Matter

CellPress

Matter of Opinion

NLP meets materials science: Quantifying the presentation of materials data in literature

Hasan M. Sayeed,[1] Wade Smallwood,[1] Sterling G. Baird,[1,2] and Taylor D. Sparks[1,*]

Large language models (LLMs) revolutionized how we engage with information. In materials science, we aim to leverage natural language processing to transform progress and discovery. Analyzing diverse materials science papers, we annotate data types and sources, laying the groundwork for targeted information extraction and LLM development.

Introduction

Materials science, an interdisciplinary field merging principles from physics, chemistry, and engineering, plays a pivotal role in technological advancement and societal progress. At its core, materials science focuses on the exploration and development of new materials. While the availability of structured data, particularly through advancements in machine learning, plays a pivotal role in expediting the discovery process,[1] materials science lags behind other physical sciences in this regard. This lag is mainly due to the diverse nature of data sources and the lack of centralization. The solution to this data challenge in materials science lies in making use of scientific publications, which are the primary means of communication in the field.[2] The extensive academic literature in this domain holds a wealth of data concerning material compositions, processing conditions, and performance properties. However, this valuable information is often scattered across the textual content, tables, and figures of research papers, posing a formidable challenge for data extraction and subsequent analysis.

Methods for data extraction in materials science encompass a combination of manual curation and automated techniques. Manual curation offers precision but is labor intensive and struggles to keep pace with the growing volume of materials science literature. Conversely, automated techniques are still evolving, grappling with the intricacies and diversity of data found in materials science papers.[3]

This informal study aims to delve into the patterns of data expression in materials science papers and uncover the relationships between data sources, laying the foundation for more efficient data extraction methods as a stepping stone toward our broader objective: to extract data from materials science papers and train a natural language processing (NLP) model to transform these data into a more machine-readable format. The insights gained from this analysis will play a pivotal role in informing the development of our model and may contribute to more efficient data extraction and analysis in materials science research.

Method

We randomly selected ten research papers spanning diverse subfields within materials science. This selection ensured a representative sample, encompassing a wide range of topics from superionic conductors to lattice anisotropy. The goal was to capture the breadth of data presentation practices in materials science literature.

The source or vehicle by which the data were presented within the papers was categorized by one of the following three definitions:

Text: This source refers to information presented within the narrative text of the research papers. It comprises descriptions, explanations, and discussions related to the materials, their compositions, processing conditions, and performance properties.

Tables: Organized tables in the papers were another key source of data. These tables typically offer structured and tabulated data, making it easier to access and interpret key information such as numerical values, experimental results, and comparative data.

Figures: Figures, including graphs, charts, diagrams, and micrographs, represent another significant data source. These visual representations often convey complex information succinctly and can include details about material properties, experimental results, and graphical depictions of scientific findings.

Within each paper, we focused on identifying four key data types or categories:

Chemical composition: Information related to the materials used, including chemical compositions, elements, and compounds.

[1]Department of Materials Science and Engineering, University of Utah, 122 Central Campus Dr, Salt Lake City, UT 84112, USA

[2]Acceleration Consortium, University of Toronto, 80 St George St, Toronto, ON M5S 3H6, Canada

*Correspondence: sparks@eng.utah.edu

https://doi.org/10.1016/j.matt.2023.12.032

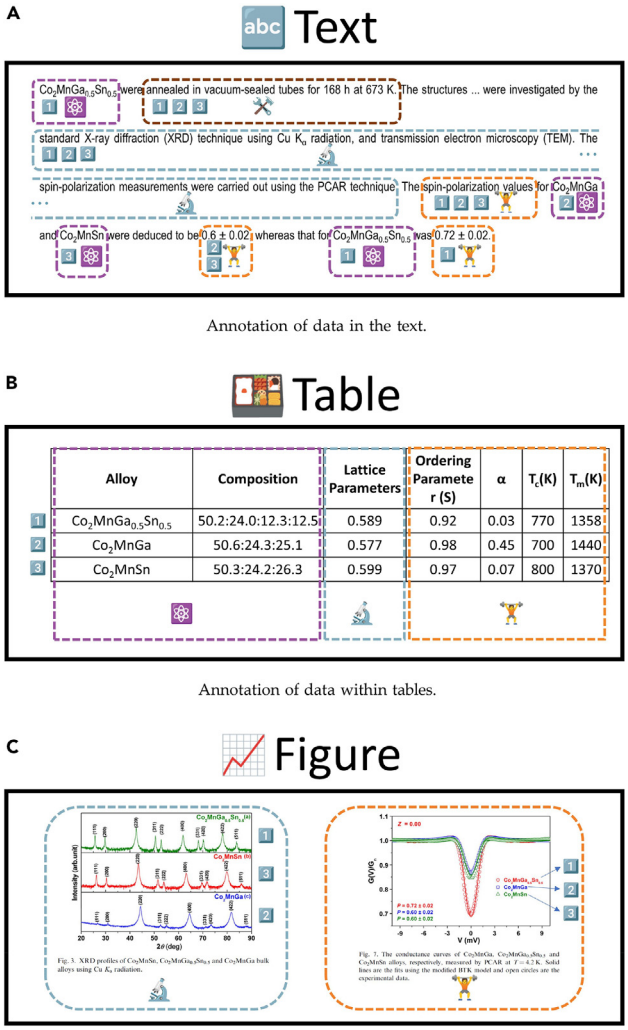

Figure 1. Example annotation format output

Data are distributed via three sources: (A) text, (B) table, and (C) figure. The type of data is either composition (⚛), processing conditions (🛠), characterization (🔬), or performance property (🏋). Multiple compositions were often discussed in the same paper. In this illustration, for example, we found three different compositions and denoted them as 1 ($Co_2MnGa_{0.5}Sn_{0.5}$), 2 (Co_2MnGa), and 3 (Co_2MnSn).

Processing conditions: Information related to the conditions under which materials were synthesized or processed, encompassing variables such as temperature, pressure, and synthesis methods.

Characterization data: Information pertaining to the characterization of materials, encompassing their structural, physical, and chemical properties.

Performance parameters: Information related to the performance properties of the materials, including mechanical, electrical, thermal, or other relevant performance metrics.

For clarity and brevity, we used emojis to represent data types. We have represented four types of data using emojis: ⚛ for composition, 🛠 for processing conditions, 🔬 for characterization, and 🏋 for performance property.

The data were hand-annotated by systematically isolating each sentence throughout the paper including the main body of the text, supplementary material, tables, and figures, then determining whether a relevant expression of data was presented. If it was determined that relevant data were presented, the source through which the data were presented along with the type or category of that data point was annotated. The most consistent test for whether relevant information was presented in a given sentence was whether the meaning of that sentence related to or progressed understanding with respect to one of the four defined data types or categories.

The output of this process is illustrated in Figure 1. This approach had the benefit of providing a visual display of the interplay of data types across text, tables, and figures while promoting consistency across each paper.

Results

A summary of the presence of data types in each of the data sources (text, tables, and figures) is provided in Table 1.

Similarly, Figure 2 provides a visual representation of the distribution percentages of each of the four data types across the three data sources, both individually across each paper and then in a combined view with the data sources as the x axis. This summary provides a holistic view of how the four key data types are distributed across these sources within the ten papers. Figure 2 serves as a valuable source for understanding and visualizing the broader patterns in data presentation.

Our investigation into the distribution of data within materials science papers revealed several noteworthy patterns and interconnections between data sources.

Data types across sources: In our analysis, we observed that textual content almost universally contains all four

Table 1. The presence of materials science data in various manuscripts on a distinct material topic

DOI	Material topic	Count	Text	Table	Figure
1. 10.1016/j.actamat.2009.02.024	spin polarization	3			
2. 10.1038/nmat3066	superionic conductivity	1			
3. 10.1103/PhysRevB.85.144109	pressure-dependent band gap	1			
4. 10.1016/j.jallcom.2012.01.122	bulk modulus and thermal expansion	14			
5. 10.1016/j.matchemphys.2011.07.080	specific heat and magnetic susceptibility	4			
6. 10.1103/PhysRevB.76.085110	thermoelectricity	1			
7. 10.1016/j.jallcom.2011.01.006	thermoelectricity	2			
8. 10.1038/nature09120	superconductivity	1			
9. 10.1016/j.physc.2010.12.001	superconductivity	4			
10. 10.1103/PhysRevB.87.064509	superconductivity	1			

Composition ( ) count is listed as more than one where a single manuscript provided relevant information across multiple chemical compositions. The data types determined as relevant for the purposes of this informal study were composition ( ), processing conditions ( ), characterization ( ), performance property ( ), or "missing" ( ). For example, the tables within the spin polarization manuscript (10.1016/j.actamat.2009.02.024) contain compositions, characterization data, and performance properties but lack any information about processing conditions. This is then labeled " " within the table ( ) column.

types of data under investigation, namely: chemical composition, processing conditions, characterization, and performance parameters. Tables are predominantly used to report the performance properties of materials while occasionally including characterization data. Figures, on the other hand, are predominantly utilized to present characterization data and performance properties.

Data isolation and interconnections: A critical observation from our data source analysis was that only processing conditions were consistently isolated in text alone. For other data types, particularly characterization and performance data, we noted a high degree of interconnection across all three sources. For instance, researchers commonly reported a performance parameter of interest in the textual content, such as ionic conductivity at room temperature, while providing detailed data, like a range of ionic conductivity values across various operating temperatures, in organized tables or figures. This interconnected presentation of data emphasizes the multifaceted nature of materials science research and the importance of accessing comprehensive information across various sources.

Challenges in textual content: An intriguing challenge emerged when considering the perspective of language model (LM) applications. It was evident that in the introduction sections of papers, researchers often discussed materials beyond the specific material of interest in the paper. These discussions typically provided context, such as the properties of related materials from similar crystal structures or otherwise recent findings from other research teams relevant to the informal study at hand. This context-setting practice, while valuable for human readers, may pose challenges for large language models (LLMs), potentially leading to confusion. LMs might have difficulty discerning whether the material discussed in the introduction is the primary focus of the paper or serves a secondary contextual purpose.

These findings underscore the intricate nature of data distribution within materials science papers and highlight the complexity of interconnections across different sources. Additionally, the potential challenges identified in textual content emphasize the need for further research and the development of more sophisticated NLP models tailored to the unique characteristics of materials science research papers.

Discussion

The predominant presence of all four data types within textual content is indicative of the role of narrative in materials science papers. Researchers often provide comprehensive details, descriptions, and contextual information about materials, processes, and properties in the text. This practice ensures that essential information is readily accessible to readers. Tables and figures, in contrast, are frequently employed for succinctly summarizing performance properties and presenting visual representations of data such as characterization and performance parameters.

The high degree of interconnected data across all three sources underscores the multidimensional nature of materials science research. Researchers often convey a performance parameter of interest in the text while offering a detailed dataset, including multiple operating conditions or characterizations, in tables or figures. This interconnected approach aids in presenting a comprehensive view of the material's behavior under various conditions. It promotes a holistic understanding of material properties, which is vital for advancing materials science research.

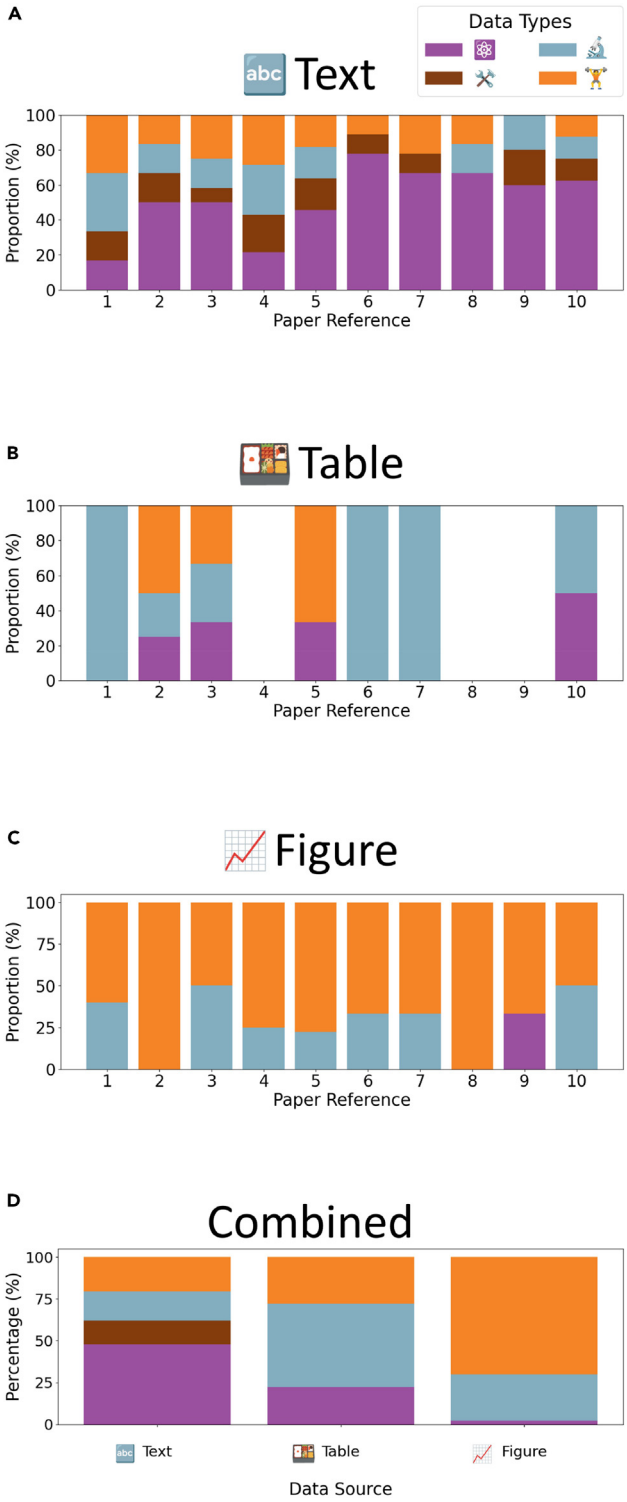

Figure 2. Per-paper summary of data distribution

(A–C) Per-paper summary of data distribution across three sources: (A) text, (B) table, and (C) figure. (D) A view of the combined distribution of data types when considering all sources (text, tables, and figures). Data type is either composition (🛠️), processing conditions (🔬), characterization (🏋️), or performance property (1️⃣), shown as the percentage of each data type present in the data source.

Notably, our informal study revealed a potential challenge in textual content. The practice of discussing materials other than the primary focus of the paper in the introduction may pose difficulties for NLP models. NLP models could misconstrue the intended focus of the paper, potentially impacting their ability to accurately extract and interpret data.

Conclusion

In conclusion, this informal study unveils essential insights into the intricate landscape of data distribution within materials science papers. Key takeaways include the prevalence of data in textual content, the focused use of tables and figures, and the strong interconnections between data types.

The importance of our informal study extends beyond academic curiosity. It can have profound implications for the broader materials science research community. The efficient extraction and integration of data, enabled by a deeper understanding of data sources and interconnections, enhances the accessibility and utility of valuable information contained in research papers. This is especially crucial in a field where insights can drive technological innovation and scientific progress.

Future directions

Building upon this preliminary analysis, several potential areas for further research and improvements in data extraction techniques emerge:

Semantic interoperability: Exploring techniques to enhance the semantic interoperability of data across text, tables, and figures may yield improvements in data integration. Developing standardized formats for data presentation could facilitate more efficient data extraction and analysis.

Multi-modal information extraction models: As a major finding from this informal study is that much of the bulk of any structured data is presented through tables and figures, development of optical character recognition (OCR) tools will be paramount to extracting the full extent of information presented in any given materials science literature.

Token reduction: A limitation of current LLMs is token quantity input, or the number of words you can enter into an LM in a single entry. The present informal study's findings have implications for the development of machine learning techniques to automate data extraction from text. These machine learning algorithms could be fine-tuned to accurately identify and categorize data types across diverse sources, which can in turn become an abbreviated text input for an LLM, ensuring maintained context.

In summary, this informal study serves as a stepping stone toward improved data accessibility and integration in materials science research. It paves the way for future endeavors, with the potential to shape the development of more sophisticated NLP and machine learning techniques tailored to the unique challenges and opportunities presented by materials science research papers.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) under grant number DMR 2334411. We extend our appreciation to the NSF for their financial support, which made this informal study possible.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT to rephrase and achieve coherence. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Baird, S.G., Liu, M., Sayeed, H.M., and Sparks, T.D. (2022). Data-driven materials discovery and synthesis using machine learning methods. Preprint at arXiv. https://doi.org/10.1016/B978-0-12-823144-9.00079-0.

2. Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.-J., and Hiszpanski, A.M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Appl. Phys. Rev. 7.

3. Hong, Z., Ward, L., Chard, K., Blaiszik, B., and Foster, I. (2021). Challenges and advances in information extraction from scientific literature: a review. JOM 73, 3383–3400.