Relation Extraction of Biomedical Text

Under Guidance of

- Professor Carlos Rojas

- Neeharika Yeluri (neeharika.yeluri@sjsu.edu)
- Pranav Chellagurki(pranav.chellagurki@sjsu.edu)
- Rahul Raghava Peela(rahulraghava.peela@sjsu.edu)
- Sai Prasanna Kumar Kumaru (saiprasannakumar.kumaru@sjsu.edu)

Computer Engineering Department

G45 Relation Extraction of Biomedical Text

Project Advisor: Carlos Rojas

Chellagurki, Pranav (MS Artificial Intelligence)
Kumaru , Sai Prasanna Kumar (MS Artificial Intelligence)
Peela, Rahul Raghava (MS Artificial Intelligence)
Yeluri, Neeharika (MS Artificial Intelligence)

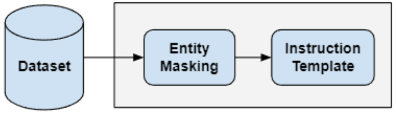
Introduction

In the face of over a million new life sciences and biomedical papers annually, extracting key information has become a necessity, making use of advanced NLP techniques. Recent progress in NLP, driven by transformer-based models like BERT and newer entries like GPT-4, Llama2, and Gemma, shows promise. However, these models often fall short with the specialized language of biomedical texts unless fine-tuned specifically on such content. Our research applies this fine-tuning approach, using smaller, open-source LLMs such as Gemma-2b, Gemma-7b, and Llama2-7b, which are manageable on our High-Performance Compute infrastructure. We focus on the relation extraction task in biomedical NLP, classifying relationships between entities like genes and diseases. Additionally, we explore the potential of knowledge graphs for discovering multi-hop relationships and enhancing biomedical research. Our findings include a comparative evaluation of these models against BioBERT using the GAD and EU-ADR datasets and the implications of knowledge graphs in identifying new entity relationships.

Methodology

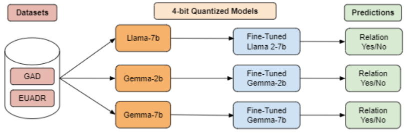
An overview of our approach to biomedical relation extraction using machine learning models, focusing on the datasets used for fine-tuning, our evaluation metrics, and experimental setup. We employ two benchmark datasets, the Gene Association Database (GAD) and the European Annotated drugs, diseases, targets, and their Relationships (EU-ADR), chosen for their common use in benchmarking biomedical NLP tasks. GAD documents gene-disease relationships with 53,300 relations, while EU-ADR focuses on drug-disease relations with 3,550 entries.

For data preprocessing, both datasets were standardized and underwent entity masking to align with model input requirements. This involved replacing specific gene and disease names with placeholders like "@GENE$" and "@DISEASE$".


Dataset
Entity Masking
Instruction Template

We used BioBERT, pre-trained on biomedical texts, and fine-tuned newer models like Llama2-7b and Gemma (in 2b and 7b sizes) on these datasets to improve their biomedical relation extraction capabilities. Due to computational limits, we trained on a subset of 20,000 GAD samples, ensuring a balanced representation of positive and negative relations and efficient use of computational resources.

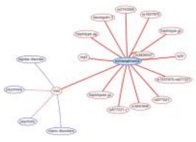
Our experimental setup included constructing a knowledge graph from the datasets to explore new relationship identification through multi-hop searches. This setup demonstrates the capability of our models to leverage extensive biomedical data for deeper insights into gene and disease interactions.

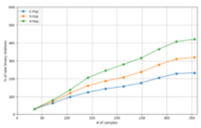

Datasets
4-bit Quantized Models
Predictions
GAD
EUADR
Llama-7b
Gemma-2b
Gemma-7b
Fine-Tuned Llama 2-7b
Fine-Tuned Gemma-2b
Fine-Tuned Gemma-7b
Relation Yes/No
Relation Yes/No
Relation Yes/No

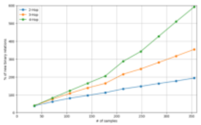
Knowledge Graph

We developed a knowledge graph using Python, Flask, and the vis-network JavaScript library to synthesize and visualize data from the GAD and EU-ADR datasets. Python handled data manipulation with pandas for data cleaning and structuring, while Flask facilitated web deployment, allowing dynamic graph rendering for interactive user engagement. The vis-network enabled real-time, interactive graph visualizations, linking entities like genetic markers and drug reactions with edges that represent their relationships.

This knowledge graph aids in exploring new relationships by visually representing connections and the absence thereof. It serves as a crucial tool for researchers and clinicians to intuitively explore and derive insights from complex biomedical data, enhancing decision-making in genetic research and drug safety.







Experiment 1: Binary Relation Extraction - Initial tests without fine-tuning showed poor model performance. After fine-tuning, BioBERT excelled particularly with the EU-ADR dataset, while Llama2-7b showed superior overall performance. The results confirmed the necessity of fine-tuning and suggested that combining different models could leverage their complementary strengths for better results.

Experiment 2: Knowledge Graph-Based Relation Extraction - We created interactive knowledge graphs using Python, Flask, and the vis-network library to visualize and analyze relationships from the GAD and EU-ADR datasets. Multi-hop searches within these graphs revealed significantly more relationships, demonstrating the value of deeper analytical approaches. For example, relationships discovered increased exponentially with the number of hops, especially notable in the GAD dataset.

These experiments utilized advanced computational setups, including NVIDIA GPUs and optimized training protocols like LoRA and 4-bit quantization, highlighting the effectiveness and scalability of our methods in biomedical text mining.

Analysis and Results

In Experiment 1, we initially tested our models un-fine-tuned on biomedical datasets, leading to poor performance. Post fine-tuning, BioBERT showed substantial improvement, outperforming Gemma-2b and underperforming against Llama2-7b across both GAD and EU-ADR datasets. Notably, BioBERT had a higher recall on the EU-ADR dataset, while larger models struggled, suggesting potential benefits from model ensembling. A vocabulary analysis revealed only 58.3% of EU-ADR's terms were present in the pre-trained model vocabularies, compared to 66% for GAD, explaining some dataset-specific challenges.

In Experiment 2, using a knowledge graph for relation extraction on the EU-ADR dataset revealed that multi-hop searches could identify significantly more relations, with numbers increasing up to four times as the hops increased from two to four.

Datasets	Metric	BioBERT	G-2b	G-7b	LlaMA2-7b
GAD	P	78.83	72.02	**97.86**	**99.78**
	R	78.18	70.00	**97.94**	**99.80**
	F1	78.27	69.76	**97.89**	**99.79**
EU-ADR	P	67.87	**93.11**	**98.33**	89.44
	R	70.14	65.35	63.35	**94.88**
	F1	68.73	74.39	74.39	**93.18**

The graph's connectivity and the potential for discovering new relations also grew substantially, indicating the effectiveness of multi-hop searches, especially in the GAD dataset. Despite computational challenges due to the vast number of relations in GAD, a controlled number allowed for meaningful comparisons.

Summary/Conclusions

In this study, we compared the performance of BioBERT with fine-tuned LLMs (Gemma-2b, Gemma-7b, Llama2-7b) on GAD and EU-ADR datasets. Llama2-7b showed superior performance in the GAD dataset, while Gemma-7b led in precision for EU-ADR but with lower recall. Using a knowledge graph, we demonstrated that multi-hop searches enhance the discovery of new relationships, indicating the potential of these models to reveal deeper biomedical connections.

Key References

1. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining" .

2. K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database" .

3. E. M. van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors, and L. I. Furlong, "The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships" .

4. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" .

5. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt .

Acknowledgements

The authors are deeply indebted to Professor Carlos Rojas and Professor Jorjeta Jetcheva for their invaluable assistance in the preparation of this study and Dr. Dan Harkey for his guidance throughout the course