Drug-Disease Relation Extraction with BERT(s)

Kevin Lu
UC Berkeley MIDS
kklu78@berkeley.edu

Spencer Song
UC Berkeley MIDS
spencersong@berkeley.edu

Chetan Munugala
UC Berkeley MIDS
c.munugala@berkeley.edu

December 4, 2021

Abstract

With the advent of BERT and its domain-specific successors, state-of-the-art standards for relation extraction tasks have greatly improved over the last few years. In the case of the drug-treatment relation extraction task, recent papers have applied semantic annotation and complex feature engineering. In an effort to make the process less complex, we fine-tuned various domain specific pre-trained transformer models on a labeled drug-treatment dataset, and applied a combination of techniques to combat imbalanced data. Although we were unable to raise the overall state-of-the-art F1 scores, we were able to offer solutions that required far less preprocessing with a minimal tradeoff in F1 score.

1 Introduction

Relation extraction between medical concepts is a challenging yet important task, as these relations (protein-protein, gene-protein, drug-drug, etc.) are valuable to scientists and health professionals. Additionally, this information can aid common civilians who want to learn from dense biomedical literature. Typically, the complexity of vocabulary, as well as the high variance in sentence structure in biomedical text has led to poor model performance.

Fortunately, biomedical literature exists through massive free, online databases which can be leveraged to create powerful machine learning datasets. For example, the Medline/PubMed baseline database alone contained over 24.3M records of data (2016). By combining these datasets and NLP techniques, we can make strides towards making biomedical relation extraction more feasible.

We found that gene-disease and protein-protein relation extraction had been more extensively studied as compared to drug-disease relations[23]. For this reason, we decided to pursue drug-disease relations, specifically by using pre-trained transformers to circumvent complex feature engineering and other preprocessing steps that require specific domain knowledge. Our goal for this work is to improve drug-disease relation extraction models using pre-trained transformers, while simultaneously creating a solution that requires less domain knowledge. Concretely, we want to achieve a similar F1 score to the model presented in "Drug Disease Relation Extraction from Biomedical Literature Using NLP and Machine Learning"[4] with a more streamlined approach and minimal performance tradeoffs.

2 Background

Drug-Disease Relation Extraction:

We used Rosario and Hearst's[1] work to guide our setup for multi-classification relation extraction. They proposed a method that distinguishes seven relations between the semantic entities "treatment" and "disease." They compare five graphical models to a neural network built on feature engineering and present three relations and their accuracies (Cure: 92.6, prevent: 38.5, and side effect: 20), showing that the latter model leads to higher accuracies.

Muzaffar et al.[2] used a hybrid feature set Unified Medical Language System (UMLS) and a ranking algorithm. They then used an SVM and Naive Bayes to classify the relations on Rosario and Heart's categorized relations. The following F-scores were reported: Cure, 98.05; Prevent, 93.55; and Side Effect, 88.89.

Karaa et al.[4] use NLP and ML to approach the labeled MEDLINE 2001 dataset. They identify specific features (lexical, semantic, morphological, etc.), use UMLS ontology for semantic annotation, and use an SVM to classify drug-disease relations. The results confirmed that the features related to medical concepts were relevant in classifying conceptual relationships. The following F-scores were reported: Cure, 98.19; Prevent, 85.71; Side effect, 79.37.

The above sources are approaching the same dataset and task that we are but choose to use feature engineering over transformer / LSTM models. On the other hand, the sources below focus on the broader biomedical relation extraction task with these modern NLP techniques.

Biomedical / BERT Relation Extraction: Su et al.[12] fine-tune biomedical BERT models for relation extraction by using the entire layer as opposed to partial knowledge from the last layer. Typically, a BERT model for classification problem is based on the [CLS] token of the last layer, dropping the other outputs. They added a new module to summarize last layer outputs and concatenated this summarized information with the [CLS] output. They found improvements in the protein-protein task, but not drug-disease or ChemProt tasks.

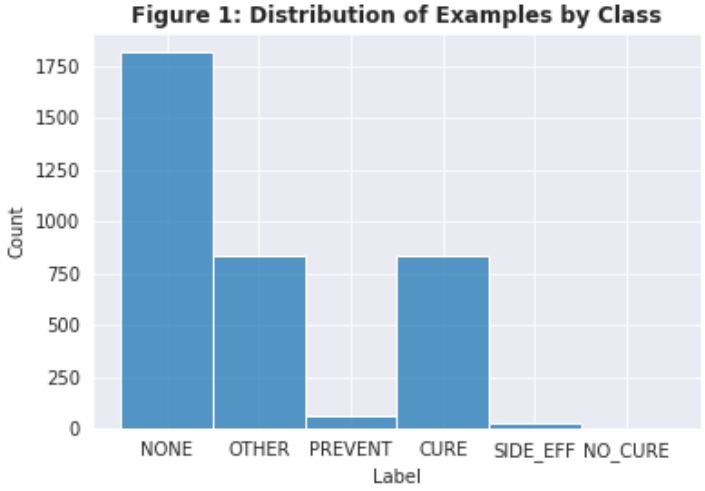
Jin et al.[10] conduct probing experiments on domain-specific corpora to see what information is carried in trained embeddings. They compare BERT, ELMo, BioBert, and BioELMo in NER and NLI tasks and probing tasks with mixed results. BioELMo outperforms BioBERT in their probing tasks but not NER / NLI tasks.

A. Roy et al.[15] examine various techniques to involve domain knowledge into BERT clinical relation extraction tasks. They incorporate UMLS knowledge sources (The Metathesaurus, Semantic network, and Lexicon tools) to test Knowledge Graph Embedding (KGE) models with ClinicalBERT.

3 Methods

3.1 Data

We are using a labeled MEDLINE 2001 dataset, a drug-treatment classification dataset annotated by a domain expert. The labeled dataset uses the first 100 titles and 40 abstracts from MEDLINE 2001. The data was labeled by Kaichi Sung, a former UC Berkeley SIMS master student with a biological background, labeling text for drug and treatment entities sentence by sentence, as well as a classification for the relationship between the entities in each sentence. There are a total of 8 types of defined relationships[1], each meant to be approached independently of its context. We reduced the number of relationships to classify to follow the problem setup from the work of Karaa et al[4].



We are aware that there is a disparity in the number of examples per class (Fig 1) and will place more emphasis on the results of relevant classes such as “PREVENT”, “CURE”, “SIDE_EFF”, “NO_CURE”.

3.2 Preprocessing

Before training any models, we first generated two versions of the dataset. In the former (untagged), manual semantically annotated tags were removed, leaving only the words as they would appear in the original text. In the latter (tagged), tagged words or phrases were converted into a tokenized form of the annotation, replacing the tagged words with a standard token in place of the disease/ailment and treatment/cure respectively. Specifically, we replaced entities tagged in the text with a singular @DIS$ for words or phrases tagged as a disease and a singular @TREAT$ for entities that were tagged as confirmed treatments. The tagging method was initially introduced by the NIH (National Institutes of Health) for their GAD[23] dataset. The GAD is an archive of human genetic association studies of complex diseases and disorders, further updated in BeFree GAD[24] dataset.We then split our MEDLINE 2001 data into a training, dev, and test set using a 80/10/10 ratio. We used these subsets for the training, tuning, and evaluation of all subsequent models.

3.3 Models

The first model we implemented was a generic BERT model (BERT-Large-Cased) to establish a baseline for performance. Originally developed by Google Research, BERT is trained on the entirety of English wikipedia as well as the Brown corpus. This BERT model, implemented through the ‘simpletransformers’ package (https://simpletransformers.ai/docs/classification-models/), included a classification layer downstream of

the BERT model that allowed us to perform multi-class classification. We trained the BERT model on both tagged and untagged versions of data to assess the capacity of BERT to extract relationships from medical text. We then evaluated the performance of these models on our test set. These results can be found in Table 1.

Next, we implemented domain specific versions of BERT, specifically BioBERT and ClinicalBERT[25]. BioBERT[22] is a domain-specific variation of BERT[14] built on the foundational embeddings of BERT with extensive training done using PUBMED, hosted by the National Library of Medicine and comprised of over 33 Million citations for biomedical literature. ClinicalBERT[25] is an extension of BioBERT that was additionally trained on a database of health data from over 40,000 patients that stayed in the critical care units of the Beth Israel Deaconess Medical Center, also known as MIMIC-III. We hypothesized that these models would outperform the generic BERT model given that it has been trained on domain-specific vocabulary.

The BioBERT model was initialized with weights from BioBERT-Base v1.1 updated and maintained by DMIS (Data Mining and Information Systems) Lab in Korea University. The ClinicalBERT model was initialized with weights from the ClinicalBERT repository[25]. Once again, we trained these models on both untagged and tagged versions of our dataset. We evaluated each model's performance on the test set. The performance metrics can be found in Table 4.

3.4 Imbalanced Dataset

The pre-annotated data from The BioText Project[1] was heavily skewed towards certain classes. Specifically, the "NONE" label took up 50.8% of our data and the "OTHER" label took up X% of our data. NO CURE, on the other hand, took up only 0.11% of our data. We addressed imbalances in our dataset to improve model performance for poorly performing classes. We referenced other publications[27] that also faced an imbalanced dataset and applied some of their strategies. First, we undersampled the high sample classes to balance the dataset, reducing the number of samples in any given class to 63 samples or less (the number of samples in the "PREVENT" class). Separately, we applied a method from Yong et al.[28] "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm"[28] to further enhance the variation of the samples taken from the majority class. This method involved converting each sentence into a vector within the top 50% of majority classes with a TFIDVectorizer. We then implemented a k-means clustering algorithm from sklearn to initialize k number of centroids equal to the number of samples we looked to limit in each majority class. We were then able to extract samples from the subset that had the closest cosine similarity with each centroid. Finally, we introduced a hyperparameter that allowed us to continue to optimize the sample size to ensure that all k samples taken from the majority class were unique.

4 Results / Model Evaluation

To analyze the performance of our models, we used the metrics of accuracy, precision, recall, and F1 score as defined by the relation extraction task outlined in DDRel (Drug Disease Relation Extraction from Biomedical Literature Using NLP and Machine Learning) (Karaa et al., 2021). Using these metrics to evaluate our models allows us to understand each of our models' performance compared to the DDRel approach as well as metrics generated from previous literature.

Model	Accuracy	Precision	Recall	F1 Score	NONE	OTHER	PREVENT	CURE	SIDE_EFF	NO CURE
(1A) BERT untagged	0.87	0.87	0.87	0.87	0.92	0.77	0.67	0.91	0.71	0.00
(1B) BERT tagged	0.95	0.93	0.95	0.94	1.00	1.00	0.00	0.84	0.00	0.00

Table 1: BERT Models

We can see from Table 1 that the BERT model set an impressive baseline of 89.4% accuracy with the untagged dataset, and performed even better when tags were added to the data. Furthermore, the tagged version of the model outperformed the untagged model in overall precision, recall, and F1 score. This gives us a strong indication that tagging words as a treatment or a disease helps improve model performance. However, we did notice that on a class by class basis there were discrepancies in performance. For example, the BERT tagged model never predicted "PREVENT" or "SIDE EFFECT" classes, while the BERT untagged model occasionally classified these sentences accurately. In addition, we noticed that the "NO CURE" class was never predicted in either model.

In Table 2/3, we see the performance metrics for the domain-specific BERT models that we implemented. We see that the untagged versions BIOBERT and ClinicalBERT both outperform their BERT counterparts in Table 1

Model	Total Accuracy	Precision	Recall	F1 Score	NONE	OTHER	PREVENT	CURE	SIDE_EFF	NO CURE
(2A) BioBERT untagged	0.89	0.89	0.89	0.89	0.95	0.78	0.67	0.92	0.71	0.00
(2B) BioBERT tagged	0.96	0.95	0.96	0.95	1.00	1.00	0.33	0.90	0.33	0.00

Table 2: BioBERT Models

in overall accuracy, precision, recall, and F1 score, albeit by a small margin. However, the untagged BERT model performs better than the untagged ClinicalBERT model for specific classes, such as "PREVENT" and "SIDE_EFF". The tagged versions of BIOBERT and ClinicalBERT have slightly higher overall accuracy scores, comparable recall scores, and slightly lower precision and F1 scores when compared to the tagged BERT model. We see this result because the BERT tagged model has a slightly lower overall false positive rate than the domain-specific tagged models. Once again, we see that the "NO CURE" label is never predicted.

To further explore the differences between the generic BERT architecture and domain-specific versions, we look at sentences that BERT misclassified and BIOBERT classified correctly. These sentences are shown below. Both these models refer to the tagged versions.

Sentence	True Label	BioBERT Prediction	BERT Prediction
The protective effect of @TREAT$ and @TREAT$ against @DIS$.	2	2	3
Initially, all eyes that had @TREAT$ without @TREAT$ also remained clear, but after 6 months, four of five developed @DIS$.	4	4	3
@TREAT$ and @DIS$.	1	1	3
@TREAT$ with @DIS$, protects from a subsequent vaginal challenge with the homologous serovar.	2	2	3

Table 3: BERT/BioBERT Sentence Comparison

From Table 3, we see that BioBERT was able to classify "PREVENT", "SIDE_EFF", and "OTHER" correctly in cases where BERT classified them as "CURE". This led us to believe that BERT was predicting "CURE" too often and thus generating many false positives within this class. This phenomenon would not be noticeable by simply looking at accuracy scores by class. Upon further investigation, we observed that the precision score for the "CURE" label with BioBERT was 90.32%, while for BERT it was only 84.47%. This result confirmed our suspicion and highlighted a benefit of using BioBERT over BERT.

Model	Accuracy	Precision	Recall	F1 Score	NONE	OTHER	PREVENT	CURE	SIDE_EFF	NO CURE
(3A) ClinicalBERT untagged	0.88	0.88	0.88	0.88	0.94	0.79	0.57	0.90	0.33	0.00
(3B) ClinicalBERT tagged	0.95	0.95	0.95	0.95	1.00	0.98	0.40	0.90	0.33	0.00

Table 4: ClinicalBERT Models

In Table 5, we see the results of the experiments we performed to combat class imbalances. It is clear that the clustering method outperformed the non-targeted clustering method across all performance metrics. While overall accuracy, precision, recall, and F1 scores were lower than our previous models, we see that the accuracy scores for our BIOBERT Clustered Model are more balanced between classes. This confirms our hypothesis that class imbalances, specifically a large proportion of "NONE" and "OTHER" examples, cause models to perform poorly on other classes. Additionally, it highlights the fact that overall performance metrics in previous models were inflated by the fact that there were so many "NONE" and "OTHER" examples, which are ultimately less useful labels than "PREVENT", "CURE", "SIDE EFFECT" and "NO CURE".

Model	Total Accuracy	Precision	Recall	F1 Score	Acc:NONE	Acc:OTHER	Acc:PREVENT	Acc:CURE	Acc:SIDE_EFF	Acc:NO CURE
(4A) BioBERT undersampling no tags	0.759	0.809	0.759	0.752	1.0	0.600	0.889	0.571	0.667	1.0
(4B) BioBERT undersampling tagged	0.759	0.754	0.759	0.741	1.0	1.000	0.857	0.500	0.500	0.0
(5A) BioBERT Clustering no tags	0.690	0.679	0.690	0.674	1.0	0.333	0.857	0.750	0.375	0.0
(5B) BioBERT Clustering tagged	0.759	0.715	0.759	0.735	1.0	0.750	0.667	0.750	0.600	0.0

Table 5: BioBERT Undersampling/Clustering Models

Using a smaller number of examples for the "NONE" and "OTHER" classes in these experiments still resulted in perfect accuracy scores for these classes. This shows that the reduction in the number of samples did not decrease performance within these two classes. While the accuracy score for "CURE" with the BIOBERT Clustered Model was lower than that of previous tagged models, we believe that the balanced performance of our clustered model is beneficial as one looks to utilize these models in real-world settings.

SOURCE	CURE	PREVENT	SIDE EFFECT	NO CURE
Biobert Tagged approach	93.33	40.00	20.00	0.00
DDRel approach	98.19	85.71	79.37	0.00
Abacha and Zweigenbaum	95.00	15.15	0.00	NA
Frunza et al.	93.60	76.50	50.00	NA
Suchitra and Sudah	90.30			
Muzaffar et al	98.05	93.55	88.89	NA
Wang et al.	90.49	NA	87.56	NA

Table 6: Comparison of F-score measures

Finally, we compare our approach (Biobert Tagged approach) to the work of Karaa et al.[4] along with other recent work on the same problem(see table 6). We decided to focus on classes that were not overrepresented in the data. Despite not achieving a comparable F1 score when compared to the DDRel approach, we believe that our approach is useful and informative.

5 Conclusion / Next Steps

With pre-trained transformers, we were able to create models that performed reasonably well, while doing significantly less preprocessing than others who tackled this specific problem set did. Although our model did not perform as well as the model created by Kaara et al.[4], we feel we made a strong case for the use of pre-trained transformers in biomedical relation extraction. Unlike the work done by Karaa et al., we were able to avoid complex feature engineering that remains impractical to apply to real-world problem sets. Furthermore, our success using simpler tags (just TREAT or DIS) justifies using a simpler annotation process in the future, which will enable easier labeling of future data. Lastly, the performance of our models after clustering highlights the utility of this method when dealing with imbalanced data.

Going forward, the field will need to improve upon methods of annotation and utilize unsupervised methods if we are to extract relations from biomedical text at scale. That being said, results from relation extraction and text classification provide hope that these tasks are solvable.

For next steps, we would like to work with a more complex and difficult dataset. The high accuracy of the baseline made fine-tuning and experimental results difficult to attribute to noise vs. true impacts. To build an even more powerful model, we would attempt to preprocess with UMLS and MetaThesaurus packages.

References

[1] B. Rosario and M. Hearst, "Classifying Semantic Relations in Bioscience Text", in the proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona. July 2004.

[2] W. Muzaffar, F. Azam, and U. Qamar, "A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set," Computational and Mathematical Methods in Medicine, vol. 2015, Article ID 910423, 12 pages, 2015.

[3] A. Thillaisundaram, T. Togia, "Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture," International Workshop on BioNLP Open Shared Tasks. 2019.

[4] W. Ben Abdessalem Karaa, E H lkhammash, A. Bchir, "Drug Disease Relation Extraction from Biomedical Literature Using NLP and Machine Learning," Mobile Information Systems, Article 9958410. 2021.

[5] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova1, "A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction" Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019.

[6] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," Journal of Biomedical Informatics, vol. 106, Article ID 103451, 2020.

[7] M. M. Balipa and R. Balasubramani, "Disease-treatment relationship extraction for psoriasis from online healthcare forums using NLP and classification techniques," International Journal of Applied Engineering Research, vol. 13, no. 6, pp. 3568–3573, 2018.

[8] M. Elberts, A. Ulges, "Span-based Joint Entity and Relation Extraction with Transformer Pre-Training." 2019.

[9] N. Perera, M. Dehmer, F. Emmert-Streib, "Named Entity Recognition and Relation Detection for Biomedical Information Extraction" REVIEW article, Front. Cell Dev. Biol., Article 103451. 2020.

[10] Q. Jin, B. Dhingra, WW. Cohen, X Lu, "Probing biomedical embeddings from language models," NAACL-HLT 2019 Workshop on Evaluating Vector Space Representations for NLP. 2019.

[11] R. Xing, J. Luo, T. Song. "BioRel: towards large-scale biomedical relation extraction." BMC bioinformatics 21.16: 1-13. 2020.

[12] P. Su, K. Vijay-Shanker, "Investigation of BERT Model on Biomedical Relation Extraction Based on Revised Fine-tuning Mechanism" 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020.

[13] P. Shi, J. Lin. "Simple BERT Models for Relation Extraction and Semantic Role Labeling." (2019)

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding." 2018.

[15] A. Roy, S. Pan, "Incorporating medical knowledge in BERT for clinical relation extraction," 2021.

[16] M. A Weinzierl, R. Maldonado, and S. M Harabagiu, "The impact of learning unified medical language system knowledge embeddings in relation extraction from biomedical texts." Journal of the American Medical Informatics Association, 27(10): 1556–1567. 2020.

[17] Q. Wei, Z. Ji, Y. Si, J. Du, J. Wang, F. Tiryaki, S. Wu, Cui Tao, Kirk Roberts, and Hua Xu. Relation extraction from clinical narratives using pre-trained language models. In AMIA Annual Symposium Proceedings, volume 2019, page 1236. American Medical Informatics Association. 2019.

[18] S. Yadav, S. Ramesh, S. Saha, A. Ekbal, "Relation Extraction from Biomedical and Clinical Text: Unified Multitask Learning Framework" 2020.

[19] P. Wang, T. Hao, J. Yan, and L. Jin, "Large-scale extraction of drug-disease pairs from the medical literature," Journal of the Association for Information Science and Technology, vol. 68, no. 11, pp. 2649–2661, 2017.

[20] A. Suchitra A, Sudha R, "Extraction of Semantic Biomedical Relations from Medline Abstracts using Machine Learning Approach. " National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications". 2012.

[21] Y. Peng, CH Wei, Z. Lu, "Improving chemical disease relation extraction with rich features and weakly labeled data." J Cheminform 8, 53. 2016.

[22] J. Lee, W. Yoon et al, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". Bioinformatics. 1–7. 2019.

[23] K. Becker, K. Barnes, T. Bright, S. Wang, "The genetic association database." Nature genetics 36.5 (2004): 431-432. 2004.

[24] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research" 2015.

[25] E. Alsentzer, J. Murphy, W. Boag, W. Weng, and D., and T. Naumann, and M. McDermott, "Publicly Available Clinical BERT Embeddings" 2019.

[26] A. Johnson, T. Pollard, R. Mark "MIMIC-III Clinical Database Demo" 2019.

[27] W. Hou, Y. Chen "Sentence-Level Propaganda Detection Using BERT with Context-Dependent Input Pairs" 2019.

[28] Y. Yong "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm" 2012.

[29] O. Frunza, D. Inkpen, and T. Tran, "A machine learning approach for identifying disease-treatment relations in short texts," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 801–814. 2010.

[30] O. Frunza, D. Inkpen, "Extracting relations between diseases, treatments, and tests from clinical data, advances in artificial intelligence," in Canadian Conference on Artificial Intelligence, pp. 140–145, Springer, Berlin, Germany. 2011.

[31] B. Abacha, P. A. Zweigenbaum, "Hybrid approach for the extraction of semantic relations from MEDLINE abstracts," in Computational Linguistics and Intelligent Text Processing, pp. 139–150, Springer, Berlin, Germany. 2011.

[32] P. Wang, T. Hao, J. Yan, and L. Jin, "Large-scale extraction of drug-disease pairs from the medical literature," Journal of the Association for Information Science and Technology, vol. 68, no. 11, pp. 2649–2661. 2017.