Initial Experiments on the DRSM Corpus

Aakash Bhatnagar*, Nidhir Bhavsar*, Tirthankar Ghosal†
*Navarachna University, India
†Oak Ridge National Laboratory, US

January 26, 2022

Abstract

This report documents the initial experiments we perform on the Disease Research State Model (DRSM) corpus. DRSM is a dataset containing abstract of rare disease articles. Here, 'rare disease' are referred to those diseases that affects less than 200,000 people in the USA. However, there are less than 5% approved therapies. The number of approved/possible therapies would increase if we can leverage the scientific knowledge encoded in rare disease research papers.

The DRSM corpus [3] consists of six distinct labels:

- Clinical characteristics or disease pathology
- Disease mechanism
- Therapeutics in the clinic
- Patient-based therapeutics
- Other
- Irrelevant

These labels represents the state of the research of each disease. As these classes are manually annotated, it takes a good amount of time to curate thousands of instances. So far, the DRSM corpus have a total of 8919 instances. The dataset is imbalanced as the number of instances are not evenly distributed amongst all classes. For example, the 'Irrelevant' class contains only 109 instances, which makes a ML model prone to overfit and produce unstable results. Similarly 'patient-based therapeutics' and 'other' classes also have fewer instances.

Since manual curation is time consuming, we tried to automate the exercise of 'abstract classification'. We experimented with variations of transformer-based neural models. We chose three BERT[5] models that are trained on large volume of biomedical articles.

1. BioBERT [8]
2. PubMedBERT [7]
3. SPECTER [4].

Along with these models, we experimented with dual-attention and label-wise-attention models. Dual attention[6] is a mechanism of applying self-attention twice. Using dual-attention model increased our scores significantly as seen in table1. Also to further improve our results, we added a label-wise-attention(LWAN)[1]

network that gives special attention to each word according to the labels. We adopted the same approach in the 'LitCovid' track of the BioCreative Challenge [2] to perform multi-label classification of COVID-19 articles. Our code is publicly available here[1].

	Precision	Recall	F1-score
clinical characteristics or disease pathology	0.9296	0.9392	0.9344
disease mechanism	0.9269	0.9664	0.9462
irrelevant	0.6154	0.4706	0.5333
other	0.8333	0.3846	0.5263
patient-based therapeutics	0.7857	0.8627	0.8224
therapeutics in the clinic	0.9180	0.8571	0.8865
micro avg	0.9161	0.9058	0.9109
macro avg	0.8348	0.7468	0.7749
weighted avg	0.9138	0.9058	0.9061
sample avg	0.9021	0.9058	0.9033

Table 1: precision, recall and f1-score of the best performing system i.e. Specter dual-attention LWAN

As we mention earlier, the DRSM corpus suffers from class imbalance. E.g., after splitting the training and testing set to 80:20, we had only 17 instances of the 'Irrelevant' class. We are further experimenting to tackle the class imbalance problem.

References

[1] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, S. Schockaert, and H. Saggion. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1508. URL https://aclanthology.org/D18-1508.

[2] A. Bhatnagar, N. Bhavsar, M. Singh, and T. Ghosal. Team cuni-nu at biocreative vii litcovid track: Multi-label topical classification of scientific articles using specter embeddings with dual attention and label-wise attention network. URL https://biocreative.bioinformatics.udel.edu/resources/publications/bc-vii-workshop-proceedings/.

[3] G. Burns, M. Torkar, A.-M. Istrate, H. Zaydens, L. Prins, E. Chou, D. Li, and S. Scovanner. Using document classification to map 'disease research state' across rare diseases. 2021. URL https://scinlp.org/.

[4] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. Specter: Document-level representation learning using citation-informed transformers. 2020.

[1] https://github.com/Nid989/Experiments-on-DRSM-corpus

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation, 2019.

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, Jan 2022. ISSN 2637-8051. doi: 10.1145/3458754. URL http://dx.doi.org/10.1145/3458754.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.