Standard KG-Based Medical Knowledge Discovery Pipeline

Nghi Truong Xoi

Abstract:
This project idea is to demonstrate and implement the Standard KG-Based Medical Knowledge Discovery Pipeline. It focuses specifically on figure 1a of this Scorpius paper (Wang et al., 2024). My goal is to reproduce some key results and try to improve/modify the model of the disease-specific folder based on the given code that is provided in this research paper. These implementations show how important and relevant medical information is. Such as the knowledge graph, graphs that are being converted into vectors for analysis, and rankings of the drug for specific diseases. From this I was able to implement and produce rankings that make sense based on real-world medical knowledge. This gives an understanding of how a malicious paper can be inserted into the medical system making it seem like legit and how manipulation to the system works.

1. Introduction

One of the most useful tools in biomedicine is medical knowledge graphs. This supports medical-decision making because knowledge graphs help medical professionals with organizing biomedical information and to make reliable decisions. Knowledge graphs represent entities such as drugs, diseases, and genes. It shows the differences and the relation of computational biology reasoning. Although it is useful it can be extremely vulnerable because based on this research Scorpius paper by Wang et al. it can easily be attacked even if there is only a single misleading paper inserted into the system. It will boost up to 71.3% of the rankings for targeted drugs which manipulate drug rankings.

2. Background

In medicine knowledge graphs are crucial to have so that it can organize important medical information to connect networks of entities and relations (drugs, diseases, genes). Medical knowledge can be complex but these knowledge graphs help to structure it in a way to make it easier and much more organized. It is useful for decisions making and discovering new drugs

The idea of this project is to explore embedding techniques of the knowledge graphs. Entities and relations are being connected/map together into numeral vector representations for computational understanding. I had implemented the Dismult model, it represents each of the entities and the vector is being represented as relations. The model uses simple multiplication to show the interaction of the entities' connections.

Moving on into relation extraction. This extraction process is to find out the entity's relations by analyzing text. In biomedicine this is to recognize the biomedical patterns through text or sentences. So for example, “(A Type of Drug, to Treat, A Disease)” or “(A Gene that Causes a Specific Disease).”

3. Implementation

This implementation organizes biomedical knowledge as a triplet relation such as subject, relation, and object. It shows (Metformin, Treats, Diabetes). This is shown that Metformin is a drug to use to treat diabetes.

I also implemented a knowledge graph to extract the process of relation of texts. I had done this by using a simple method of looking for patterns in sentences with dependency parsing. What I did was I implemented it to show this: Dependency path: DRUG | nsubj | treats | dobj | DISEASE. This extract shows the relations of drugs and diseases.

The Dismult model was implemented so that it can convert entities and the relations into vectors of the mathematical shared vector space. This helps the model to compute how likely a vector relationship of the score is and this is done by multiplicative interactions.

For each of the given diseases, the drug is being ranked by the relevance scores which is calculated by the Dismult model. If it is a higher score this means that the model predicts a stronger connection of drugs and diseases.

T-SNE dimensionality reduction was used for a visualization of the embedding space. The colors entities are being represented by type which is drug, disease, and gene. I also added a bar chart to represent and visualize the drug rankings for specific diseases.

4. Results

By t-SNE visualization of the embedding space, the clustered type which is represented by drugs (green), diseases (orange), and genes (blue). This group together is a distinct region. This show that the embedding model did well on reflecting on telling the different types apart, and understanding the meanings

Since I specifically did diabetes, the testing model shows that it accurately ranked Metformin as the most relevant drug with the score of 0.0203. This result aligned well with medical knowledge since Metformin was really the most common first time use to treat Type 2 diabetes. Although this implementation was a small scale it was still able to provide a very meaningful and accurate results

5. Discussion

This standard pipeline was able to do well on organizing biomedical knowledge and showing the different kinds of relevant ranking. All of the implementations are being implemented individually which leaves room for improvement or to update. As being talked about in the Scorpius research paper, this pipeline represented that the medical system of the knowledge graphs can be extremely vulnerable to poisoning attack, where an attacker can manipulate rankings and give in a malicious paper that can prevent patients from getting the full proper treatment for the specific disease that they have. This implementation is to understand how these attacks work and the foundation of it.

There are limitations to my implementation. I am using a small-scale implementation to demonstrate examples of how figure 1a works. Having a full-scale implementation would provide a much more understanding to Scorpius. But doing so would require a very large data set and advanced techniques.

If I could improve this I would implement defensive mechanisms and try a much more advanced model. I would also try to use a larger biomedical dataset to improve this current overall implementation to better advanced the research

Conclusion

Overall, my project pipeline implements the Standard KG-Based Medical Knowledge Discovery of figure 1a of the Scorpius paper. I showed how the system can accurately rank drugs for a specific disease. This represents how the medical system of the knowledge graph would work. It gives insight into the normal functioning and the poisoning attacks of the knowledge graph. This helps provide an understanding of malicious inputs and how these graphs can easily be manipulated which can affect healthcare professionals with their decision making to treat a disease.

References:

Wang, S., Yang, J., Xu, H., Mirzoyan, S., Chen, T., Liu, Z., Liu, Z., Ju, W., Liu, L., Xiao, Z., Zhang, M. (2024). Poisoning medical knowledge using large language models. Nature Machine Intelligence, 6(10), 1156-1168. https://doi.org/10.1038/s42256-024-00899-3