BIOMEDBERT: LANGUAGE MODEL TO ADVANCE DISCOVERABILITY FOR BIOMEDICAL 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. 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

The late 2019 outbreak of the novel coronavirus strain of disease SARS-CoV-2 causing COVID-19 brought again to the fore the need for a tool that biomedical researchers, doctors and virologists can employ to augment their ability to sift through biomedical knowledge and existing research to extract novel insights and help them make new drug discoveries. The rate of new publications in the biomedical field is on the increase. PubMed reports that more than 1 million biomedical research papers are published each year, amounting to about two papers per minute [1]. Nature reports that as of January 20th, 2020, more than 50 papers have been published on COVID-19 [2]. For papers mentioning COVID-19 alone, more than 8000 peer-reviewed publications have been made on PubMed. With the rate of scientific papers on COVID-19 doubling every fourteen days [3], it becomes imperative to have a language understanding tool that can extract relevant information from the literature such as the research methodology, data, authors, results and citations for paper evaluation in light of extant research [4].