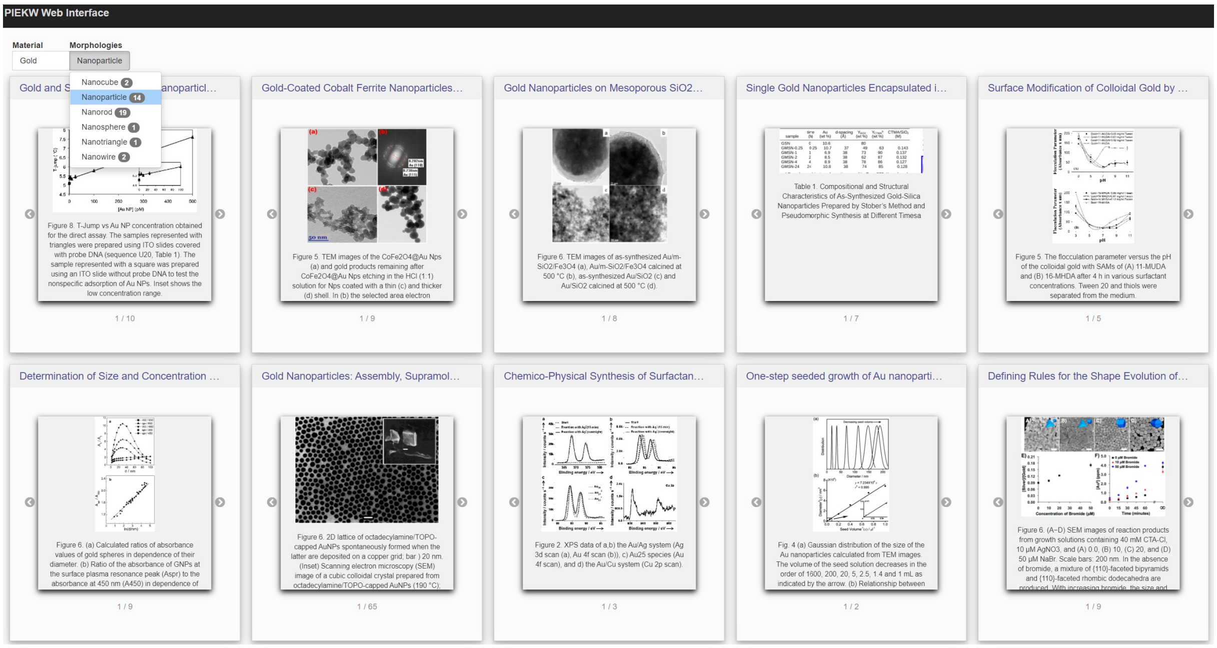
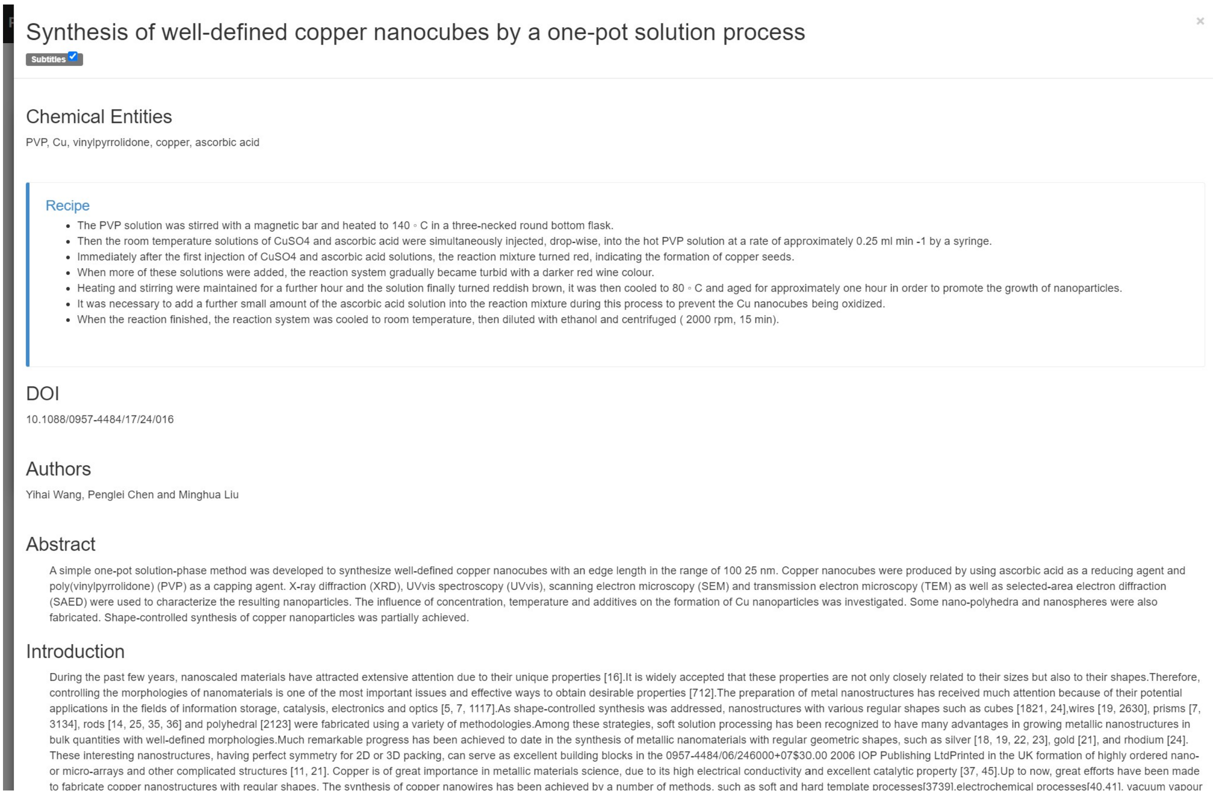

PIEKW Web Interface
Material
Gold
Morphologies
Nanoparticle
Nanocube 2
Nanoparticle 14
Nanorod 19
Nanosphere 1
Nanotriangle 1
Nanowire 2
Gold-Coated Cobalt Ferrite Nanoparticles…
Gold Nanoparticles on Mesoporous SiO2…
Single Gold Nanoparticles Encapsulated i…
Surface Modification of Colloidal Gold by …
Determination of Size and Concentration …
Gold Nanoparticles: Assembly, Supramol…
Chemico-Physical Synthesis of Surfactan…
One-step seeded growth of Au nanoparti…
Defining Rules for the Shape Evolution of…
1 / 10
1 / 9
1 / 8
1 / 7
1 / 5
1 / 9
1 / 65
1 / 3
1 / 2
1 / 9

Figure 3: Search result page showing extracted papers and a preview figure browser



Synthesis of well-defined copper nanocubes by a one-pot solution process

Subtitles ☑

Chemical Entities

PVP, Cu, vinylpyrrolidone, copper, ascorbic acid

Recipe

- The PVP solution was stirred with a magnetic bar and heated to 140 ◦ C in a three-necked round bottom flask.
- Then the room temperature solutions of CuSO4 and ascorbic acid were simultaneously injected, drop-wise, into the hot PVP solution at a rate of approximately 0.25 ml min -1 by a syringe.
- Immediately after the first injection of CuSO4 and ascorbic acid solutions, the reaction mixture turned red, indicating the formation of copper seeds.
- When more of these solutions were added, the reaction system gradually became turbid with a darker red wine colour.
- Heating and stirring were maintained for a further hour and the solution finally turned reddish brown, it was then cooled to 80 ◦ C and aged for approximately one hour in order to promote the growth of nanoparticles.
- It was necessary to add a further small amount of the ascorbic acid solution into the reaction mixture during this process to prevent the Cu nanocubes being oxidized.
- When the reaction finished, the reaction system was cooled to room temperature, then diluted with ethanol and centrifuged (2000 rpm, 15 min).

DOI

10.1088/0957-4484/17/24/016

Authors

Yihai Wang, Penglei Chen and Minghua Liu

Abstract

A simple one-pot solution-phase method was developed to synthesize well-defined copper nanocubes with an edge length in the range of 100 25 nm. Copper nanocubes were produced by using ascorbic acid as a reducing agent and poly(vinylpyrrolidone) (PVP) as a capping agent. X-ray diffraction (XRD), UVvis spectroscopy (UVvis), scanning electron microscopy (SEM) and transmission electron microscopy (TEM) as well as selected-area electron diffraction (SAED) were used to characterize the resulting nanoparticles. The influence of concentration, temperature and additives on the formation of Cu nanoparticles was investigated. Some nano-polyhedra and nanospheres were also fabricated. Shape-controlled synthesis of copper nanoparticles was partially achieved.

Introduction

During the past few years, nanoscaled materials have attracted extensive attention due to their unique properties [16].It is widely accepted that these properties are not only closely related to their sizes but also to their shapes.Therefore, controlling the morphologies of nanomaterials is one of the most important issues and effective ways to obtain desirable properties [712].The preparation of metal nanostructures has received much attention because of their potential applications in the fields of information storage, catalysis, electronics and optics [5, 7, 1117].As shape-controlled synthesis was addressed, nanostructures with various regular shapes such as cubes [1821, 24],wires [19, 2630], prisms [7, 3134], rods [14, 25, 35, 36] and polyhedral [2123] were fabricated using a variety of methodologies.Among these strategies, soft solution processing has been recognized to have many advantages in growing metallic nanostructures in bulk quantities with well-defined morphologies.Much remarkable progress has been achieved to date in the synthesis of metallic nanomaterials with regular geometric shapes, such as silver [18, 19, 22, 23], gold [21], and rhodium [24]. These interesting nanostructures, having perfect symmetry for 2D or 3D packing, can serve as excellent building blocks in the 0957-4484/06/246000+07$30.00 2006 IOP Publishing LtdPrinted in the UK formation of highly ordered nano- or micro-arrays and other complicated structures [11, 21]. Copper is of great importance in metallic materials science, due to its high electrical conductivity and excellent catalytic property [37, 45].Up to now, great efforts have been made to fabricate copper nanostructures with regular shapes. The synthesis of copper nanowires has been achieved by a number of methods, such as soft and hard template processes[3739],electrochemical processes[40,41], vacuum vapour

Figure 4: Chemical entities, recipe, and full content of extracted literature

6 Conclusion

This work presents a machine learning-based procedural information extraction and knowledge management system, namely PIEKM, for the materials science domain. PIEKM system integrates multiple functionalities, such as procedural information extraction, chemical entities extraction, information retrieval capabilities, and statistics interactive visualization, into a single web interface. This system provides an efficient way for researchers to gain insights from an enormous number of well-established literature and offers a feasible way to manage knowledge and publications in not only materials science but also other domains.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. 2011. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3(1):1–13.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Haoyan Huo, Ziqin Rong, Olga Kononova, Wenhao Sun, Tiago Botari, Tanjin He, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Computational Materials*, 5(1):1–7.

Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. 2017. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444.

Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A Olivetti, and Gerbrand Ceder. 2021. Opportunities and challenges of text mining in materials research. *Iscience*, 24(3):102155.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):1–11.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

F Maria, De La Torre, Carlos A Aguirre, BreAnn M Anshutz, and William H Hsu. 2018. Matesc: Metadata-analytic text extractor and section classifier for scientific publications. In *KDIR*, pages 259–265.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317.

Maksym Polshcha. 2020. Pdfreader. `https://github.com/maxpmaxp/pdfreader`.

Yusuke Shinyama. 2007. Pdfminer. `https://github.com/pdfminer/pdfminer.six`.

Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):1–11.

Huichen Yang, Carlos A Aguirre, F Maria, Derek Christensen, Luis Bobadilla, Emily Davich, Jordan Roth, Lei Luo, Yihong Theis, Alice Lam, et al. 2019. Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 41–46. IEEE.

Huichen Yang and William H Hsu. 2021. Named entity recognition from synthesis procedural text in materials science domain with attention-based approach. In *SDU@ AAAI*.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Online. Association for Computational Linguistics.