Knowledge Graph Construction Using Bio-Medical Dataset

1st Abhimanyu S
Department of Computer Science and Applications
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Amritapuri, India
amscu3cds20003@am.students.amrita.edu

2nd Athul S Kumar
Department of Computer Science and Applications
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Amritapuri, India
amscu3cds20019@am.students.amrita.edu

3rd Athulraj
Department of Computer Science and Applications
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Amritapuri, India
amscu3cds20020@am.students.amrita.edu

4th Balasankar P
Department of Computer Science and Applications
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Amritapuri, India
amscu3cds20021@am.students.amrita.edu

***Abstract*—Our project focuses on creating a Knowledge Graph within the biomedical field, utilizing the NCBI dataset, to establish relationships between disease-disease, chemical-chemical, and disease-chemical entities. Such Knowledge Graphs can be used in the development of chatbots, question-answering systems, and relation extraction systems, providing a means to answer complex biomedical queries and facilitate drug discovery.**

I. Introduction

A Knowledge Graph(KG) is a powerful tool used in artificial intelligence and machine learning to represent knowledge in a structured format. It involves organizing information into entities, attributes, and relationships, creating a web-like structure of interconnected data. This enables machines to understand and reason about the relationships between different pieces of information, providing a more intuitive and efficient way to process complex data. KGs are increasingly being used in a variety of applications, including search engines, chatbots, and recommendation systems, to provide more personalized and accurate results to users.

The use of biomedical KGs in Natural Language Processing (NLP) has gained increasing attention in recent years due to their ability to provide a structured representation of biomedical information. Biomedical KGs are knowledge-based systems that store and represent biomedical concepts as nodes and their relationships as edges, allowing for the integration of diverse biomedical data sources. They have been employed to aid in various NLP tasks such as entity recognition, relationship extraction, and machine learning. One major advantage of biomedical KGs is their potential to enhance data integration in biomedical research. With the vast amount of biomedical data available from multiple sources, including electronic health records, clinical trials, and scientific literature, there is a growing need for effective data integration to support comprehensive analysis of biomedical concepts. Biomedical KGs can provide a framework for integrating such data, allowing researchers to explore complex relationships between various entities and gain a deeper understanding of biomedical concepts.

In tasks such as information extraction from scientific literature, where entities such as genes, diseases, and drugs are frequently mentioned, the use of Biomedical KGs can help identify the correct entities by considering their relationships with other entities. This enhances the accuracy of NLP systems and enables more effective knowledge discovery. Finally, Biomedical KGs can be used to train machine learning models in NLP tasks. By using a KG as a structured representation of biomedical information, researchers can improve the accuracy of machine learning models in various NLP tasks.

This research paper explores the importance of Biomedical KGs in NLP in data science. We examine the different applications of KGs in NLP and how they aid in data integration, entity recognition, relationship extraction, and machine learning. We also discuss the current challenges and opportunities in the field and explore the potential for future research in this area.

When compared to relational models or other alternatives, utilizing a graph-based abstraction for representing information offers several advantages. Graphs offer a lucid and comprehensible abstraction, particularly for various domains, where the edges and paths represent complex relationships between the constituents of the domain. Humans possess exceptional abilities in information processing, reasoning, and knowledge interpretation due to their accumulated knowledge over the years. For a considerable period, computer technology and AI have been pursuing a similar objective, and the means for machines to interpret this knowledge is via knowledge

representation. The Semantic Network of KG is a representation of a community of real-world entities such as items, events, conditions, or concepts and depicts the relationships between them. Typically, this data is stored in a graph database and visualized as a graph structure. What sets a KG apart is its ability to re-reason itself and generate new insights and inferences, unlike a regular database that only stores populated knowledge. An KG is dynamic, and the graph itself understands the connections between entities, eliminating the need for manual programming of every new piece of data. The success of supervised machine learning heavily depends on the quantity and quality of available training data, sometimes even more than the selection of specific learning algorithms.

A graphical community comprises nodes that represent items and arcs that depict the relationships between those objects. The following example illustrates the visualization of the networks:

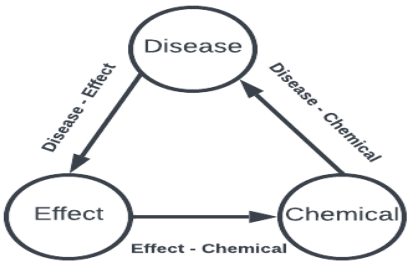

Fig. 1. Relation between entities

The above Fig1 depicts a graphical representation illustrating the triplet relation between Disease, Effect and Chemical. It features three interconnected nodes: Disease, Effect, and Chemical, with labelled arrows indicating the relation between them.

This paper proposes useful approaches for creating a KG from the NCBI dataset. A Biomedical KG is a powerful data structure that stores structured and unstructured information from diverse sources related to the Biomedical domain. It is designed to create a comprehensive and well-organized representation of medical concepts, such as diseases, treatments, drugs, and medical procedures. This wealth of information can then be utilized to generate valuable insights and support decision-making in the medical field.

II. Related Works

Biomedical knowledge graphs have been extensively studied in recent years, and there is a growing body of research on their development and applications.

A. Domain specific

DisGeNET is a knowledge graph that focuses on disease-related knowledge, including disease-gene associations, genomic variants, and drug-disease relationships. It has been used for various applications, including drug discovery and precision medicine.

HPO is a knowledge graph that represents human phenotype information, including abnormal phenotypes observed in patients with genetic disorders. It has been used in various applications, including clinical decision support and genetic diagnosis.

OBO is a collection of interoperable ontologies that covers various biomedical domains, including anatomy, diseases, and genomics. It has been used in various applications, including data integration and semantic annotation.

B. Open domain

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a widely used knowledge graph in the field of bioinformatics. It provides comprehensive information on biological pathways, diseases, and drugs, and has been used for various applications, including drug discovery and systems biology.

BioPax is another widely used knowledge graph in bioinformatics, representing biological pathways and their interactions. It is based on a community-driven standard and has been used in various applications, including pathway analysis and network visualization.

III. Methodology

The detailed pipeline model of the proposed system is shown in Figure **??**. Knowledge Graphs can be created using different resources, such as text or pre-existing databases. Existing datasets are commonly used to develop Knowledge Graphs in the biomedical field. Domain experts use a variety of methods, including text mining and manual curation, to create these databases. Manual curation is a time-consuming process that involves experts reading papers and identifying sentences that provide relevant information.

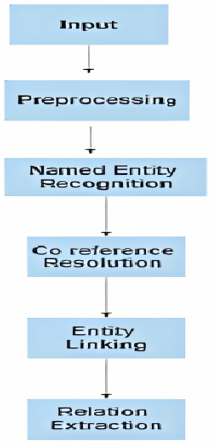

Fig. 2. Flowchart

Automated methods, such as machine learning or natural language processing, can identify pertinent sentences more

quickly than manual curation. Also, it allows for the rapid and efficient construction of Knowledge Graphs.

The text data which contains 92890 words is pre-processed which contains the process of cleaning, reduction, and simplification is done. Which is then used to identify all the coreferent expressions in the biomedical texts and resolved them to their respective entities. Which is then linked to create a CSV file this CSV is fed into a graphical database to create a graph

To model relationships between two entities in a Knowledge Graph, the triplet format is commonly employed. The relationship between two entities is represented as:

- Disease – Effect
- Disease – Chemical
- Effect – Chemical

Our model utilized different algorithms for the project, and in the upcoming sections, we will provide a comprehensive breakdown of each algorithm, explaining their workings in a clear and concise manner.

A. MinIE (Information Extraction)

For our project, which focuses on creating a KG, we utilized the MinIE algorithm. Specifically, MinIE was employed to extract triplets from a text file, including Disease-Effect, Disease-Chemical, and Effect-Chemical relationships.

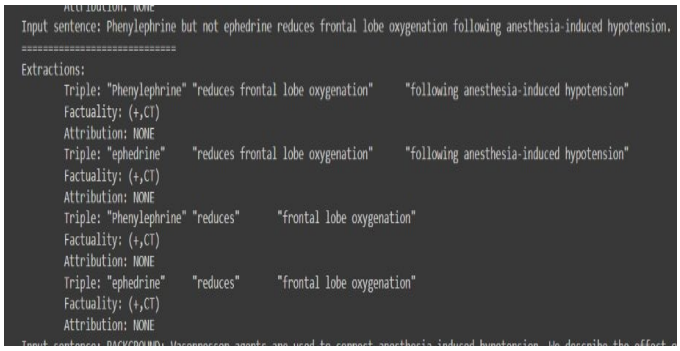

Fig. 3. Relation between entities

MinIE, or Minimally Supervised Information Extraction, played a pivotal role in our project by automatically identifying and extracting pertinent information from the text data. Through efficient text scanning and advanced natural language processing techniques, MinIE successfully identified and extracted the desired triplets.

The implementation of MinIE proved invaluable to our project, streamlining the process of KG creation by automating the extraction of important relationships between diseases, effects, and chemicals. This algorithm significantly reduced the time and effort required for the manual extraction and categorization of this information.

In summary, the utilization of the MinIE algorithm facilitated the extraction of disease-effect, disease-chemical, and effect-chemical triplets from the text file, contributing to the successful construction of our KG.

B. Coreference Resolution

Coreference resolution is a task in natural language processing that identifies all the expressions in a text that refer to the same entity. This can be challenging in the biomedical domain due to technical language, but it can help create knowledge graphs by linking related entities.

In our research project, we employed Coreference resolution, which played a vital role in enhancing the quality and coherence of our KG. Coreference resolution is the process of determining the referential relationships between different expressions in a text, specifically identifying instances where different expressions refer to the same entity.

By utilizing Coreference resolution techniques, we were able to resolve ambiguous references within the text and establish a more accurate representation of the relationships between diseases, effects, and chemicals in our KG. This process enabled us to connect related information across different parts of the text, ensuring a more comprehensive and coherent knowledge representation.

Coreference resolution proved to be highly beneficial for our project as it improved the overall accuracy and consistency of the extracted triplets. By resolving references to the same entity, we were able to avoid duplication and create a more concise and informative KG. Additionally, it enhanced the interpretability and usability of the KG by providing a clearer understanding of the interconnectedness between different entities and their associated information.

In summary, the application of Coreference resolution in our project significantly improved the quality and coherence of our KG. By resolving ambiguous references and establishing accurate relationships, we were able to create a more comprehensive and interpretable knowledge representation.

C. Sentence Simplification

Splitting sentences into clauses can be a useful technique for easier processing of compound and complex sentences in natural language processing tasks. This can be achieved by using the dependency parse of the sentence, which identifies the grammatical relationships between words in the sentence.

To split a sentence into its component clauses, follow these steps:

- Identify the root of the sentence (usually a verb).
- Traverse the dependency tree of the sentence from the root, identifying any dependent clauses that are attached to the root.
- For each dependent clause, repeat the process, identifying the root of the clause and any dependent clauses attached to it.
- Continue recursively until all clauses have been identified.

By splitting sentences into clauses, it becomes easier to perform tasks such as sentiment analysis, text classification, and information extraction, as each clause can be analyzed independently.

D. Named Entity Recognition (NER)

Named Entity Recognition (NER) is a natural language processing technique used in the biomedical domain to identify and categorize specific entities such as genes, proteins, diseases, and chemicals within a text. It is a crucial component in the creation of a KG in this field as it allows for the extraction of relevant information from large volumes of text data.

The following are the steps involved in performing NER:

- Detecting the entities from the text: In this step, the algorithm scans the text to identify potential entities. This is usually done by looking for patterns in the text or using statistical models.
- Classifying the entities into different categories: Once the entities have been detected, they are classified into different categories. The categories depend on the application and can range from generic categories like person, organization, and location, to more specific categories like disease, drug, or gene. The classification can be done using rules-based approaches, machine learning models, or a combination of both.

E. Relation Extraction

Relation extraction is a natural language processing (NLP) technique that involves identifying and extracting meaningful relationships or associations between entities in text. This technique is used to build KGs, which represent knowledge in a structured and machine-readable form. In the biomedical domain, relation extraction is particularly important for knowledge discovery and integration from large volumes of biomedical literature.

In our project, we used relation extraction to automatically extract and classify relationships between biomedical entities such as drugs, genes, diseases, and symptoms from unstructured text. We employed a supervised machine learning approach that leveraged pre-existing knowledge in the form of annotated training data. Our system was able to accurately identify and classify various types of relationships, including disease-effect, disease-chemical, and effect-chemical indications. Coreference resolution proved to be highly beneficial for our project as it improved the overall accuracy and consistency of the extracted triplets. By resolving references to the same entity, we were able to avoid duplication and create a more concise and informative KG. Additionally, it enhanced the interpretability and usability of the KG by providing a clearer understanding of the interconnectedness between different entities and their associated information.

In summary, the application of Coreference resolution in our project significantly improved the quality and coherence of our KG. By resolving ambiguous references and establishing accurate relationships, we were able to create a more comprehensive and interpretable knowledge representation.

F. Creating the Knowledge Graph

To create a KG using Neo4j, we can follow the following steps:

- Load the preprocessed data in CSV format into Neo4j.
- Use the Mapping editor to define the relationship types between entities. This involves identifying which columns in the CSV file correspond to entities and which correspond to relationships between them.
- Once the mapping is complete, we can create the knowledge graph by clicking on the ”Create” button. Neo4j will then use the mapping information to create a graph database with nodes representing entities and edges representing relationships between them.
- We can use queries in the Cypher query language to extract insights from the knowledge graph. For example, we could query for all the diseases that are related to a particular chemical or all the treatments that are associated with a particular disease.
- We can also use tools like GraphXR to visualize the KG in 3D and explore the relationships between entities more easily.

We can also use tools like GraphXR to visualize the knowledge graph in 3D and explore the relationships between entities more easily.

G. Limitations of knowledge graph in machine learning

- Limited scope: KGs are limited to specific domains, and their effectiveness is dependent on the quality and completeness of the data sources available for that domain. They are not well-suited for handling broader or more abstract concepts.
- Data quality issues: The accuracy and completeness of the data used to build a KG can have a significant impact on its effectiveness. Incomplete or inaccurate data can lead to incorrect inferences and inconsistent results.
- Maintenance costs: KGs require ongoing maintenance and updates to ensure that the data is accurate and up-to-date. This can be time-consuming and expensive, particularly for large and complex graphs.
- Semantic ambiguity: Natural language is often ambiguous and context-dependent, and different sources may use different terminology or conceptual frameworks. Resolving these semantic differences can be challenging and require significant effort.
- Lack of context: KGs may not capture the full context in which a particular piece of information is relevant, leading to misunderstandings or incorrect inferences.
- Difficulty with inference: KGs may struggle to make inferences or draw conclusions based on incomplete or ambiguous data. This can limit their ability to provide useful insights or predictions.
- Limited ability to handle uncertainty: KGs may struggle to represent uncertain or probabilistic information, such as estimates or predictions. This can limit their usefulness in some domains.
- Limited ability to learn: While KGs can be updated and refined manually, they may struggle to adapt to new information or learn from experience in the way that humans do.

H. Result

a) : After preprocessing the text data, we created a knowledge graph that serves as the input for several chatbots. It can be used to capture different relations to help different systems to fetch knowledge. The key insight of KG is that graphs provide a simple, flexible, intuitive and yet powerful abstraction for representing and integrating diverse data at a large scale.

b) : Graphs have long been used to represent data and knowledge in areas such as Graph Algorithms and Theory, Graph Databases, Information Extraction, Knowledge Representation, and Machine Learning. The advances in these areas can now be unified and applied to KGs.

c) T: he result of a biomedical knowledge graph is a comprehensive, organized map of medical concepts and their relationships. It can help to identify potential connections between diseases, and drugs, as well as uncover important insights. This information can be used to improve patient care, develop new treatments, and advance medical research. Additionally, the information contained in a biomedical knowledge graph can be used to inform decisions within the medical field and to support healthcare professionals in their work.

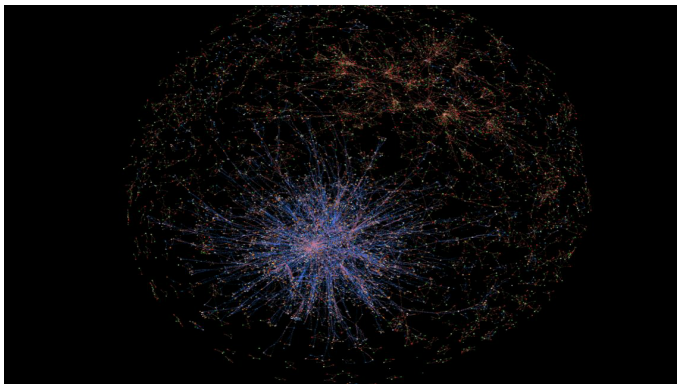

Fig. 4.

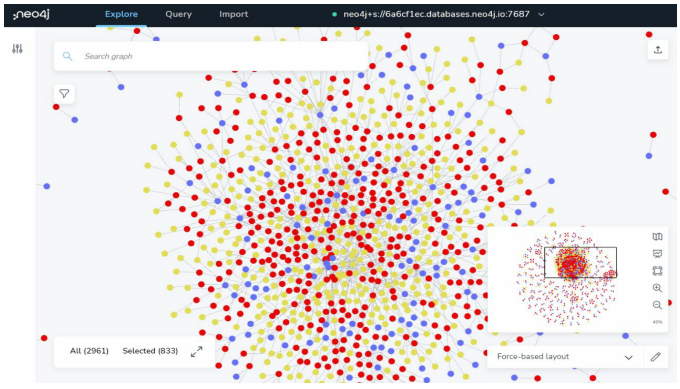

Fig. 5.

A comprehensive analysis of the input data was conducted. The dataset consisted of 170 pages, encompassing a total of 92,890 words. The textual content of the dataset was further characterized by its character count, with 559,342 characters excluding spaces and 651,231 characters including spaces. Additionally, the dataset contained 999 paragraphs and was structured into 6,914 lines. These statistics provide valuable insights into the scale and complexity of the dataset, laying the foundation for subsequent KG construction and analysis. A visual representation of these statistics can be found in Figure 6, which enhances the overall presentation and understanding of the research paper.

Fig. 6. Input Data

In the context of creating a KG in Neo4j, the output data of this project has been analyzed and the statistics are as follows: The KG comprises a total of 3,890 nodes, which accounts for 2 percentage of the total anticipated nodes (200,000). Furthermore, there are 6,171 relationships in the KG, representing 2 percentage of the expected relationships (400,000). These statistics provide an overview of the progress made in constructing the KG, indicating the current state of completion and giving insights into the scale of the knowledge representation. The visual representation of these statistics can be found in Figure 7, which serves as a valuable visual aid to enhance the understanding and presentation of the research paper.

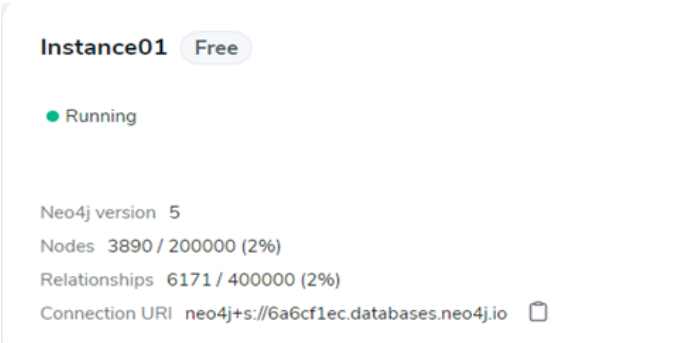

Fig. 7. Output Data

IV. Acknowledgment

We express our heartfelt gratitude to all the individuals and organizations who have contributed to the success of this project. Without their kind support and assistance, this endeavor would not have been possible.

We would like to extend our sincere thanks to Veena G, Assistant Professor at the School of Computing, Amritapuri, for her invaluable guidance, constant supervision, and provision of necessary information throughout the project. Her expertise and support were instrumental in completing the project successfully.

We are deeply grateful to Amrita Vishwa Vidyapeetham for providing us with the platform and resources to undertake this project. Their continued support and encouragement have been instrumental in our research journey, and we sincerely acknowledge their contribution.

REFERENCES

[1] Jianbo Yuan, Zhiwei Jin, Han Guo, Hongxia Jin, Xianchao Zhang, Tristram Smith and Jiebo Luo

[2] Sousa, R.T., Silva, S. and Pesquita, C. Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics 21, 6 (2020).

[3] David N. Nicholson, Casey S. Greene, Constructing knowledge graphs and their biomedical applications, Computational and Structural Biotechnology Journal, Volume 18, 2020,

[4] Chen X and Güttel S. (2022). An Efficient Aggregation Method for the Symbolic Representation of Temporal Data. ACM Transactions on Knowledge Discovery from Data. 17:1. (1-22). Online publication date: 28-Feb-2023.

[5] Bingcong Xue, Lei Zou, "Knowledge Graph Quality Management: A Comprehensive Survey", IEEE Transactions on Knowledge and Data Engineering, vol.35, no.5, pp.4969-4988, 2023.

[6] Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, Xin Li, Weijia Xu, Vetle I. Torvik, Yi Bu, Chongyan Chen, Islam Akef Ebeid, Daifeng Li, Ying Ding

[7] Hakala, K., Kaewphan, S., Salakoski, T. and Ginter, F. Syntactic analyses and named entity recognition for PubMed and PubMed Central—up-to-the-minute. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing 102–107, https://doi.org/10.18653/v1/W16-2913 (2016).