Connor Brown
Connorb8
CS410
Tech Review

Overview of BioBERT: A Domain-specific Language Representation Model Pre-trained on Large-scale Biomedical Corpora

BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, is a pre-trained bio-medical language representation model created for a variety of bio-medical text mining tasks.

On average, 3000 biomedical research papers are published every day. This volume of information is so high, it's not feasible for researchers to read through every paper to stay up to date on the latest knowledge being released. To address this problem, automating the retrieval and analysis of information from these papers could be very valuable to researchers around the world.

Biomedical texts contain lots of domain-specific terminology and content that is only understood by domain experts; there are terms used that are not likely to be well-understood by a general language model. BioBERT transfers the knowledge of large amount of biomedical texts into biomedical text mining models.

BioBERT is trained with different combinations of general and biomedical domain words.

BERT is a pre-trained model that was trained on 2.5 billion words from Wikipedia and 0.8 billion words from BooksCorpus. BioBERT is trained on an additional 4.5 billion words from **PubMed Abstracts** and 13.5 billion words from **PMC Full-text articles,** which contain content specific to the biomedical domain.

The BioBERT authors use the Wordpiece Model to handle the out of the vocabulary problem.

This involves splitting rare words into pieces, and representing words by frequent subwords. An example is the word**, “**Immunoglobulin”, which is split into “I ##mm ##uno ##g ##lo ##bul ##in”.

The three tasks that BioBERT was tested on are named entity recognition, relation extraction, and question answering.

Named Entity Recognition is the recognition of numerous domain specific proper nouns in biomedical corpus.

Relation Extraction is a task of classifying relations of named entities occurring in the biomedical corpus. (ex: a sentence classification task)

Question Answering is a task of answering questions written in natural language given related passages. In biomedical text mining tasks, this includes answering patients’ questions such as ‘‘In which breast cancer patients can palbociclib be used?’’, or questions raised by biomedical researchers such as ‘‘Where is the protein Pannexin1 located?’’

BioBERT outperforms BERT on the three biomedical text mining tasks including:

1. biomedical named entity recognition (1.86% absolute improvement over BERT)
2. biomedical relation extraction (3.33% absolute improvement over BERT)
3. biomedical question answering (9.61% absolute improvement over BERT)

The results clearly demonstrate that utilizing BioBERT over BERT is beneficial when conducting named entity recognition, relation extraction, and question answering tasks within the biomedical domain.

References:

https://www.arxiv-vanity.com/papers/1901.08746/

https://medium.com/@raghudeep/biobert-insights-b4c66fde8fa7

https://arxiv.org/pdf/1609.08144.pdf

https://academic.oup.com/bioinformatics/article/36/4/1234/5566506