Energy Proceedings
Vol 44, 2024
ISSN 2004-2965

Construction of Knowledge Graph for Material and Energy Flow Integration in Biochemical Industry

Yishen Tew[1], Ling Fu[2], Jing Li[2], Tianrui Sun[2], Lanyu Li[1*], Xiaonan Wang [1*]

1 Department of Chemical Engineering, Tsinghua University, China

2 Siemens Technology, Siemens Ltd., China
(*Corresponding Author: lilanyu@mail.tsinghua.edu.cn | wangxiaonan@tsinghua.edu.cn)

ABSTRACT

Amid the escalating concerns of climate change and the mounting research papers on carbon neutrality and waste-to-energy solutions, comprehending crucial knowledge and technological trends in the chemical and energy sectors has become challenging. This study presents a novel approach combining large language models (LLM) and knowledge graphs (KG) to facilitate AI-supported knowledge retrieval. This work establishes a knowledge graph in the biochemical industry with 6,461 nodes and 8,969 relationships, emphasizing material and energy flow integration with the autonomous AI workflow. The graph's node attributes and relationships are analyzed using cosine similarities, with the capability to trace back to original literature through DOIs. This method not only underscores the relevance of node pairs in the graph but also links their similarities to the physical and chemical properties of materials. To sum up, this work provides an AI-enhanced tool that enables researchers and decision-makers to quickly build a knowledge base, learn about trends, and gain insights into specific fields.

Keywords: natural language processing, knowledge graph, interpretability, material flow integration, sustainable development

NOMENCLATURE

Abbreviations	
LLM	Large Language Model
KG	Knowledge Graph
AI	Artificial Intelligence
API	Application Programming Interface
NLP	Natural Language Processing
NER	Named Entity Recognition
RE	Relationship Extraction

1. INTRODUCTION

The pressing concern of climate change necessitates a paradigm shift in industries, especially in the chemical and energy sectors, towards sustainable practices. Central to this transformation is the circular economy model, which prioritizes efficient resource utilization, waste minimization, and energy efficiency. This approach aims to enhance production outputs while concurrently mitigating environmental degradation. However, an ever-expanding volume of research articles addressing waste-to-energy solutions, carbon neutrality, and integration techniques has made comprehensively navigating this knowledge sphere daunting.

In light of this, various innovative methodologies have emerged to tackle this challenge. For instance, Trokanas et al. [1] used an ontological semantic method, which was specifically designed to discern substance flow matching opportunities within industrial symbiosis. Complementing this, Davis and Aid [2] introduced an automated approach that utilized word vectors to pinpoint potential waste streams that could serve as alternative feedstocks, by gleaning information from a vast corpus of waste valorization literature and patents.

In this context, recent advancements in Artificial Intelligence (AI) provide useful tools, addressing the overflow of information mentioned earlier. Notably, cutting-edge language models, such as ChatGPT, have exhibited strong capability in tasks ranging from natural language understanding to generating nuanced human-like text. Furthermore, when armed with apt ontology and tailored prompts, these models show the potential to build comprehensive knowledge graphs [3]. Such innovations have great potential in accelerating research endeavors, facilitating swifter information extraction

This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

from extensive literature collections, and thereby helping to address complex issues like integrating material and energy flows in the biochemical field.

To make the most of this opportunity, our research utilizes the capabilities of ChatGPT (gpt-3.5-turbo) for information extraction and creates an automatic workflow to construct a knowledge graph. This AI-assisted method searches through academic literature, gathering important data, and using it to automatically build a detailed knowledge graph. The main benefit of this graph is its ability to combine various fields of knowledge related to material and energy flow integration within the biochemical sector. The constructed knowledge graph not only acts as a comprehensive knowledge base but also offers useful insights into potential integration strategies, further supporting the principles of the circular economy within the industry.

2. METHOD

Figure 1 shows the overall framework of this work. This research framework starts with sourcing literature articles. Through Elsevier's Text Mining API, relevant bio-chemical waste valorization articles are retrieved. Next, ChatGPT is prompted to perform natural language processing (NLP) tasks such as named entity recognition (NER) and relationship extraction (RE) to extract pertinent details about material properties and the valorization process. While the data aids in forming an initial knowledge graph, further manual refinement of the knowledge graph is still necessary. To enrich the graph further, data from Google's KG is used, focusing on the informational content of the material flow nodes. Finally, the Node2Vec algorithm is utilized to produce node embeddings to calculate cosine similarity and identify nodes with similar topological structures. More information is detailed in the below subsections.

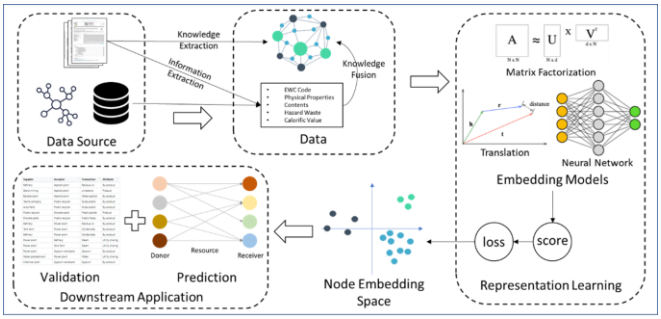

Fig. 1. Framework of automated information extraction by ChatGPT, knowledge graph construction and prediction workflow

2.1 Data Collecting and Processing

To curate the foundational literature for our research, we used the Text Mining API provided by Elsevier. The following query keywords are used to target biochemical waste valorization-related literature: "title-abs-key(Organic Waste OR Biomass OR Biological waste) AND title-abs-key(Material Exchange OR Waste Valorization)".

2.2 Information Extraction with ChatGPT

Following multiple rounds of prompt refinements, we have identified two primary NLP tasks: Named Entity Recognition (NER) and Relationship Extraction (RE). Both tasks are designed to yield improved results.

As shown in Table 1, for the NER task, we have specifically outlined ChatGPT's responsibilities and provided a detailed description of its intended purpose. We chose JSON as the output format because it is easily

Table 1. Prompts used for NER and RE tasks

Named Entity Recognition Prompt: Task: Extract and categorize named entities from a given text. Task 1: Extract Named Entities from the text. Task 2: Categorize each entity using Wikidata-style classes (e.g., ClassName). Task 3: Identify properties describing the entities. Name properties in camelCase (e.g., propertyName). Output Format: {"Entities": [{"entity": "exampleEntity", "abbreviation": "exampleAbbreviation", "class": "exampleClass", "properties":{"PropertyName1": ["value1"], "PropertyName2": ["value2a", "value2b"]}}]}
Relationship Extraction Prompt: Task: Identify relationships between entities based on given text. Direction: Ensure correct relationship direction from context. Format: Name relationships in UPPERCASE_SNAKE_CASE (e.g., USED_IN). Entities and Classes: Provided in the list below. <Output from NER task> Output Format: {"relations": [{ "relation": "exampleRelation", "from": {"class": "exampleClass1","entity": "exampleEntity1"}, "to": {"class": "exampleClass2", "entity": "exampleEntity2"}}]}

imported into Python and Neo4j for building knowledge graphs.

For the RE task, we leverage the entities recognized during the NER phase. ChatGPT's roles and goals are also clearly defined in this context, along with information about the expected output format.

2.3 Knowledge Graph Construction

We used data from the NER and RE tasks and imported them into Neo4j to build our knowledge graph. A key component in building this graph is ontology, which defines the types of entities and how they relate to each other. Given the context provided, we are concentrating on material transformation. Mainly, we recognize two entities: "Matter" and "Process." The "Matter" entity covers resources, waste, intermediate products, and final products. The "Process" entity describes how materials change, identified by attributes like name, type, and related technologies.

Using ChatGPT, we generate entity tags based on the context of the original text, which sometimes leads to variations or multiple tags for the same concept. This calls for aligning or merging tags that refer to the same thing. As an example, different relationship tags like “USE_IN”, “USED”, and “USED_FOR” might be combined into one, and similar node labels like “ChemicalSubstance” and “ChemicalSubstances” could be unified.

Optimizing downstream tasks for material flow prediction, the node labels generated by ChatGPT underwent further refinement. Nodes pertaining to materials are categorized under the "Material" label, encompassing all tags related to materials. Conversely, nodes concerning processes are collated under the "Process" label, as shown in Figure 2.

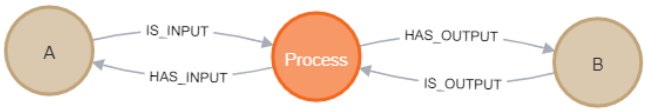

Fig. 2. Ontology of material valorization process

2.4 Knowledge Fusion with Google KG

Google's Knowledge Graph is a vast repository of factual information stored in graph format. The information in this graph is sourced from open datasets, web pages, books, etc. Google offers the Knowledge Graph Search API [4], which developers can use to query the Google Knowledge Graph and retrieve related information. By leveraging Google's Knowledge Graph, descriptions, definitions, and web links for entities in the local knowledge graph can be enhanced.

2.5 Calculating Node Similarity

In our work with knowledge graphs, we used Node2Vec to create node embeddings. Node2Vec is based on deep learning concepts and is adapted from the Word2Vec model [5]. But instead of words, it creates vector representations for nodes using random walks on graphs.

After generating these embeddings, we measured the similarity between nodes by computing the cosine similarity between their embeddings [6]. This allows us to identify nodes with similar structures or characteristics in the knowledge graph, giving us valuable insights.

3. MATERIAL AND METHODS

3.1 NER and RE task performed by ChatGPT

Using the application of ChatGPT for the construction of the knowledge graph related to the valorization of bio-engineering waste, this research automated the extraction of entities and their attributes and relationships from the literature. The utilization of ChatGPT considerably simplified the named entity recognition (NER), relationship extraction (RE), and attribute extraction processes.

Post multiple prompt engineering iterations, notable stability in extraction results was achieved. A total of 11,065 entities were identified. The unique labels of these entities summed up to 3,089, out of which 145 labels appeared more than ten times. The relationship extraction task produced a total of 20,654 relationships.

To optimize downstream tasks related to material flow prediction, nodes related to materials are labeled as “Material” while those related to processes are labeled as “Process”. The final subgraph consists of 6,461 nodes and 8,969 relationships, with 1,548 unique node tags and 964 unique relationship tags. Figure 3 shows the word cloud of named entities and relationships generated by ChatGPT.

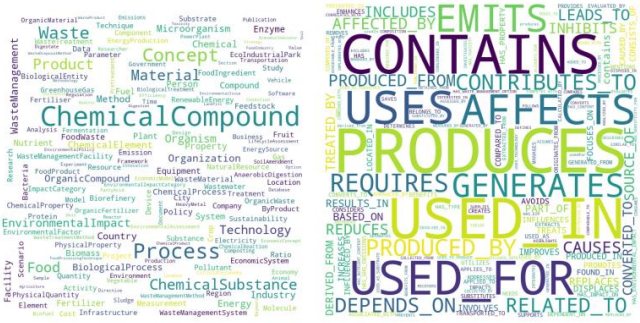

Fig. 3. Word cloud of named entities (left) and relationships (right) generated by ChatGPT

3.2 Predicting and explaining similar nodes

Node embeddings generated by Node2Vec are utilized to calculate cosine similarities between nodes, revealing some shared properties.

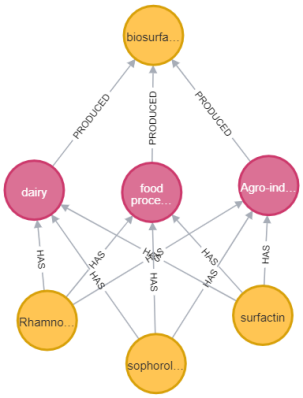

Fig. 4. Graph of relationships between surfactin and sophorolipids

Some node pairs with high cosine similarity can be explained by analyzing their structure. For example, "Surfactin", "Sophorolipids", and "Rhamnolipid" have a similar structure in the knowledge graph as shown in Figure 4. They can be obtained from dairy products, food processing industries, and agricultural fertilizers and can be further-processed to produce biosurfactants. In addition, they also have a broad spectrum of applications including in water treatment.

4. CONCLUSIONS

In this work, we developed an AI-supported workflow that uses ChatGPT to extract information from a collection of literature articles and maps material flow in the bio-chemical engineering field in the form of a knowledge graph, and calls Node2Vec to analyze the graph's structure. It is aimed to assist stakeholders amidst growing discussions on waste valorization and carbon neutrality. However, the literature scope of this study was limited, which affected the depth of the knowledge graph. In addition, although ChatGPT was effective in extracting entities, it struggled with consistent labeling and handling of large-scale data. In future endeavors, we will broaden our literature base, fine-tune NLP techniques, and diversify our analytical approaches. We will also use other analytical methods besides Node2Vec and cosine similarity to gain more insights into material flows in biochemical engineering and other energy-saving and carbon-neutrality-related fields. Such progressions would offer invaluable insights and enriched resources for industry stakeholders.

DECLARATION OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors read and approved the final manuscript.

REFERENCE

[1] Trokanas, N., Cecelja, F., & Raafat, T. (2014). Semantic input/output matching for waste processing in industrial symbiosis. Computers & Chemical Engineering, 66, 259–268. https://doi.org/10.1016/j.compchemeng.2014.02.010

[2] Davis, C., & Aid, G. (2022). Machine learning-assisted industrial symbiosis: Testing the ability of word vectors to estimate similarity for material substitutions. Journal of Industrial Ecology, 26(1), 27–43. https://doi.org/10.1111/jiec.13245

[3] Pan, J. Z., Razniewski, S., Kalo, J.-C., Singhania, S., Chen, J., Dietze, S., Jabeen, H., Omeliyanenko, J., Zhang, W., Lissandrini, M., Biswas, R., de Melo, G., Bonifati, A., Vakaj, E., Dragoni, M., & Graux, D. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges (arXiv:2308.06374). arXiv. https://doi.org/10.48550/arXiv.2308.06374

[4] Google Inc. (2022, November 15). Google Knowledge Graph Search API. https://developers.google.com/knowledge-graph

[5] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks (arXiv:1607.00653). arXiv. https://doi.org/10.48550/arXiv.1607.00653

[6] Han, J., Kamber, M., & Pei, J. (2012). 2—Getting to Know Your Data. In J. Han, M. Kamber, & J. Pei (Eds.), Data Mining (Third Edition) (pp. 39–82). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00002-2