CLASS MATE: Cross-Lingual Semantic Search for Material Science driven by Knowledge Graphs

Aleksandr Perevalov[1,2], Jiveshwari Chinchghare[1], Mouli Krishna[1], Shivam Sharma[1], Amal Nimmy Lal[1], Aryman Deshwal[1], Andreas Both[2,3], and Axel-Cyrille Ngonga-Ngomo[1]

[1] Paderborn University, Paderborn, Germany
[2] Leipzig University of Applied Sciences, Leipzig, Germany
[3] DATEV eG, Nuremberg, Germany
alpe@mail.uni-paderborn.de

Abstract. As diverse linguistic backgrounds contribute valuable insights to scientific research, effective Cross-Lingual Semantic Search (CLSS) mechanisms, which often remain overlooked, become crucial. This paper introduces CLASS MATE[4]—a CLSS application working over material science knowledge graphs (KGs). Our work aims to bridge the digital language divide in the research community by employing advanced knowledge representation techniques. In particular, (1) we acquire our KG containing chemical substances with multilingual entity labels; (2) we implement a symbolic similarity-based named entity recognition algorithm; and (3) we develop a demo application employing the previous steps for retrieving information requested by a user from our KG and LOD sources in multiple languages. Our industry partner Springer Nature provided us with a KG as an information source to understand information needs. To the best of our knowledge, we made the first contribution to CLSS within material science.

Keywords: Cross-Lingual Semantic Search · Question Answering.

1 Introduction

Many research fields, including the field of material science, are undergoing a paradigm shift towards cross-lingual information retrieval due to the increasing international collaboration of researchers. While established commercial (e.g., Springer Materials [5]) and research (see Section 3) solutions exist for information search within the field, they often overlook the information access in different languages. This gap hinders researchers from fully benefiting from the diverse knowledge of the international scientific community. We introduce a pioneering exploration into a cross-lingual semantic search over knowledge graphs tailored specifically for material science, focusing on multilingual functionality. First, we enrich our RDF-based knowledge graph (KG) with domain-specific information (chemical substances and properties) from public knowledge graphs like Wikidata. We focus on supporting the following *languages*: English, German, Chinese,

[4] https://lass-kg.demos.dice-research.org/
[5] https://materials.springer.com/

Japanese, Arabic, Persian, Hindi, and Russian. Secondly, we develop a domain-specific algorithm for named entity recognition (NER) followed by fuzzy string matching for named entity linking (NEL) within our KG. Given the linked entities, a SPARQL query is executed on a KG to find the information requested by a user. Finally, the retrieved information is presented to a user in the language of their input query by injecting named entity labels of the respective language into a Large Language Model (LLM) through a translation prompt.

This work is a collaboration with Springer Nature, which provided us with their internal data for better domain understanding. The accessibility of our research is ensured by releasing the demo and open-source data[6], accompanied by a video tutorial[7]. To the best of our knowledge, this is the first contribution to tackle cross-lingual semantic search within the material science domain.

2 CLASS MATE

We define the *objective of semantic search* as to identify answers $\mathcal{A}$ that fulfill an informational need of a NL query q, utilizing a knowledge graph $\mathcal{KG}$. In its turn, cross-lingual semantic search systems *provide a possibility of searching for information in several languages* $l \in \mathcal{L}$, $|\mathcal{L}| > 1$. Hence, a user may pose a query in different languages: $q_{l_1}, \ldots, q_{l_n}$, where $n = |\mathcal{L}|$. At the same time, a system may search for answers in KGs in different languages: $\mathcal{KG}_{l_1}, \ldots, \mathcal{KG}_{l_n}$ (if multilingual information is not merged into one KG instance). For example, a user writes the question q_{l_i} (written in language l_i) and a system finds an answer $\mathcal{A}$ in the $\mathcal{KG}_{l_j}$ (instantiated for language l_j) [4].

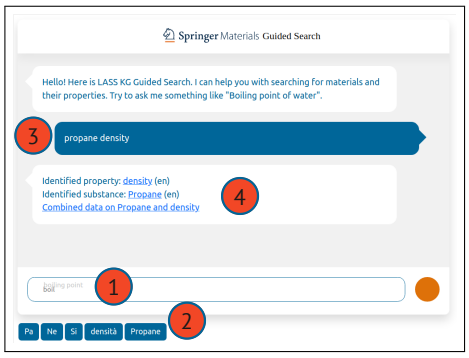
For instance, when asking "What is the density of Propane?" one expects to see a particular density *value* or a *phase diagram* of propane. While the values or the phase diagrams are provided by our project partner within their data, our task here is to correctly navigate a user to the right point where this data is available (independently from the question's language).

Knowledge Graph Enrichment As the initial *Springer Materials KG* (SM KG) from our industry partner contains only English entity representations, we enriched the graph with multilingual representations taken from Wikidata—an open collaborative KG. First, we narrowed down our search space to the only relevant entities using the entity types (e.g., Chemical Property (Q764285[8])). Second, we used English entity labels from the SM KG as an identifier for matching the respective entities in Wikidata. If there was a match, we connected it to the matched Wikidata entity via an `owl:sameAs` property. After that, the now integrated multilingual representations can be simply retrieved using SPARQL.

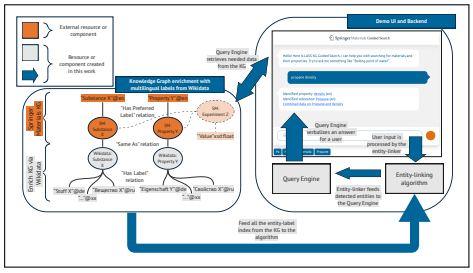
The Demo Engine The *user interface* is a Web application written using the React framework[9]. We designed it as a chat window with an input field, auto-completions, and auto-suggestions. Figure 1a demonstrates the user interface in

[6] `https://doi.org/10.6084/m9.figshare.24592428.v1`
[7] `https://youtu.be/2-YwbFWMW7Q`
[8] `https://www.wikidata.org/wiki/Q764285`
[9] `https://react.dev/`

detail. The *entity-linking algorithm* has a straightforward implementation. For



(a) User Interface: (1) the input field with auto-completions, (2) the auto-suggestions section, (3) the user's message, (4) the system's answer.



(b) Big picture of the workflow in our paper. First, we enriched our KG via connecting it to Wikidata. Secondly, we build a demo UI and backend that recognizes and links entities as well as retrieves them from our KG.

each entity type (`SUBSTANCE` (e.g., Propane), `PROPERTY` (e.g., density)) we use a separate entity-label index. Given a user's input query, we iterate through all the items of the label index and measure the ratio of the most similar sub-string. Finally, the algorithm returns the linked entity (ID, language, label, similarity), which label has the highest similarity within a user's search term (unless the similarity is lower than a pre-defined threshold). The *query engine* deals with the following cases: (1) Both `SUBSTANCE` and `PROPERTY` were recognized; (2) Either `SUBSTANCE` or `PROPERTY` was recognized; (3) No entities were recognized. In the first case, based on the entity ID obtained from the entity-linking step, we query the SM KG to identify whether (a) data on `SUBSTANCE` is available; (b) data on `PROPERTY` is available; (c) combined data on and `PROPERTY` is available. In the second case, we query either (a) or (b), In the last case, we do not perform any queries on the SM KG.

Finally, we verbalize the answer in a user-friendly form. We use HTML to include hyperlinks if the queries (a), (b), and (c) were successfully executed. We translate the HTML into the language of a user's initial input, which is detected upon the language tag of the identified entities, by using OpenAI's `gpt-3.5-turbo`[10] with the following prompt template: "`Translate to {lang} (ISO-639-1) while keeping HTML consistent: {HTML}`" (the entity labels are injected in the original language). Figure 1b demonstrates the overall "big picture" of the workflow within this paper.

3 Related Work

Text2Mol [3] uses a combined model to link natural language and molecular structures, enhancing performance significantly. SynKB [1] introduces an open-

[10] `https://platform.openai.com/docs/models/gpt-3-5`

source knowledge base that extracts chemical synthesis procedures from patents, providing accessible information for chemists. Unlike commercial databases, it is freely available and excels in query performance. MolT5 [2] is a framework addressing tasks like generating captions for molecules or creating molecules from text descriptions, displaying promising results in advancing molecule-language understanding. However, a critical gap exists in applying these techniques to material science, emphasizing the need for domain-specific semantic search functionalities tailored to its unique characteristics and complexities.

4 Conclusion

CLASS MATE introduces a significant step towards bridging linguistic barriers within the material science domain. However, certain limitations persist, such as reliance solely on preferred labels from a KG, and neglecting synonyms, typos, and other fluctuations in language when doing NEL. In addition, this approach requires proper and trustworthy evaluation. Despite these constraints, our demo application marks a crucial advancement, fostering collaboration and exploration in the complex landscape of material science. Moreover, considering synonyms, typos, and linguistic fluctuations in the NEL process and exploring the pretraining of domain-specific multilingual LMs and the fine-tuning of LLMs could significantly improve the effectiveness of our approach.

Acknowledgements. We would like to thank Springer Nature for supporting this work, in particular, we would like to say thank you to Stefan Scherer, Alexander Eckl, Volha Bryl, Marcel Karnstedt-Hulpus, and Harald Wirsching. This research has been funded by the Federal Ministry of Education and Research (BMBF) under grant 01IS17046. as part of the Software Campus project "LASS KG: Language Agnostic Semantic Search driven by Knowledge Graphs".

References

1. Bai, F., Ritter, A., Madrid, P., Freitag, D., Niekrasz, J.: SynKB: Semantic search for synthetic procedures. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 311–318. Association for Computational Linguistics, Abu Dhabi, UAE (2022). https://doi.org/10.18653/v1/2022.emnlp-demos.31
2. Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., Ji, H.: Translation between molecules and natural language. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 375–413. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.emnlp-main.26
3. Edwards, C., Zhai, C., Ji, H.: Text2Mol: Cross-modal molecule retrieval with natural language queries. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 595–607. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.47
4. Perevalov, A., Both, A., Ngomo, A.C.N.: Multilingual question answering systems for knowledge graphs—a survey (2024), https://www.semantic-web-journal.net/system/files/swj3633.pdf, under review