TEMPHI

Text Mined Pathogen-Host Interactions

CMPE492 Graduation Project, Spring 2023

Bogazici University - Department of Computer Engineering

Yavuz Samet Topcuoglu & Sinan Kerem Gündüz

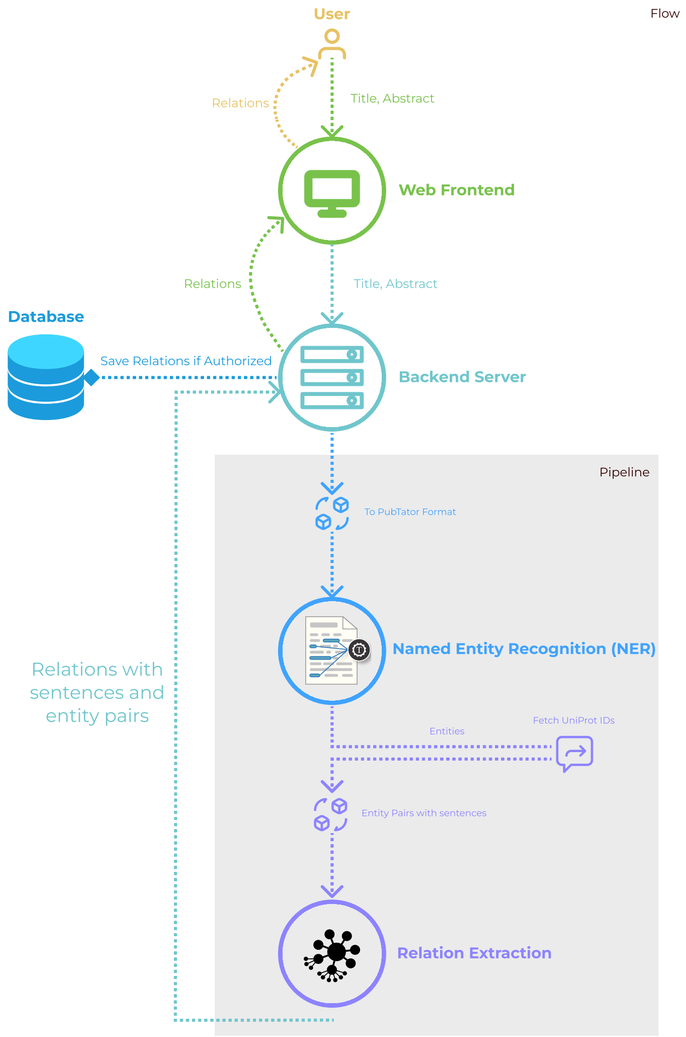
06.06.2023

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.

Technical Details

Pipeline



Technical Details

Datasets

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.

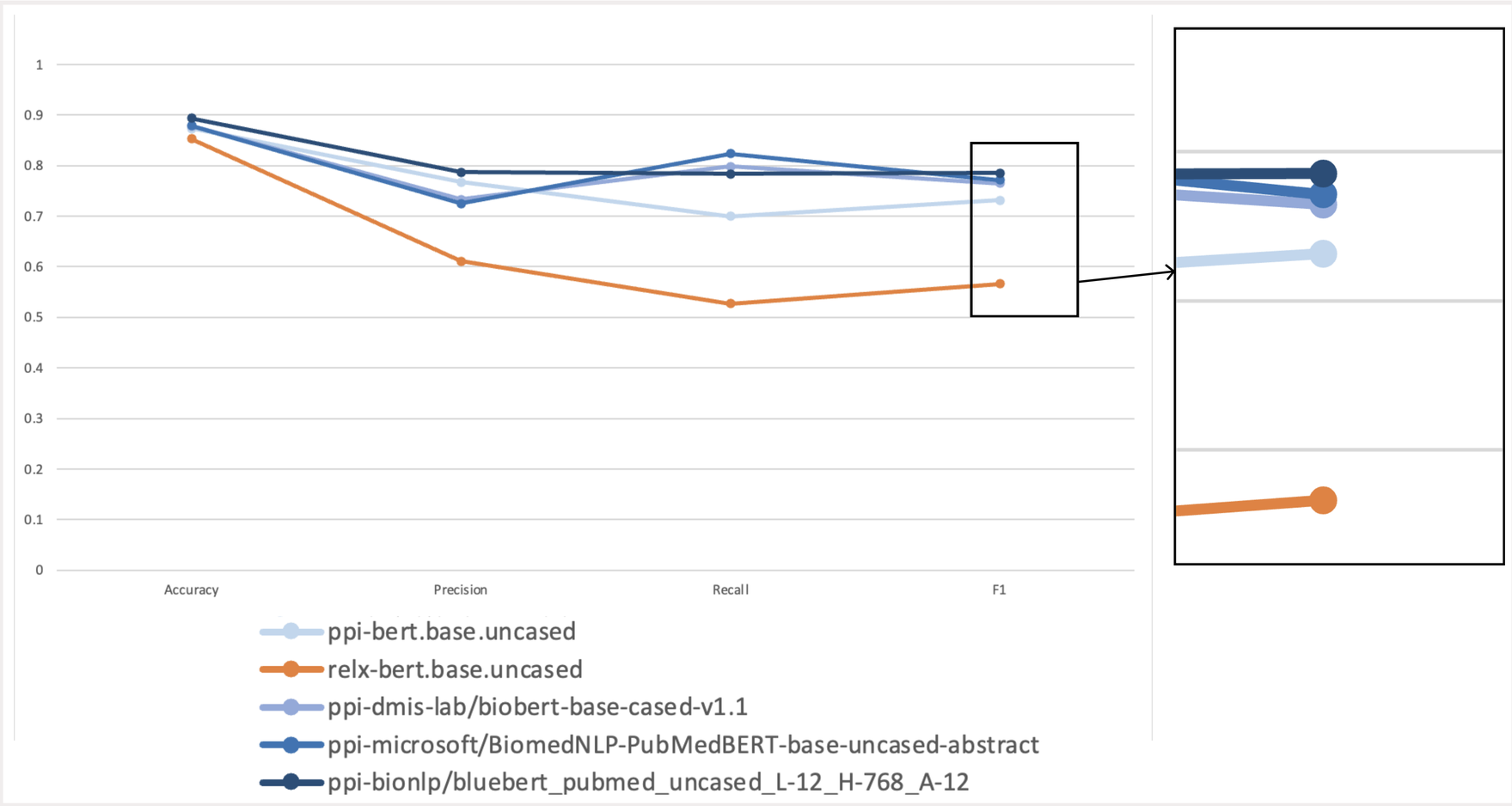
   Combined dataset has 16263 relations which 12173 of them labeled as 0 and 4090 of them as 1.

Sentence	Label

Table 3. Altered BioRED Dataset Column Names

Technical Details

Method Comparison



Challenges

- Finding high quality of labeled domain-specific data.
- Long time needed to fine-tune the pre-trained model unless you have GPU supporting device.
- Problems with deployment due to incompatibility in differences between operating systems and library versions.

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