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)

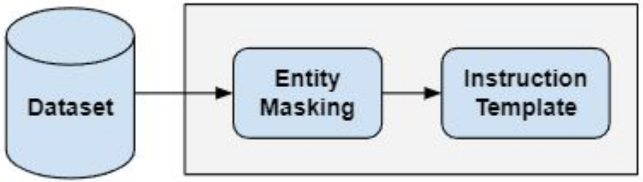
Introduction

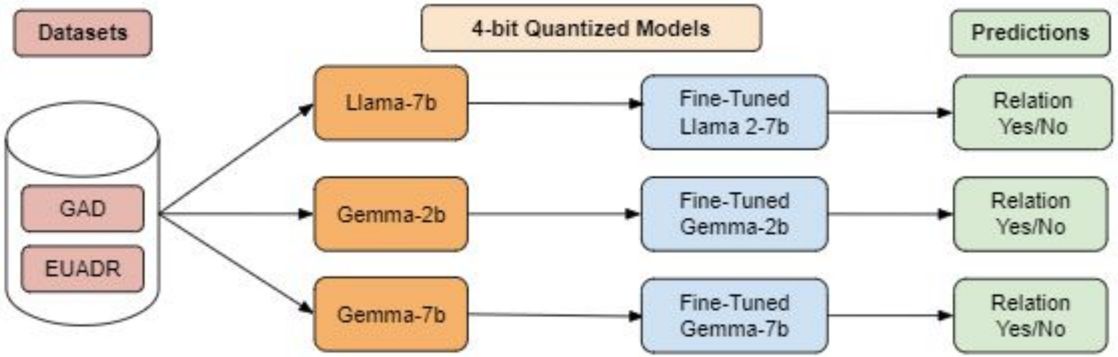
- Exploring the rapid increase in biomedical research papers; over a million new papers published annually.
- Emphasizes the need for advanced tools to extract and synthesize information from the vast amount of data.
- Introduces the use of Large Language Models (LLMs) and Knowledge Graphs as innovative solutions for biomedical Natural Language Processing (NLP).
- Aims to enhance biomedical NLP capabilities, making it possible to extract significant relationships from texts and discover insights into gene-disease interactions and drug efficacy.

Related Work

- Reviews early statistical NLP models leading up to advanced deep learning approaches including transformers and LLMs.
- Highlights the specific evolution of biomedical NLP, noting the adaptation of general NLP tools for biomedical applications, such as BioBERT.
- Discusses previous challenges in the field, particularly the lack of effective tools that integrate both high-level computational models and scalable knowledge architectures like knowledge graphs.
- Points out that while general-purpose LLMs are increasingly used, their application in specialized fields like biomedicine remains complex due to unique vocabulary and contextual demands.
- Concludes with the observation that recent works have begun to bridge these gaps through specialized datasets and model tuning but more focused efforts are needed.

Methodology





Datasets and Processing

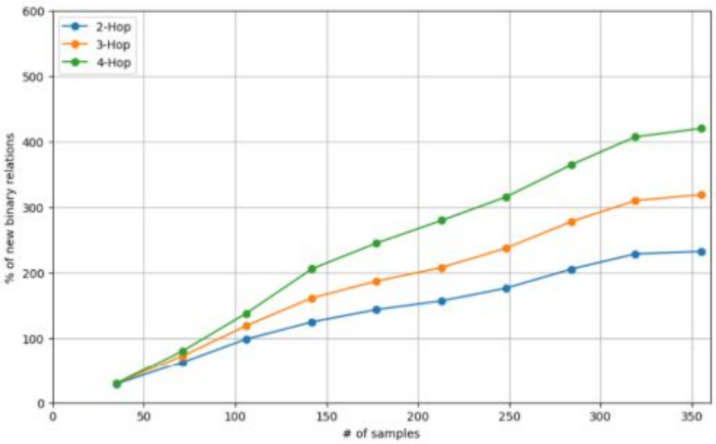
- GAD Dataset: 53,300 relations describing gene-disease associations, labeled as positive (1) or negative (0) relations.
    a. GAD: "Mutations in the BRCA1 gene can cause breast cancer" labeled as 1.
- EU-ADR Dataset: 3,550 relations focused on drug, disorder, and gene targets, similarly labeled for positive or negative relationships.
    a. EU-ADR: "LRP5 genetic variants as possible susceptibility factors for osteoporosis" labeled as 1.
- Structured input template with masked entities and binary relation indicators to reduce ambiguity and enhance learning consistency.

Experiments and Results

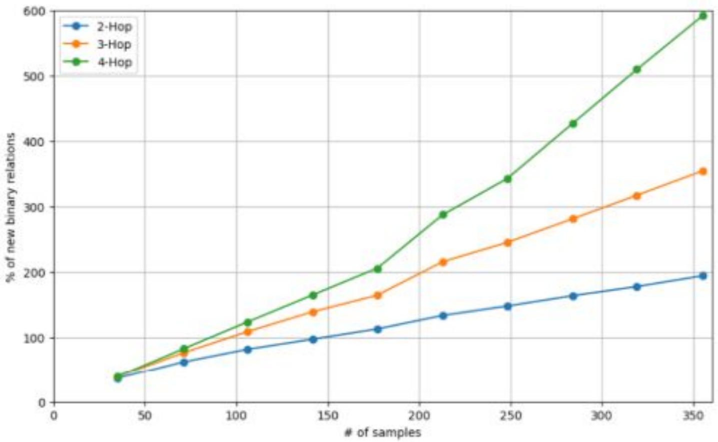
Datasets	Metric	BioBERT	G-2b	G-7b	Llama-7b
GAD	P	52.84	24.03	47.10	27.47
	R	50.11	40.58	49.80	49.37
	F1	33.23	30.00	35.62	34.20
EU-ADR	P	40.73	34.65	36.86	36.56
	R	42.57	40.07	49.80	48.28
	F1	29.22	37.16	42.37	41.61

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**

Knowledge Graph Insights

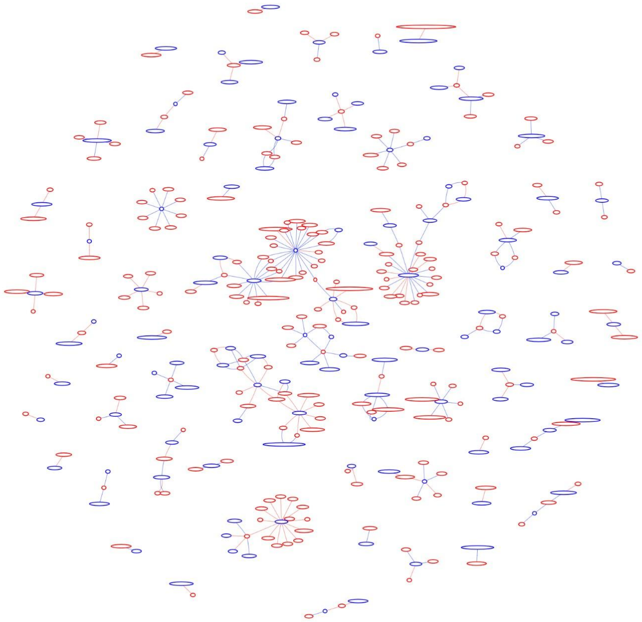


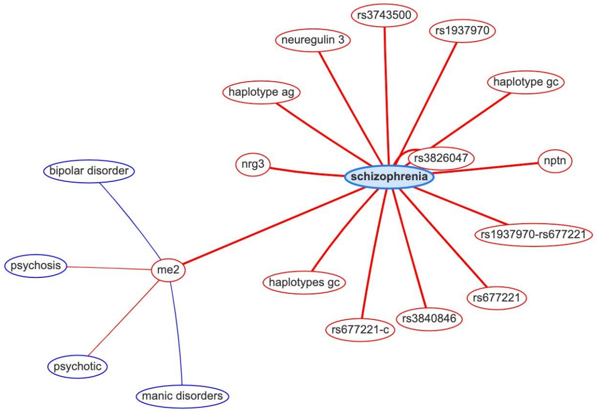
(a) % of new relations retrieved from KG for EU-ADR



(b) % of new relations retrieved from KG for GAD

Knowledge graph and Multi-hop connections





Challenges and Future Work

1. Model Size vs. Tuning Depth:
   - Smaller models offer computational efficiency and faster training, ideal for limited-resource settings.
   - Larger models have superior reasoning abilities due to extensive neural architectures but often lack comprehensive fine-tuning.
2. Challenges with Large Models:
   - Inadequate handling of specialized vocabularies due to partial fine-tuning, a significant issue in fields like biomedicine where terminology evolves quickly.
3. Future Focus on Fine-Tuning:
   - Comprehensive fine-tuning of larger models, especially their tokenizer components, to enhance their ability to process complex datasets fully.
4. Innovative Solutions to Reduce Computational Load:
   - Develop more efficient algorithms and explore advanced hardware solutions to manage and reduce the computational demands of large models.
5. Adaptive Tokenization Techniques:
   - Implement adaptive tokenization that adjusts to the specialized vocabularies of different domains, ensuring large models maintain effectiveness across varied datasets.

THANK YOU