
BioMedBERT: LANGUAGE MODEL TO ADVANCE DISCOVERABILITY FOR BIO MEDICAL RESEARCH

May 1, 2020

ABSTRACT

BioMedBERT advances the state-of-the-art language models for increased information discoverability from biomedical literature. The number of peer-reviewed papers that are published everyday is dramatically increasing. In addition, more papers are being published in pre-prints such as arXiv and biorXiv. PubMed reports that more than 1 million biomedical research papers are published each year, this amounts to about two papers per minute [1]. Since the 1st of January 2020 more than 8000 papers on COVID-19 have been published on PubMed. Our work leverages the BERT language model architecture to pre-train a large-scale BioMedical language representation model. Using transfer learning, our model yields state-of-the-art results when fine-tuned on downstream specialized language tasks. Because we train our model to capture word structures from both general and biomedical corpora, we are able to achieve state-of-the-art results for Named Entity Extraction (NER), Relation Extraction (RE) and Question Answering tasks in biomedical datasets as well as in general classification datasets.

1 Introduction

2 Related Work

3 Methodology

4 Results

5 Discussion

6 Conclusion

7 Acknowledgements

References