TEMPHI: Text Mined Pathogen Host Interactions

Yavuz Samet Topcuoglu, Sinan Kerem Gündüz - Advisor: Arzucan Özgür

Bogazici University - Computer Engineering - CMPE492 Project

Problem Statement

- **Extracting relations** from biological papers is a time-consuming and labor-intensive task.
- Manual extraction hinders **efficient access** to valuable information on **protein-protein interactions**.
- Lack of **automated tools** limits progress in understanding pathogen-host interactions.
- Researchers require a reliable and efficient solution for **automated relation extraction**.

Motivation

- Accelerate academic research and advancements in understanding and combating infectious diseases.
- Develop an **end-to-end** automated tool for relation extraction from biological papers.
- Create a **comprehensive dataset** for protein-protein interactions from PubMed documents.
- Provide a user-friendly front-end for easy access to relation data.
- Enable researchers to search for relations associated with specific proteins.

Datasets

The **availability** of high-quality datasets plays a crucial role in the success of relation extraction tasks, especially when it comes to fine-tuning pre-trained models, as their performance tends to be suboptimal without sufficient training on domain-specific data.

1. **RelX Dataset from PHISTO**

   PHISTO is a dataset containing 62340 Pathogen-Host Protein Interactions from 2317 PubMed documents.

Pathogen	Taxonomy ID	Uniprot ID	Pathogen Protein	Uniprot ID	Human Protein	Experimental Method	PubMed ID

Table 1. PHISTO Dataset Column Names

Using the same documents used in PHISTO, we have created RelX Dataset. Labels in the dataset are assigned according to existence of protein pairs in the PHISTO dataset.

PubMed ID	Sentence	Entity 1 Name	Entity 1 Uniprot ID	Entity 2 Name	Entity 2 Uniprot ID	Label

Table 2. RelX Dataset Column Names

2. **PPI Dataset from BioRED**

   Biomedical Relation Extraction Dataset is formed by merging different datasets such as:
   - **AIMed**: Sentence level relation extraction dataset containing 1101 relations.
   - **BioInfer**: Sentence level relation extraction dataset containing 2662 relations.

   Combined dataset has 16263 relations which 12173 of them labeled as 0 and 4090 of them as 1.

Sentence	Label

Table 3. Altered BioRED Dataset Column Names

Pipeline

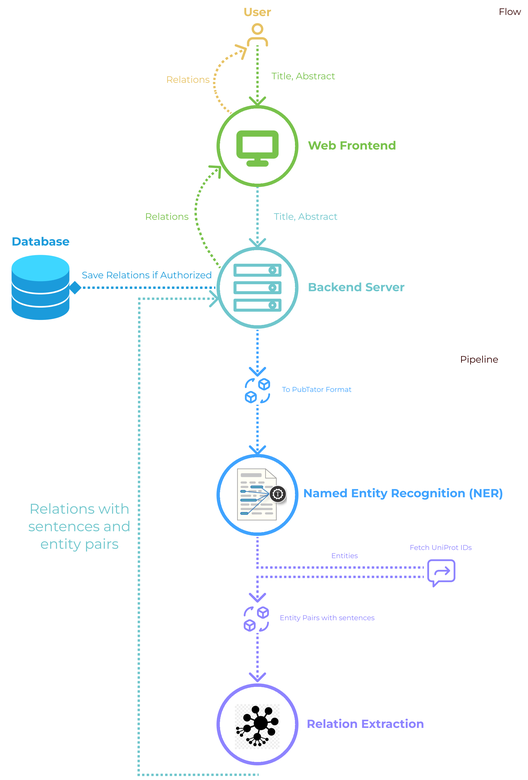

Figure 1. Flow and pipeline.

Comparison of Models

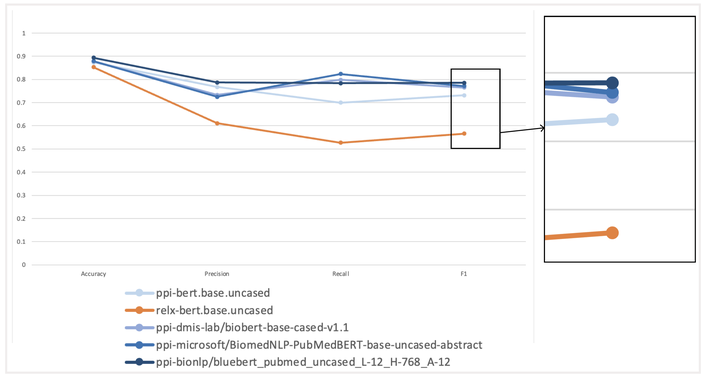

Figure 2. Model comparison on four different metrics.

Data Generation

Document Scraping

After designing and gathering the pipeline together, and determining which model to continue with, we have scraped lots of documents from PubMed on different topics using various queries such as the followings:

- protein targets for cancer therapy
- coronavirus spike protein
- proteomics of breast cancer
- proteins implicated in stroke pathophysiology
- protein-protein networks in kidney disease and many more...

Passing through the Pipeline

Scraped documents are given to pipeline and the results are stored in a database to be shown on the website.

Result

In conclusion, our project, TEMPHI: Text Mined Pathogen Host Interactions, has successfully addressed the challenges associated with relation extraction from biological papers. Through the development of our automated tool, we have achieved the following key outcomes:

- **Created a comprehensive dataset** of protein-protein interactions by extracting entity pairs from a large collection of PubMed documents.
- Developed a user-friendly front-end that allows researchers to **search for relations** associated with proteins using titles and abstracts or protein IDs.
- Utilized Named Entity Recognition (AIONER) and fine-tuned the BERT models with domain-specific labeled PPI data to **achieve improved performance** in relation extraction.
- Demonstrated **the importance of high-quality datasets** for fine-tuning pre-trained models, significantly enhancing their effectiveness.

Future Work and Possible Improvements

- Expand the training dataset further.
- Explore advanced ML models.
- Create a recurring task to retrieve relations from **newly added documents** to PubMed.

Acknowledgement

We wish to thank our advisor Arzucan Özgür for the guidance and Berke Kavak for helping us tackle the Named Entity Recognition problem as well as the support given by both of them.

References

[1] GitHub - hushee69/biobert-relation-extraction: Relation Extraction using BERT and BioBERT - using BERT, we achieved new state of the art results — github.com.
`https://github.com/hushee69/biobert-relation-extraction`.

[2] Abdullatif Köksal, Hilal Dönmez, Rıza Özçelik, Elif Ozkirimli, and Arzucan Özgür.
Vapur: A search engine to find related protein-compound pairs in covid-19 literature.
In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[3] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu.
BioRED: a rich biomedical relation extraction dataset.
Briefings in Bioinformatics, 23(5), 07 2022.
bbac282.

[4] Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu.
AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning.
Bioinformatics, 39(5), may 2023.