Assignment - 6

Name: Srilalitha Lakshmi Anusha Chebolu
Date: 03/12/25

Useful literature for our project: https://pmc.ncbi.nlm.nih.gov/articles/PMC10760044/

A Biomedical Knowledge Graph

The construction of **BioKG**, a large-scale biomedical knowledge graph, follows a structured multi-step process that integrates natural language processing (NLP), machine learning, and probabilistic reasoning. The lifecycle consists of the following key phases:

1. Data Acquisition and Preprocessing

The first step involves **collecting and curating heterogeneous biomedical data sources** to ensure comprehensive coverage. This includes:

- **Textual Data:** Over **34 million PubMed abstracts** were retrieved as the primary source of biomedical literature.
- **Structured Data:** Relation data from **40 publicly available biomedical databases** was incorporated to complement the extracted knowledge.
- **High-throughput Genomic Data:** Additional datasets were used to capture relationships inferred from large-scale biological studies.

To facilitate seamless integration, entity names and identifiers were normalized across multiple sources, reducing redundancy and ensuring consistency.

2. Information Extraction and Relation Identification

To transform unstructured textual content into a structured knowledge graph, the authors employed **state-of-the-art NLP techniques**:

- **Named Entity Recognition (NER):** Entities such as diseases, genes, chemical compounds, and species were extracted using deep learning-based NLP models.
- **Relation Extraction (RE):** Relationships between entities were identified and classified using a machine learning pipeline. This system was developed and refined based on their **LitCoin NLP Challenge-winning approach**, achieving human-level accuracy in extraction.

3. Entity Normalization and Integration

Once extracted, entities were **standardized and mapped** to existing biomedical ontologies to resolve variations in naming conventions. Relations extracted from literature were then **merged with those from public biomedical databases**, significantly expanding the coverage and accuracy of BioKG.

4. Causal Knowledge Graph Construction

Unlike traditional knowledge graphs that capture only direct relationships, BioKG was designed to **support causal inference**. This was achieved by:

- Annotating the **directionality of relationships**, indicating causal dependencies where applicable.
- Training a deep learning model to **predict the direction of relations**, transforming BioKG into a **Causal Knowledge Graph (CKG)** capable of **indirect inference**.

5. Probabilistic Semantic Reasoning (PSR) for Automated Knowledge Discovery

To enable automated hypothesis generation, the authors developed a **probabilistic reasoning algorithm** that allows the inference of relationships between entities that are not explicitly connected. This enables:

- **Drug Target Identification:** Predicting potential drug targets for diseases by analyzing indirect relationships.
- **Drug Repurposing:** Identifying novel uses for existing drugs based on inferred connections within the knowledge graph.

6. Evaluation and Validation

The accuracy and reliability of BioKG were assessed through multiple validation techniques:

- **Manual verification** of a subset of extracted relations to compare performance against human annotations.
- Calculation of **recall and observed positive rate (OPR)** for evaluating **drug repurposing predictions** and **target identification accuracy**.
- **Retrospective studies** demonstrating that BioKG could have predicted **40% of lung cancer drug targets** up to **15 years in advance**, highlighting its potential for accelerating biomedical discoveries.

7. Deployment and Accessibility

The final **BioKG system** was deployed as a **cloud-based platform** (https://www.biokde.com), providing researchers with **open access to structured biomedical knowledge**. This allows for:

- **Efficient query execution** to retrieve relationships between biomedical entities.
- **Integration with AI-powered discovery pipelines**, facilitating real-time hypothesis generation.

Conclusion

The **BioKG lifecycle** represents a **scalable, AI-driven approach** to biomedical knowledge representation. By leveraging state-of-the-art NLP, causal reasoning, and probabilistic inference, the system **bridges the gap between unstructured literature and structured databases**, making it a valuable tool for **accelerating scientific research** in drug discovery and biomedical knowledge discovery.

Resource for training a spaCY model: https://ner.pythonhumanities.com/intro.html

Resource for constructing a KG: link

Using GraphRAG: link

I would like to take up the task of training a spaCy model for our domain specific task and then give to the pretrained LLM to retrieve relation extractions or use REBEL.

Later compare this(Entity and relations) with the triples of an LLM model pretrained on pubmed abstracts.

Finally Construct the KG using LlamaIndex