Material Knowledge Integration Through Large Language Model Based Agents

Yi Shen Tew[a], **Zhihao Wang**[a], **Xiaonan Wang**[*a]

[a] *Department of Chemical Engineering, Tsinghua University, Beijing 100084, China* wangxiaonan@tsinghua.edu.cn

1. Introduction

In recent years, the advent of large language models (LLMs) such as GPT-4, Gemini, and DeepSeek has revolutionized the field of natural language processing by demonstrating unprecedented capabilities across various disciplines, including materials science[1, 2, 3]. These state-of-the-art LLMs excel at extracting structured information from unstructured text and are increasingly deployed as AI agents to automate data collection and management in scientific research. The rapid expansion of scientific literature in materials science poses significant challenges, as traditional databases struggle to capture the intricate relationships among experimental data, synthesis routes, and material properties. At the same time, this wealth of unstructured scientific data presents unprecedented opportunities to develop advanced data mining and machine learning techniques, which can extract meaningful insights and drive innovation in materials discovery and optimization[4]. This situation motivates the exploration of knowledge graphs, which provide a flexible framework for representing interconnected data, and knowledge fusion techniques that integrate heterogeneous sources into a cohesive, semantically rich repository, thereby enabling advanced semantic search and reasoning capabilities[5, 6].

2. Methodology

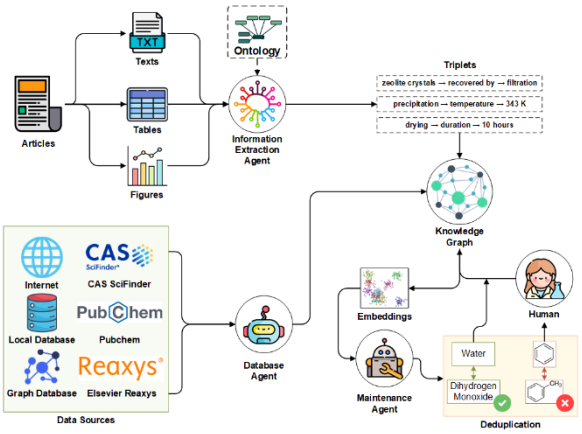

Fig. 1: The workflow of information extraction with LLM and multi-source data integration with knowledge graph

2.1 Literature Text Mining

The process begins when a researcher submits a search query to a platform such as the Web of Science API, which returns a list of article DOIs. Using web scrapers, the system retrieves the articles and processes them into a dataset in JSON format that includes full text, tables, figures, and metadata. This initial step establishes the raw material necessary for subsequent extraction and integration.

2.2 Large Language Model Information Extraction

Domain experts design a specialized ontology that guides the LLM agent in identifying and extracting relevant information from both the full text and tables of each article[7]. The agent performs named entity recognition (NER) and relationship extraction (RE) based on the predefined ontology, ensuring that key elements such as materials, synthesis methods, properties, and experimental conditions are accurately captured [8, 9]. In this process, the ontology is incorporated into the NER and RE process through prompt engineering inspired by domain knowledge in materials science, thereby enhancing extraction precision and contextual relevance[10].

2.3 Knowledge Graph Construction and Cleaning

Extracted information is converted into triplets and imported into a knowledge graph using Neo4j[11]. During this stage, an automated agent performs deduplication and entity linking to consolidate redundant entries and align semantically similar concepts under unified, canonical labels. By leveraging text embeddings, the agent identifies similar terms and assigns a confidence score to each resolution: high-confidence conflicts are automatically resolved, medium-confidence cases are flagged for human review, and low-confidence instances are rejected. This semi-automated process not only ensures that the knowledge graph maintains high data quality and consistency but also alleviates some of the considerable human labor and time required to manage a large, evolving knowledge graph.

2.4 Enrichment through AI Agents and External Databases

To further enrich the knowledge graph, AI agents perform targeted web searches and access external databases that complement the initial literature-derived dataset. In this step, the system incorporates data from PubChem, CAS SciFinder, Reaxys, NIST, and the Open Catalysis Project. CAS SciFinder contributes detailed references to related publications and patents, as well as reaction schemas and relevant experimental data, while PubChem offers chemical structure information, physical properties, and standardized compound identifiers that facilitate cross-referencing. Reaxys provides reaction pathways, yields, and step-by-step synthesis procedures, allowing the system to validate or refine existing entries in the knowledge graph. By integrating these complementary datasets, the enriched knowledge graph offers a more comprehensive view of the materials science domain.

2.5 Case Studies

One case study focuses on electrocatalytic urea synthesis from carbon dioxide and nitrate using a multi-metallic coupling site catalyst. In this study, the knowledge graph integrates detailed information on synthesis procedures, including key parameters such as reaction temperature, pH, precursor ratios, and

Material Knowledge Integration Through Large Language Model Based Agents

Yi Shen Tew[a], **Zhihao Wang**[a], **Xiaonan Wang***[a]

[a] *Department of Chemical Engineering, Tsinghua University, Beijing 100084, China* wangxiaonan@tsinghua.edu.cn

synthesis steps like hydrothermal treatment and thermal annealing. It also captures performance metrics of the catalyst, including Faradaic efficiency, urea yield, current density, and overpotential. By leveraging active learning, predictive models are trained on this knowledge graph to forecast optimal synthesis parameters and procedural steps, thereby guiding experimental design toward achieving superior catalyst performance. This integrated approach not only supports advanced machine learning tasks for material property estimation but also facilitates targeted question-answering systems, underscoring its potential to drive innovation in materials discovery and design.

3. Conclusion

This integrated methodology systematically addresses challenges inherent in materials science research, including the utilization of data in rapid expansion of unstructured literature, the complexity of domain-specific terminologies, and the heterogeneity of data sources such as textual, tabular, and image-based information. The proposed approach systematically captures, standardizes, and links domain-specific data—from experimental procedures and synthesis parameters to performance metrics—into a robust, semantically rich knowledge graph. The resulting resource not only enhances data accessibility and supports advanced machine learning applications but also provides a reliable foundation for predictive modeling and targeted question-answering systems.

Acknowledgments

This work is supported by the National Science and Technology Major Project of China (No.2022ZD0117501) and Tsinghua University Initiative Scientific Research Program.

References

[1] T. B. Brown *et al.*, Language models are few-shot learners, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 1877–1901.
[2] G. Team *et al.*, Gemini: A Family of Highly Capable Multimodal Models, Jun. 17, 2024, *arXiv*: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805.
[3] DeepSeek-AI *et al.*, DeepSeek-V3 Technical Report, Feb. 18, 2025, *arXiv*: arXiv:2412.19437. doi: 10.48550/arXiv.2412.19437.
[4] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, Opportunities and challenges of text mining in materials research, *iScience*, vol. 24, no. 3, p. 102155, Mar. 2021, doi: 10.1016/j.isci.2021.102155.
[5] L. Ehrlinger and W. Wöß, Towards a Definition of Knowledge Graphs, presented at the International Conference on Semantic Systems, https://www.semanticscholar.org/paper/Towards-a-Definition-of-Knowledge-Graphs-Ehrlinger-W%C3%B6%C3%9F/b18e4272a7b9fa2e1c970d258ab5ea99ed5e2284, 2016. Accessed: 27.02.25.
[6] X. L. Dong *et al.*, From data fusion to knowledge fusion, *Proc VLDB Endow*, vol. 7, no. 10, pp. 881–892, Jun. 2014, doi: 10.14778/2732951.2732962.
[7] T. R. Gruber, Toward principles for the design of ontologies used for knowledge sharing?, *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5, pp. 907–928, Nov. 1995, doi: 10.1006/ijhc.1995.1081.
[8] B. Babych and A. Hartley, Improving Machine Translation Quality with Automatic Named Entity Recognition, in *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, https://aclanthology.org/W03-2201/,2003. Accessed: 27.02.25.
[9] M. Eberts and A. Ulges, Span-based Joint Entity and Relation Extraction with Transformer Pre-training, 2020. doi: 10.3233/FAIA200321.
[10] H. Liu, H. Yin, Z. Luo, and X. Wang, Integrating Chemistry Knowledge in Large Language Models via Prompt Engineering, Apr. 22, 2024, *arXiv*: arXiv:2404.14467. doi: 10.48550/arXiv.2404.14467.
[11] J. Miller, Graph Database Applications and Concepts with Neo4j, in *SAIS 2013Proceedings*, https://aisel.aisnet.org/sais2013/24, 2013.