Biomedical Information Extraction For Disease Gene Prioritization

Jupinder Parmar[1,2,], William Koehler[2,*], Martin Bringmann[2], Katharina Sophia Volz[2], Berk Kapicioglu[2,*]*

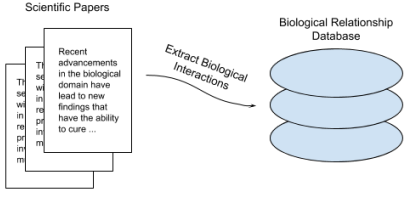
[1]Stanford University, jsparmar@stanford.edu

[2]OccamzRazor, {william, martin, volz, berk}@occamzrazor.com

arXiv link

Biomedical Information Extraction (IE)

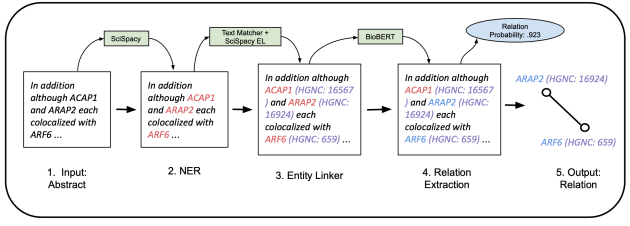
- Understanding diseases and developing curative therapies requires **extracting** and **synthesizing** relevant knowledge from vast swaths of biomedical information.



Our Contribution

- We built an end-to-end biomedical IE pipeline that **outperforms** existing state of the art for biomedical IE.
- Ran our pipeline over the PubMed corpus to extract protein-protein interactions (PPIs).
- Augmented an existing biomedical knowledge graph, DisGeNet, that already contains PPIs from STRING with our extracted PPIs and demonstrate that the augmentation yields a **20% relative increase** in hit@30 for predicting novel disease-gene associations.

Information Extraction Pipeline



- Augmented leading NLP models to achieve better performance for the biomedical domain.

System	Precision	Recall	F1
v1	**43.24**	45.71	44.44
v2	41.17	50.00	**45.16**
v3	31.37	68.57	43.04
Masked BioBERT	29.87	**70.00**	41.88

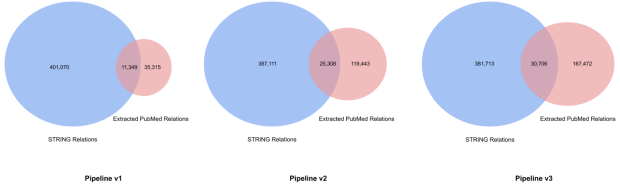
RE Results

System	Precision	Recall	F1
Our Model	**78.41**	**73.87**	**76.08**
PubTator	58.96	49.20	45.76
ScispaCy	37.81	57.96	53.64

NER Results

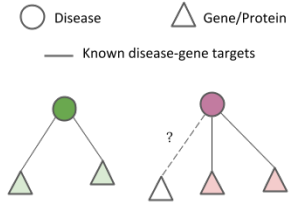
Extracting Relations from PubMed

- Ran three versions of the pipeline over 10 million PubMed abstracts that each extract **more PPIs** than previous information extraction attempts.



Disease Gene Prioritization

- Despite DisGeNet containing PPIs from STRING, a structured database, our extracted relations **boost performance** in the task of disease gene identification.



	MR	MP	hit@30	hit@3	hit@1
IE v3 + STRING + DisGeNET	**1418.397**	**92.484**	**37.367%**	**15.302%**	**7.829%**
IE v2 + STRING + DisGeNET	1441.802	92.262	35.409%	14.057%	7.473%
IE v1 + STRING + DisGeNET	1829.548	89.869	32.74%	13.701%	6.762%
STRING + DisGeNET	1952.084	89.362	31.139%	13.879%	7.651%
DisGeNET	7422.117	59.544	0.356%	0.178%	0.178%

Discussion

- Our pipeline not only is able to identify a large amount of PPIs, but these relations are **high quality** as they improve performance on a downstream task.

Ongoing and Future Work

- Train pipeline to extract additional biomedical relationships.
- Utilize extracted PPIs in additional tasks.