Leveraging Deep Learning Techniques for Biomedical Entity Extraction

Caroline He, Kunxiao Gao, Safiya Alavi, Sammy Suliman, Yujie Li

{c_x_he, kunxiaogao, safiyazalavi, shsuliman, yujie_li}@ucsb.edu

Abstract

In this project, we constructed and fine tuned a **NLP model** utilizing a BERT (Bidirectional Encoder Representations from Transformers) based machine learning model trained on a dataset called BioRED. Our project essentially benchmarks the performance of the NER task using this specific type of model on free text coming from published articles and research papers on FierceBiotech.com. So far in our project, we have gotten through the initial task of NER; although, moving forward, we can work towards the competition of an algorithm which produces a **Knowledge Graph** from free text by additionally completing the NEL task and RE task. In this report, we will discuss more about key details and concepts within our project, our data source, methodologies, and final results.

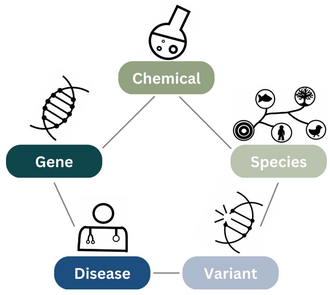


FIGURE 1. Visualization of the labels assigned to specific biological entities in the BioRED dataset. Distinct entities (words) are identified as biologically important and categorized into specific labels.

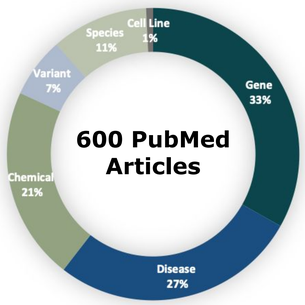


FIGURE 2. A Pie chart depicting the proportion of labeled tokens by entity type in the BioRed training dataset.

Definitions

Knowledge Graph (KG): a network model that links together named entities ('nodes') based on some common relation ('edge')

BERT (Bidirectional Encoder Representations from Transformers): a state-of-the-art Transformer-based language model most prominently used by Google to process search inquiries. Distinguished from other Transformer models by weighing the context on either side of each word in the input to better infer the meaning of each specific word in the input.

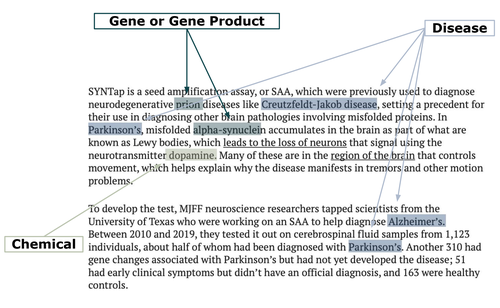


FIGURE 4. Visualization of NER tasks being performed on a news article about a new Parkinson's Test. Biological entities are identified and labeled.

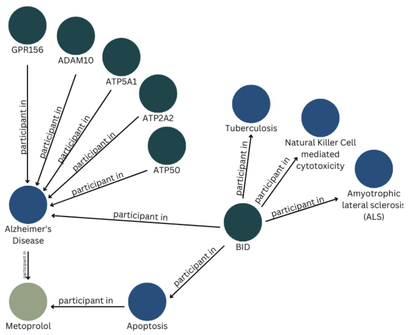


FIGURE 3. Example of a Knowledge Graph which links biological entities, colorized by label, to each other.

Named Entity Recognition (NER): training a model to parse unstructured text and classify named entities of a particular class.

Transformer: NLP model distinguished by its use of "self-attention", in which the model is able to selectively focus on individual parts of its input which it deems to be most important and uses these parts of the input to determine overall context.

Data

The entity-recognition model is trained using a pre-labeled dataset from the **BioRED** paper (Luo et al). This dataset contains:

BioRed Dataset

- 600 PubMed abstracts
- 6 entity categories (Gene, Disease, Sequence Variant, Chemical, Species, Cell Line)
- Multiple entity pairs along with their linkage

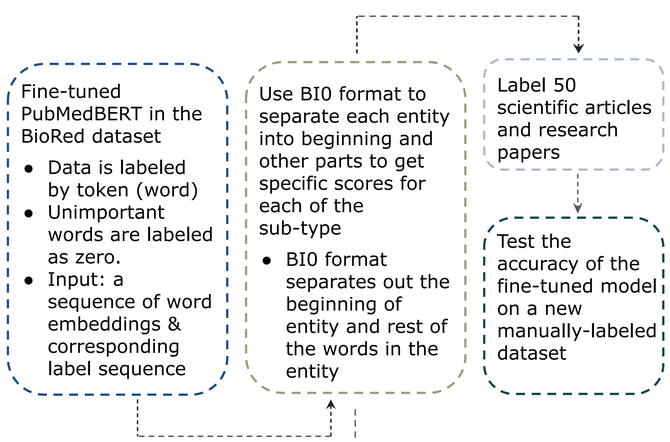
Our group manually labeled 50 scientific works (both research paper abstracts and scientific articles) to test the accuracy of our pre-trained model:

Articles & Abstracts Dataset

- 25 research abstracts
- 25 science communication articles
- 6 entity categories (Gene, Disease, Sequence Variant, Chemical, Species, Cell Line)

Methodology

To obtain an entity-recognition model, we trained PubMedBERT on the BioRed dataset. To test how applicable the model is, we labeled a novel dataset filled with abstracts and news articles and tested our model on it.



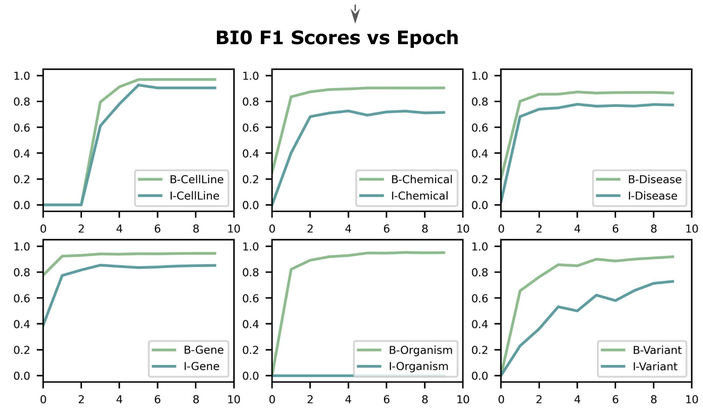


FIGURE 6: Displays the F1 scores of each entity type separated by the BI0 formatting versus epoch number in our model FIGURE 6: Displays the F1 scores of each entity type separated by the BI0 formatting versus epoch number in our model

Results

The overall **F1 score** associated with our model's performance on both the abstract and news article dataset is **40%**. In order, all the entity types had the following corresponding F1 scores:

- None (96%), B-Gene (73%), I-Gene (73%), B-Disease (64%), I-Disease (68%), B-Chemical (46%), I-Chemical (61%), B-Organism (0%), I-Organism (0%), B-SequenceVariant (6.5%), I-SequenceVariant (5.1%), B-CellLine (14%), I-CellLine (19%).

	TRUE	FALSE
POSITIVE	Tokens 524	Tokens 478
NEGATIVE	Tokens 5264	Tokens 138

PRECISION: 52.3% | RECALL: 79.2%

FIGURE 6. A confusion matrix depicting the raw number of true positives, true negatives, false positives, and false negatives from all of the manually annotated paper abstracts.

	TRUE	FALSE
POSITIVE	Tokens 161	Tokens 97
NEGATIVE	Tokens 1999	Tokens 51

PRECISION: 62.4% | RECALL: 75.9%

FIGURE 7. A confusion matrix depicting the raw number of true positives, true negatives, false positives, and false negatives from all of the manually annotated news articles

Model Results On 10 Tokens Of An Article In Our Dataset

Tokens	Predictions	True Labels
'to'	'None'	'None'
'diagnose'	'None'	'None'
'neurodegenerative'	'BBDiseaseOrPhenotypicFeature'	'None'
'prion'	'IIDiseaseOrPhenotypicFeature'	'None'
'diseases'	'IIDiseaseOrPhenotypicFeature'	'None'
'like'	'None'	'None'
'cre'	'BBDiseaseOrPhenotypicFeature'	'BBDiseaseOrPhenotypicFeature'
'##utz'	'IIDiseaseOrPhenotypicFeature'	'IIDiseaseOrPhenotypicFeature'
'##feld'	'IIDiseaseOrPhenotypicFeature'	'IIDiseaseOrPhenotypicFeature'
'##t'	'IIDiseaseOrPhenotypicFeature'	'IIDiseaseOrPhenotypicFeature'

FIGURE 8: Compares the model's predictions versus our team's true labeling for a random 15 tokens from an article in our manually-labeled dataset.

Conclusion

The model performed equally well in identifying entities in news articles and research objects. However, the overall F1 score associated with our model's performance on both the abstract and news article dataset (40%) is lower than the F1 scores associated with the model's performance on the BioRed dataset. We believe that this is because our group defined entity labels differently than the researchers who labeled the BioRed dataset. Specifically, as shown in Fig. 8, the issue is likely that our group had more limited definitions for what we considered to be different types of entities. As a result, entities that were technically labeled correctly by our model according to the BioRed definition were subsequently identified as 'false positives' in our dataset. It appears that most of these false positives are related to less common labels such as Organism and Sequence Variant.

References

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, Zhiyong Lu, BioRED: a rich biomedical relation extraction dataset, Briefings in Bioinformatics, Volume 23, Issue 5, September 2022, bbac282, https://doi.org/10.1093/bib/bbac282

Acknowledgements

Thank you to our Amgen sponsors Maxim Ivanov and Bonnie Jin, as well as our instructor Trevor Ruiz, for their help during the course of the project.