DrugProt: Drug-Protein Relation Extraction Using Transformer-Based Models

Volkan Bulca, İbrahim Özgürcan Öztaş

Instructor: Assoc. Prof. Arzucan Özgür

Computer Engineering, Boğaziçi University

Introduction and Motivation

- Our goal is to express relations between drugs and proteins in an easily accessible, reshapable structures.
- Our motivation is to extract relations between drugs and proteins to:
  - Discover new drugs by using the existing ones
  - Extract information about diseases
  - Reorganize drugs for different purposes

Related Work

- Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature
  Vapur is empowered with a relation-oriented inverted index that is able to retrieve and group studies for a query biomolecule with respect to its related entities. [1],
- BioBERT: a pre-trained biomedical language representation model for biomedical text mining, which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. With almost the same architecture across tasks, BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora.[2].

Methods

- Preprocess Stage
  - Splitting abstract sentences using GENIA Sentence Splitter
  - Entity pair construction by finding entity pairs in each sentence and expanding the sentences by that number
  - Mark each entity, sentence start and end locations with special tags: Mark each Drug as <e1>*Drug*</e1> and each Protein as <e2>*Protein*</e2>
  - Enumerate each relation type with numbers between 0-12
- Relation Extraction Stage
  - Tokenization of each abstract sentence marked above
  - Pre-trained transformer-based BioBERT model: "dmis-lab/biobert-v1.1"
  - Linear Layer and softmax activation
  - Prediction of the relation labels
- Training Hyperparameters
  - **Loss Function:** Multiclass Cross-Entropy Loss
  - **Learning Rate:** 1e-5
  - **Weight Decay:** 3e-2
  - **Optimizer:** Adam

Contact Information

- **volkan.bulca98@gmail.com**
- **oztas.ozgurcan@gmail.com**

Dataset

#Abstracts	#Entities	#Relations
3500	89529	17288

Table: Dataset Statistics

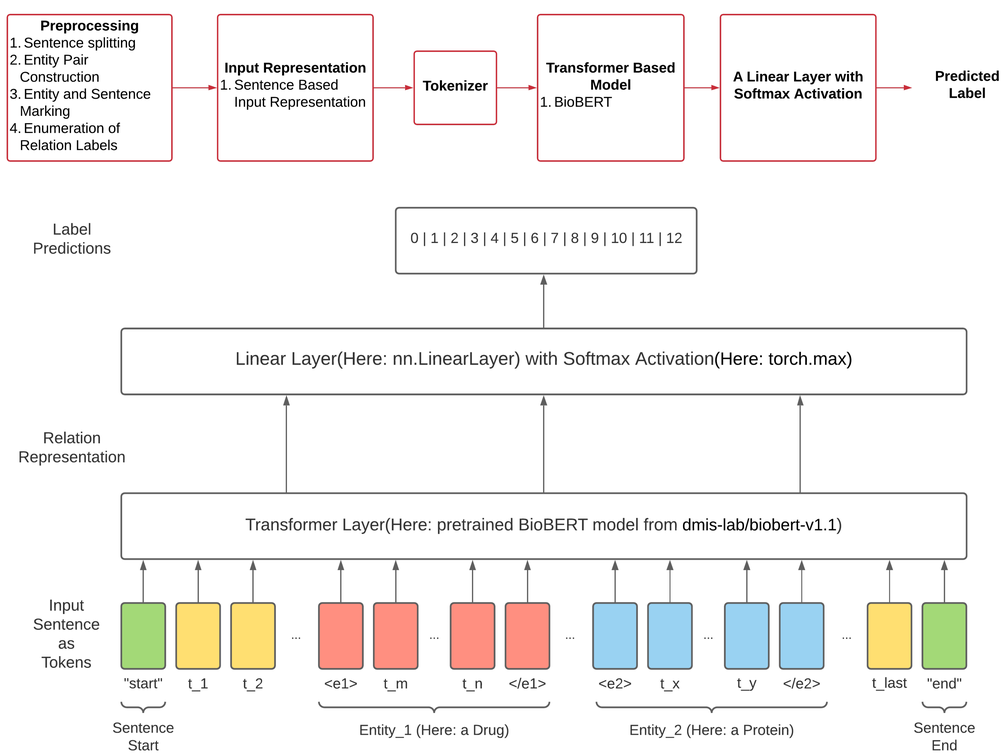
	#Sentence Count
Train Set	13819
Test Set	3455

Table: Train & Test Set Statistics

	#Relation	#Count	#Relation	#Count	#Relation	#Count
Train	INHIBITOR	4371	PART-OF	738	SUBSTRATE	1587
	ACTIVATOR	1185	IDOWNREGULATOR	1064	ANTAGONIST	768
	IUPREGULATOR	1135	AGONIST	515	DREGULATOR	1746
	PRODUCT-OF	775	AGACTIVATOR	22	AGINHIBITOR	12
			SPRODUCTOF	15		
Test	INHIBITOR	1121	PART-OF	148	SUBSTRATE	416
	ACTIVATOR	244	IDOWNREGULATOR	266	ANTAGONIST	204
	IUPREGULATOR	244	AGONIST	144	DREGULATOR	504
	PRODUCT-OF	146	AGACTIVATOR	7	AGINHIBITOR	1
			SPRODUCTOF	10		

Table: Relation Group Input Sentence Statistics

Model Pipeline



Results

Evaluation Results on 0.2-ratio test set generated by given training dataset: **3455** total predicted labels

- **0.77** overall accuracy
- **0.76** weighted precision
- **0.77** weighted recall
- **0.77** weighted f1-score

Discussions & Conclusion

- Transformer based models are highly dependent on random token guessing, thus running the model 10 times with the same hyperparameters and averaging the results to lessen the effect of randomization is necessary.
- Using more generalized transformer based models would result in a lesser accuracy than our score due to the precision in the pre-training of the models. Generalized models are trained on vast data sets with different topics, thus its power to focus on one topic is highly decreased.
- Biomedical text mining highly depends on pre-trained models with high-precision fine tuning and our project shows one instance of such an occasion that BioBERT can be utilized and provides fairly strong results.

Future Work

- Using different transformer-based pretrained models such as SciBERT, which is also a suitable for Relation Extraction tasks in scientific fields
- Consulting different techniques for representing clustered relations such as graph networks
- Training and testing our model with different hyperparameters to be able to eliminate the bias, deviations and randomness in the pre-trained model we used
- We will continue working on this project for the BioCreative VII competition and seek to improve our results by trying the aforementioned suggestions of improvements.

References

[1] 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*, Online, December 2020. Association for Computational Linguistics.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.
Biobert: a pre-trained biomedical language representation model for biomedical text mining.
Bioinformatics, Sep 2019.

Acknowledgement

We'd like to thank **Hilal Dönmez** for helpful discussions.