MATKB: SEMANTIC SEARCH FOR POLYCRYSTALLINE MATERIALS SYNTHESIS PROCEDURES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we present a novel approach to knowledge extraction and retrieval using Natural Language Processing (NLP) techniques for material science. Our goal is to automatically mine structured knowledge from millions of research articles in the field of polycrystalline materials and make it easily accessible to the broader community. The proposed method leverages NLP techniques such as entity recognition and document classification to extract relevant information and build an extensive knowledge base, from a collection of 9.5 Million publications. The resulting knowledge base is integrated into a search engine, which enables users to search for information about specific materials, properties, and experiments with greater precision than traditional search engines like Google. We hope our results can enable material scientists quickly locate desired experimental procedures, compare their differences, and even inspire them to design new experiments.

1 INTRODUCTION

Materials science is a rapidly growing and evolving field, with discoveries and innovations always being made. As the field grows, so does the amount of published research, making it increasingly challenging for researchers to keep up with the latest developments and find the information they need. This is especially true for researchers working in specialized areas, where the sheer volume of research can make it difficult to find relevant information. Therefore, there has been growing interest in applying machine learning for automatically extracting information from tons of publications Kim et al. (2017a); Olivetti et al. (2020).

Traditionally, researchers have relied on search engines like Google to find information. While these search engines are powerful and widely used, they can be limited in their ability to search within specific fields, such as materials science. Additionally, they often return many irrelevant results, making it time-consuming to sort through the results and find the information one needs.

To address these challenges, this paper presents a new approach to knowledge extraction and retrieval using NLP techniques. Our approach leverages the advances in NLP to automatically extract relevant information from research articles, such as materials, properties, and experiments, and build a large knowledge base. This knowledge base is then integrated into a search engine that allows users to search for information about specific materials and experiments with greater precision and speed than traditional search engines.

Recently, there has been a released corpus PcMSP Yang et al. (2022) for entities and relations extraction from polycrystalline materials synthesis procedure. We utilize their data to build our search engine as a first step. We leave the extension to the whole materials domain for future work.

In general, from a collection of 4.9M and 4.6M publications in physics and material science domain in S2ORC Lo et al. (2020), we retrieve 5,846 relevant articles. Based on this, we extract 269,808 desired entities for constructing our semantic search platform MatKB. Compared with the human expert-curated commercial application like Reaxsys provided by Elsevier, we will make our platform freely available to the public.

2 RELATED WORK

The application of Natural Language Processing (NLP) techniques in materials science has gained significant attention in recent years. The main objective of using NLP in materials science is to extract information from unstructured text sources such as scientific articles, patents, and technical reports. This information can be used for various purposes such as knowledge discovery, material design, and performance optimization.

One of the earliest studies on NLP for materials procedures extraction was performed by Kim et al. (2017a), who used NLP techniques to extract materials processing information from the literature. They proposed a system that used rule-based and machine learning-based methods to identify and extract materials processing information and make predictions based on it. Similar work has also been reported in Jensen et al. (2019); Kim et al. (2017b); He et al. (2020).

In conclusion, using NLP techniques for materials procedure extraction has shown promising results and has the potential to revolutionize the way information is extracted and utilized in materials science.

3 METHODS

We aim to build a publicly available knowledge base for the semantic search of experimental sections focused on Polycrystalline materials.

Corpus collection: Since most scientific publications can only be accessed on specific journals, their results can not be publicly distributed, thus not satisfying our needs. We turn to the largest open-access scientific publications, S2ORC Lo et al. (2020) dataset, for acquiring all available full-text articles, specifically focusing on the subdomains of materials science and physics. However, most articles only provide abstract parts, and we obtain 838k, and 213k full text, respectively. Finally, all paragraphs are parsed by the Chemdataextractor Swain & Cole (2016) specifically designed for the scientific domain.

Data Filtering: To obtain relevant information, we applied predefined key phrases (see Appendix 7.1) suggested by materials experts to filter all relevant paragraphs from the result in the previous step, which gives us $5,846$ articles with full text. To test the recall rate of our filtering mechanism, we also test this filtering process to the full article of the test set in PcMSP Yang et al. (2022), where we successfully retrieve 230 relevant paragraphs from 290 original examples, achieving a recall of 80%.

Named Entity Recognition: To extract semantic entities within the filtered paragraphs, we utilized the Named Entity Recognition (NER) model proposed by Zhong & Chen (2021). We follow the training setups in Yang et al. (2022) and obtain an overall F1 score of 79% using the MatBERT trained on 50 million materials science paragraphs by Walker et al. (2021).

Semantic Search: The extracted information was then loaded into our intelligent search engine powered by Elasticsearch [1], enabling fast and flexible search capabilities. We adopt the pipeline in SynKB from Bai et al. (2022) for interface design.

User Interface: Our interface allows researchers to search for specific information, such as temperature or pressure, by entering single or multiple keywords. The system returns all relevant paragraphs, enabling quick and easy access to the most important methods in previous research.

4 RESULT

4.1 STATISTICS

Table 1 shows the statistics of predicted entity mentions in a dataset. The entity mentions are divided into 11 categories: Descriptor, Material-target, Material-intermedium, Operation, Device, Brand, Property-time, Value, Property-pressure, Material-others, Material-recipe, and Property-temperature, following the original definition in PcMSP Yang et al. (2022). The categories are

[1] https://www.elastic.co/downloads/elasticsearch

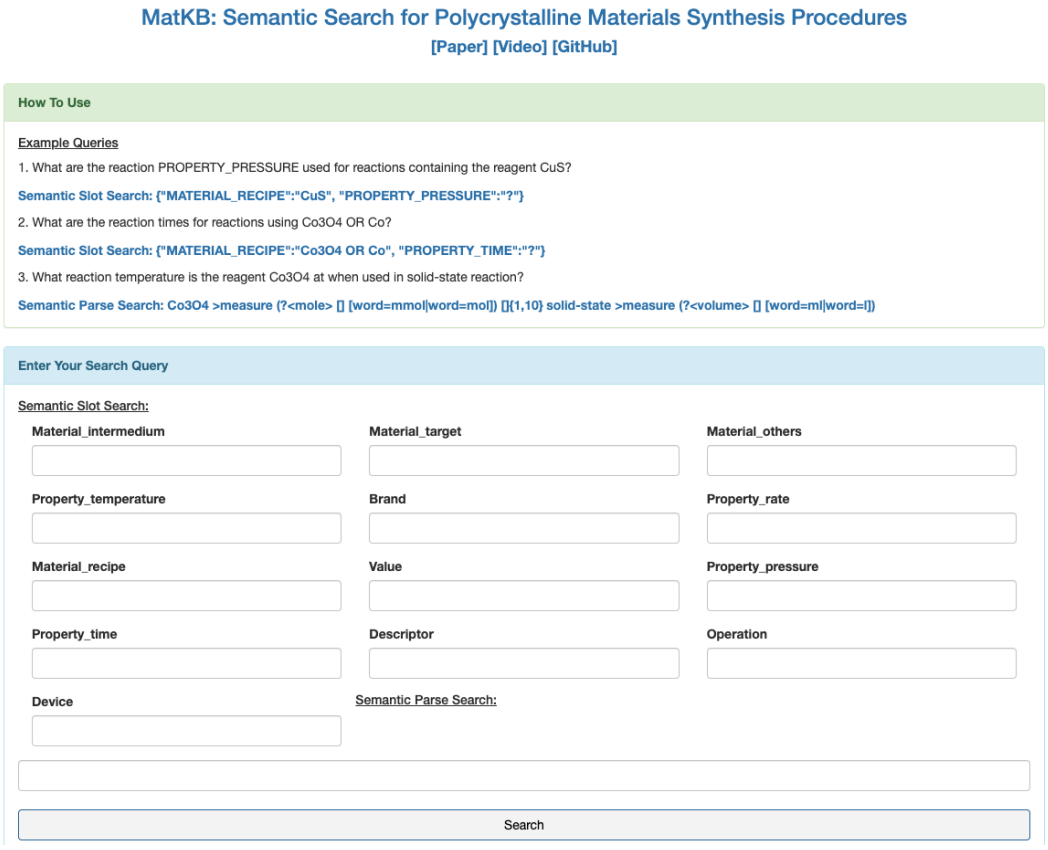

Figure 1: An overview of our MatKB semantic search interface. Different semantic slots can be combined or independently for search.

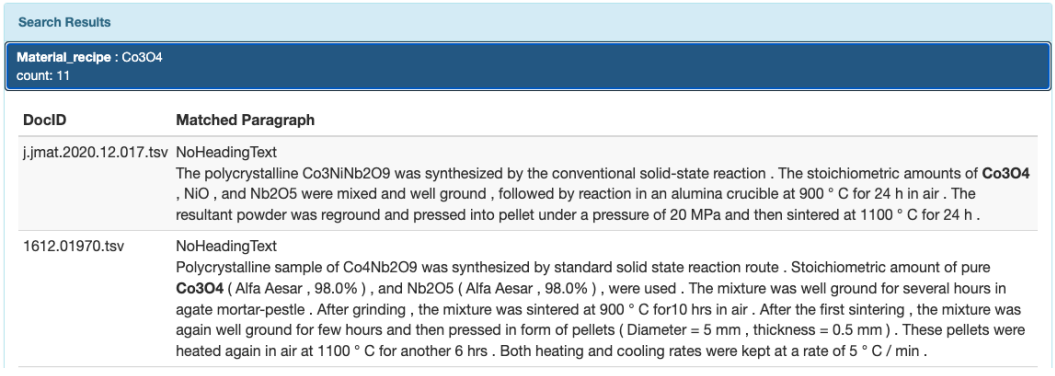

Figure 2: An example showing the search results by *Material_recipe: Co3O4*.

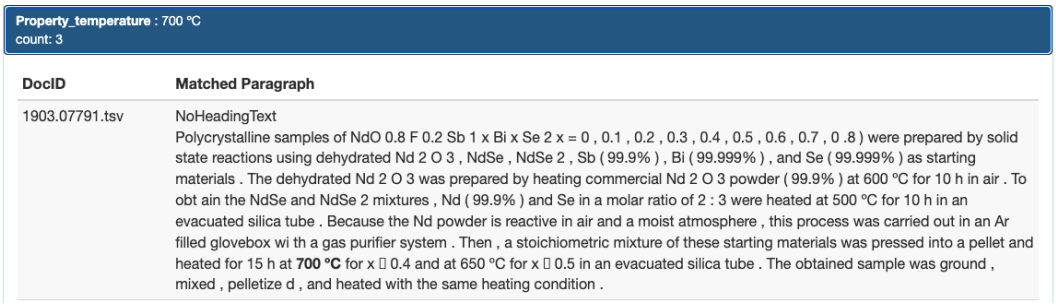

Figure 3: An example showing the search results by *Material_temperature:* 700 °C.

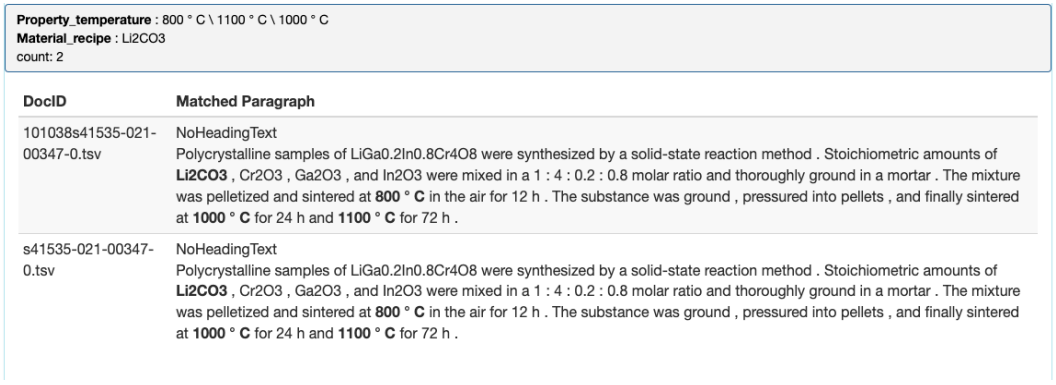

Figure 4: An example showing the search results by a combination of *Material_temperature:* 1000 °C and *Material_recipe:* $Li2Co3$.

defined based on the type of information they represent. For each category, Table 1 lists the number of counts (#Count), the number of unique mentions (#Unique), and a few examples of the mentions. The extracted dataset's total number of entity mentions is 269,808, with 29,774 unique mentions. The most frequently mentioned category is Descriptor, with 82,766 counts, followed by Operation, with 55,229 counts. The least frequently mentioned category is Property-rate, with only 2,133 counts. The information in this table provides insights into the distribution of entity mentions across the different categories, which can be useful for various data analysis and information extraction tasks.

Name	#Count	#Unique	Examples
Descriptor	82,766	7,721	polycrystalline, different, powder, single
Material-target	11,651	1,063	SiC, FeSe, ZnO, LaFeAsO
Material-intermedium	18,956	1,356	solution, grains, powders, pellets
Operation	55,229	3,993	added, arc, heat, grinding
Device	15,659	2,163	tube, furnace, ampoule, crucible
Brand	5,241	1,671	Sigma-Aldrich, Rigaku, Hitachi, Bruker
Property-time	5,103	794	24 h, 30 min, 3 h, 1.5 hours
Value	24,045	3,295	10 mg, stoichiometric amounts, 2 ml, around 3 g
Property-pressure	9,466	2,190	nitrogen, ambient pressure, air, 20 KPa
Material-others	8,294	1,338	ethanol, water, carbon, silicon
Material-recipe	18,341	1,218	Al, Si, Ga, Zn
Property-temperature	12,924	2,303	room temperature, 1000 °C, below 600 °C, about 100 °C
Property-rate	2,133	669	cooling rate, 1 K/min, approximately 2 K/min, air
Total	269,808	29,774	

Table 1: Predicted entity mention statistics and corresponding examples.

4.2 SEARCH

In Figure 1, we show an overview of our search interface, where we can perform a search according to our predefined semantic slots. For example, the results in Fig. 2 are obtained by a slot search of *Material_recipe:* $Co3O4$. Besides, we additionally show more search examples in Fig. 3 and 4. Compared with traditional search engines like Google or scholar search platform like Google scholar or Semantic Scholar, our pre-extracted entities can return us with precise experimental sections without further click-into publishers' websites and do tediously manual filtering. We hope such a tool can help materials scientists save time looking for correct references for experiments. Furthermore, since we return multiple results with different experimental procedures, material scientists can also compare the differences between those methods for designing their experiments.

5 CONCLUSION

In conclusion, we have presented a new approach for extracting structured knowledge from large amounts of research articles in materials science. Our method leverages NLP techniques to identify entities and experimental sections and builds an extensive knowledge base for easy search and retrieval. The proposed system demonstrates superiority over traditional search methods like Google by instantly returning experimental sections based on specific entity queries. Our results show that our approach can effectively extract valuable information and provide a comprehensive overview of current research in the field of materials science.

Future work will focus on expanding our knowledge base to cover a broader range of research articles and improving the accuracy of our entity recognition and experimental section extraction models. Additionally, we plan to enhance the user experience of the search website by incorporating interactive visualizations and more advanced search algorithms. We believe that this system has the potential to greatly improve the efficiency and effectiveness of research in the field of materials science and ultimately contribute to scientific advancements in this area.

6 LIMITATIONS

Data Bias: The study's results may be biased by the limited scope of the data used, and any biases present within the data.

Model Limitations: The results are only as reliable as the Named Entity Recognition (NER) model used, and any limitations or inaccuracies in the model could affect the results.

User Error: The accuracy of the results may be impacted by user error, such as incorrect keywords or misunderstandings of the system functionality.

It is important to carefully consider and address these potential risks in the study's design and implementation to ensure the results' validity and reliability.

REFERENCES

Fan Bai, Alan Ritter, Peter Madrid, Dayne Freitag, and John Niekrasz. Synkb: Semantic search for synthetic procedures. *arXiv preprint arXiv:2208.07400*, 2022.

Tanjin He, Wenhao Sun, Haoyan Huo, Olga Kononova, Ziqin Rong, Vahe Tshitoyan, Tiago Botari, and Gerbrand Ceder. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chemistry of Materials*, 32(18):7861–7873, 2020.

Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry ZH Gani, Yuriy Román-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS central science*, 5(5):892–899, 2019.

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

Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. Machine-learned and codified synthesis parameters of oxide materials. *Scientific data*, 4(1):1–9, 2017b.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447.

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

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

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

Xianjun Yang, Ya Zhuo, Julia Zuo, Xinlu Zhang, Stephen Wilson, and Linda Petzold. Pcmsp: A dataset for scientific action graphs extraction from polycrystalline materials synthesis procedure text. *arXiv preprint arXiv:2210.12401*, 2022.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5.

7 APPENDIX

7.1 KEY PHASES

Strategy 1: key_list = ['powder samples were prepared', 'powders were obtained', 'Polycrystalline ingots', 'ground together and pressed into pellets', 'starting materials were ground together', 'were prepared using bulk solid state methods', 'arc-melting stoichiometric quantities', 'ground together and pressed into pellets', 'starting materials were ground together', 'polycrystalline/Polycrystalline samples were', 'polycrystalline/Polycrystalline sample was']

Strategy 2: First it satisfies that 'polycrystalline' and 'Polycrystalline' in text and then perform a second round filtering, key_list = ['were/was synthesized/prepared', 'were/was first synthesized/prepared', 'were/was used', 'were/was first used', 'were/was obtained', 'were/was first obtained', 'were/was achieved', 'were/was first achieved']