Large scale Extraction of Composition and Properties from Materials Tables

Kausik Hira[1], **Mohd Zaki**[2],
Mausam[1], **N. M. Anoop Krishnan**[1,2]
[1]Yardi School of Artificial Intelligence, [2]Department of Civil Engineering
Indian Institute of Technology Delhi
{kausikhira, mohdzaki1995}@gmail.com
{mausam, krishnan}@iitd.ac.in

Abstract

In this study, we aim to develop the largest automated knowledge base (KB) of inorganic materials' compositions and properties by systematically extracting data from published research articles in the Materials Science (MatSci) domain. Since most material compositions and properties are reported in tables, their efficient extraction is essential for building large-scale knowledge repositories in this field. To this extent, we developed a framework combining two models, namely, DISCOMAT and PEGAMAT for extracting materials' compositions and properties respectively. Training data was generated through distant supervision using compositions and desired properties from existing databases and the corresponding journals, supplemented by rule-based parsers. Validation and test datasets were manually annotated by materials science experts. DISCOMAT achieved an F1 score of 71.49 for composition extraction, while PEGAMAT attained 86.90 for property extraction. We processed research papers published in 12 journals of the ScienceDirect database for our study and extracted more than 550,000 entries comprising around 100,000 glass material compositions with their properties, along with 137,000 compositions and 316,000 properties without their counterparts. The proposed models and the resulting database offer significant potential to advance the modeling and development of tailored materials.

1 Introduction

The discovery of novel materials has significantly accelerated human progress, with a crucial initial step being the comprehensive understanding of existing materials and their behaviors, which are fundamentally linked to their compositions, structure, and properties. This information is typically found in both structured formats (tables) and unstructured formats (text and images) across research papers, handbooks, and patents [17, 16]. Several efforts have been made in the past to develop materials domain-specific language models or other machine learning models that allow information extraction [13, 7, 18, 24, 2, 14, 23, 20–22]. Extracting and organizing this information into structured databases can accelerate the development of machine learning models, which in turn, can be used for accelerating materials discovery [12, 10].

Tables in scientific articles form a crucial repository of information [8, 12]. There have been several efforts towards information extraction from materials tables using regular expressions (regex) [9, 15], finetuning language models [24, 5, 19], or developing table specific graph neural networks [7]. There have been some additional efforts towards probing the LLMs to evaluate their understanding of tables and materials [4, 1, 22], which can then be used for extracting materials in an autonomous fashion [3]. However, these methods are limited to a smaller number of materials due to either the inherent limited generalizability of the model or the computational cost associated with inferencing employing LLMs, which makes it prohibitive to scale to large data.

AI4MAT workshop at 38th Conference on Neural Information Processing Systems (AI4MAT-NeurIPS 2024).

Here, we present a framework that allows automated extraction of materials compositions and properties from tables in scientific articles. The major contributions of our studies are as follows.

- **Unified model framework for tabular data extraction:** We introduce a comprehensive framework integrating two models, DiSCoMaT and PeGaMaT, trained on distantly-supervised data, enhanced with annotation codes and data augmentation, to facilitate large-scale extraction of compositions and properties, respectively, from tables present in the MatSci literature.
- **Comprehensive Materials Database:** Using the unified model framework, we synthesize large-scale data from MatSci literature, which comprises ≈ 550,000 entries, making it one of the largest materials databases.

2 Methodology

To build a knowledge base (KB) of compositions and properties in MatSci domain, our initial aim was to identify where these entities are reported within research articles. According to Hira et al.(2024) [8], 74% of papers reported compositions in tables, while 82% of the papers recorded the properties belonging to the experimented material in tables. Hence, we prioritized the extraction of compositions and properties from tables present in the documented literature of the MatSci field. We chose DiSCoMaT [7], a framework that handles different tables of complex structures along with other challenges and is currently the best composition extraction model in this field [22]. We upgraded DiSCoMaT by re-annotating its training data and modifying its post-processing rules. We followed the similar annotation approach used in DiSCoMaT and started working towards creating our own property extraction model from tables. When simple distant supervision on [11] proved to be insufficient for property extraction, as done in DiSCoMaT, we supplemented the training data with annotation codes followed by data augmentation. Finally, we developed our own graph attention network-based model for property extraction, namely PeGaMaT(Property Extraction using Graph Attention Networks from Material Science Tables). With PeGaMaT and an improved version of DiSCoMaT, we extracted composition and properties from tables belonging to different journals of the MatSci domain. We then integrated the extracted information to form our automated KB.

3 Extracting compositions using DiSCoMaT

We adapted DiSCoMaT from [7], and re-annotated the training data with rule-based constituent detectors, and re-trained DiSCoMaT on the newly annotated training data. We also modified the post-processing rules for partial-composition extraction and the regex parsers and introduced some less frequently seen compounds to the compound list. Lastly, we extended the scope of unit extraction from tables and captions to the "Result" and "Discussion" sections of the paper for more accurate unit extraction. These modifications led us to a gain of 6.74 points in the F1 score, as shown in Table 1. Notably, we have enhanced the extraction of materials from MCC-CI tables, which is the most frequently reported table structure in this field [8].

Table-type	Old F1 scores	New F1 scores
SCC-CI	78.21	77.07
MCC-CI	65.41	74.53
MCC-PI	51.66	52.82
All table types	63.53	70.27

Table 1: Evaluation of DiSCoMaT Performance: A comparison between previous and upgraded versions.

4 Extracting properties using PeGaMaT

4.1 Dataset Construction

We used the dataset reported in DiSCoMaT to develop our property-extraction model. We employed a similar approach to DiSCoMaT [7] for annotating training datasets. Similar to DiSCoMaT, we utilised the commercial glass database [11] and used distant supervision to align the properties mentioned in the training dataset. However, training data built on distant supervision [11] yielded low validation scores. Upon further analysis, it was observed that properties in several tables of the training data remain undetected. Thereafter, we supplemented the distantly supervised dataset with deterministic annotation codes, which were significantly more complex than those used in composition extraction. Some of the primary reasons are the presence of a huge number of tables having different properties with semantically similar abbreviations, in some cases exactly identical abbreviations are used to represent different entities and also the same properties reported with different acronyms across different articles, and many other factors mentioned in [8], makes the detection of the desired property challenging by any automated system. Despite making a substantial effort to write the supplementary codes for annotating the training data after distantly supervising it, we found that while our model was doing well for frequently studied properties such as density and glass transition temperature, it struggled with lesser seen properties like Abbe Number, Young's Modulus, Fracture Toughness. To address this issue, we employed data augmentation by integrating a column of the target property into tables that did not originally include that property but contained other properties commonly examined together. To reconcile the differences between the two table structures, assuming the table is column-oriented, we adjusted the column length of the source table (CL_s) to match the column length of the augmented target table (CL_t). Specifically, when $CL_s < CL_t$, we padded the source column with values within ±10% of the median to match CL_t, or clipped the source column if $CL_s > CL_t$. Given the inherent noise and potential inaccuracies in the distant supervision approach, the validation and test datasets were manually annotated by two co-authors.

4.2 Model Architecture

As 82% of the articles reported their findings related to properties in tables [8], our next goal was to create a model for extracting properties from tables of MatSci literature. We developed a GAT-based model, PeGaMaT(Property Extraction using Graph Attention Networks from Material Science Tables), and intended to extract 18 properties. PeGaMaT takes a table and caption text as input, upon which a directed graph is constructed where each table cell is a node, along with one additional node for the caption and a header node for each row and column. Each cell node is directed to its corresponding row header and column header node, and bidirectional edges exist between every cell node of the same row or column. The caption node propagates its information to every header node. Each node's embedding is initialized by running through the LM MatSciBERT [6]. We also introduced four structural constraints, which, when violated, lead to penalization of the loss function. This is followed by multiple post-processing rules specific to one or a particular set of properties. For instance, the Expansion Coefficient or Electrical Conductivity reported for a material often contains values in the range of $y \times 10^{-x}$, where only y is mentioned in table cells and $[10^{-x}]$ in the header cell. Post-processing rules take care of such cases to extract the correct values regarding each property. Additionally, we implemented outlier detection by verifying that the extracted property values fall within their expected range. These are some of the few examples of post-processing rules used by our model. As we have considered 18 properties, extracting the correct unit for a reported property is substantially more complex than that used in composition extraction.

4.3 Results

PeGaMaT takes the table along with its caption and text of the paper as input and gives out a list of tuples as output. The tuple consists of four elements – (Unique_ID, Property Name, Value, Unit). The Unique_ID is unique to every extracted tuple, which contains PII_TID_R_C_ID, where PII stands for article ID, and TID stands for table number belonging to that article. R and C represent the row and column index from which the information is obtained. We append the material ID if it is detected inside the table. This Unique_ID will provide us with information about the source of the generated tuple, which is essential to connect with other entities that were extracted by other models over the same table. PeGaMaT achieves a strong F1 score of 87.80 on the validation set and an F1 score of 86.90 on the test dataset. Notably, we achieved an F1 score of 90 or more for 8 out of the 18 properties on the test dataset, as detailed in Table. 2. We ignored the ID while evaluating the model, as our objective was to detect how efficiently we were able to extract the properties from each individual table. PeGaMaT detects material ID with an F1 score of 82.90. The material ID is essential when considering aligning information for a particular material across different tables.

5 Creating the MatSci Database

	Precision	Recall	F1
Materials extracted	81.55	63.65	71.49
Properties extracted	88.20	85.65	86.90
Database constructed	74.19	64.10	68.78

Table 3: Performance metrics of composition and property extracted, along with the database constructed.

To evaluate the metrics at the database level, we used the manually annotated gold standard data from PeGaMaT and DiSCoMaT, and compared it with the database generated by our models on the test dataset, obtaining a precision of 74.19 and F1 of 68.78, as described in Table. 3. We focused on scores excluding IDs because the primary compositions and

Sl. No.	Property Name	Validation F1	Test Precision	Test Recall	Test F1	Test Support
1	Density	92.9	87.7	94.7	91	75
2	Glass Transition Temp	97.2	90.4	90.4	90.4	104
3	Refractive Index	81	76.7	78.7	77.6	42
4	Abbe Number	100	100	100	100	1
5	Young's Modulus	93.3	74.2	100	85.2	23
6	Shear Modulus	100	100	81.2	89.7	16
7	Vickers Hardness	95.8	100	86.7	92.9	15
8	Poisson Ratio	82.8	64.7	73.3	68.8	15
9	Fracture Toughness	71.4	80	100	88.9	4
10	Crystallization Temp	92.6	93.5	86.7	90	83
11	Melting Temp	85.7	95.2	87	90.9	23
12	Electrical Conductivity	61.5	63.6	58.3	60.9	12
13	Temp of Softness	90.9	80	80	80	10
14	Annealing Point	83.3	75	75	75	8
15	Expansion Coefficient	88.9	100	95	97.4	20
16	Liquidus Temp	95.7	76.9	83.3	80	12
17	Bulk Modulus	100	100	71.4	83.3	14
18	Activation Energy	65.9	97.8	80.4	88.2	56

Table 2: Comprehensive evaluation of PeGaMat for each property extracted.

properties of the materials are linked by their orientation and not IDs, as in most of the articles the composition of the material along with its properties are reported in the same table.

We applied DiSCoMat and PeGaMat on 12 scientific journals of the ScienceDirect database to extract materials compositions and properties from the tables reported in these articles. The extracted information was integrated based on table orientation, assuming that the material's properties and composition would be presented in the same row for column-oriented tables or in the same column for row-oriented tables. In cases where composition and properties were reported in separate tables, material IDs were used to link the extracted data. Additionally, our database includes materials for which only compositional or property data were extracted, but not both. We obtained 76,166 entries having both composition and property of the material from the same table, followed by 21,945 similar entries obtained from inter-table with the help of material IDs. Additionally, we have extracted 137,003 only compositions, and 316,117 only properties, where we could not extract their corresponding counterparts, detailed in Table 4, where '*' refers to inter-table extraction.

Sl. No.	Journal Abbreviations	Papers	Tables	(C + P)	(C + P)*	C only	P only	Total entries
1	J. Solid State Chem.	4005	5889	3742	270	17955	15989	37956
2	Solid State Ionics	2568	3662	4582	305	6240	17892	29019
3	J. Phys. Chem. Solids	2397	3759	5652	523	6380	23293	35848
4	J. Non-Cryst. Solids	6590	10537	30805	9444	23689	45815	109753
5	Ceram. Int.	14024	21689	14727	7236	38024	89411	149398
6	Solid State Sci.	1831	2724	2288	626	6677	10019	19610
7	J. Nucl. Mater.	4609	6699	1934	968	11160	29870	43932
8	Thin Solid Films	4478	5684	2828	358	8081	24358	35625
9	Mater. Lett.	3524	3960	2450	166	5334	15368	23318
10	Solid State Commun.	1629	2339	3087	215	3338	15371	22011
11	Opt. Mater.	2610	3635	2914	1299	5308	17255	26776
12	J. Lumin.	1842	2453	1157	535	4817	11476	17985
Total		**50107**	**73030**	**76166**	**21945**	**137003**	**316177**	**551231**

Table 4: Summary of the knowledge-base formed by extracting information from different journals.

6 Conclusion and Future Work

In this article, we introduce property-extraction model PeGaMat, which extracts properties with high accuracy from tables reported in material science articles. By enhancing DiSCoMat and integrating it with PeGaMat we have developed the first large-scale automated material science database (MatSci DB), which meticulously captures compositions and properties from tables across a broad spectrum of research articles. These models have implications beyond material science, as various scientific domains like Biochemistry, Food Science, Pharmacology, and others use tables to report their findings. The models introduced and the database generated present substantial opportunities for advancing the design and development of specialized materials, along with accelerating the discovery of novel materials and improving the prediction models used in this domain.

References

[1] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[2] Markus J Buehler. Mechgpt, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities. *Applied Mechanics Reviews*, 76(2):021001, 2024.

[3] Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L Brinson. Extracting materials science data from scientific tables. In *ACL 2024 Workshop Language+ Molecules*, 2024.

[4] Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L Catherine Brinson. How well do large language models understand tables in materials science? *Integrating Materials and Manufacturing Innovation*, pages 1–19, 2024.

[5] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

[6] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, May 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00784-w. URL `https://www.nature.com/articles/s41524-022-00784-w`.

[7] Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan, and Mausam . DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.753. URL `https://aclanthology.org/2023.acl-long.753`.

[8] Kausik Hira, Mohd Zaki, Dhruvil Sheth, NM Anoop Krishnan, et al. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digital Discovery*, 3(5): 1021–1037, 2024.

[9] Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry Z. H. Gani, Yuriy Roman-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. doi: 10.1021/acscentsci.9b00193. URL `https://doi.org/10.1021/acscentsci.9b00193`.

[10] 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, 2017.

[11] Japan NGF. International glass database system, March 2019. URL `https://www.newglass.jp/interglad_n/gaiyo/info_e.html`.

[12] 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), 2020.

[13] Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264*, 2023.

[14] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. HoneyBee: Progressive instruction finetuning of large language models for materials science. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.380. URL `https://aclanthology.org/2023.findings-emnlp.380`.

[15] 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.

[16] Vineeth Venugopal and Elsa Olivetti. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217, 2024.

[17] Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7), 2021.

[18] 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.

[19] Gyeong Hoon Yi, Jiwoo Choi, Hyeongyun Song, Olivia Miano, Jaewoong Choi, Kihoon Bang, Byungju Lee, Seok Su Sohn, David Buttler, Anna Hiszpanski, et al. Matablegpt: Gpt-based table data extractor from materials science literature. *arXiv preprint arXiv:2406.05431*, 2024.

[20] Mohd Zaki, NM Anoop Krishnan, et al. Extracting processing and testing parameters from materials science literature for improved property prediction of glasses. *Chemical Engineering and Processing-Process Intensification*, 180:108607, 2022.

[21] Mohd Zaki, Sahith Reddy Namireddy, Tanu Pittie, Vaibhav Bihani, Shweta Rani Keshri, Vineeth Venugopal, Nitya Nand Gosvami, NM Anoop Krishnan, et al. Natural language processing-guided meta-analysis and structure factor database extraction from glass literature. *Journal of Non-Crystalline Solids: X*, 15:100103, 2022.

[22] Mohd Zaki, J Jayadeva, Mausam Mausam, and NM Anoop Krishnan. Mascqa: Investigating materials science knowledge of large language models. *Digital Discovery*, 2024.

[23] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

[24] Jiuyang Zhao, Shu Huang, and Jacqueline M Cole. Opticalbert and opticaltable-sqa: Text-and table-based language models for the optical-materials domain. *Journal of Chemical Information and Modeling*, 63(7):1961–1981, 2023.