BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining

Luo et al. | Briefings in Bioinformatics, 2022

Varun Venkat Rao | November 16th, 2023
EECS 598-007 Biomedical AI | University of Michigan

Introduction

Methods

Experimentation

Results

Discussion + Conclusion

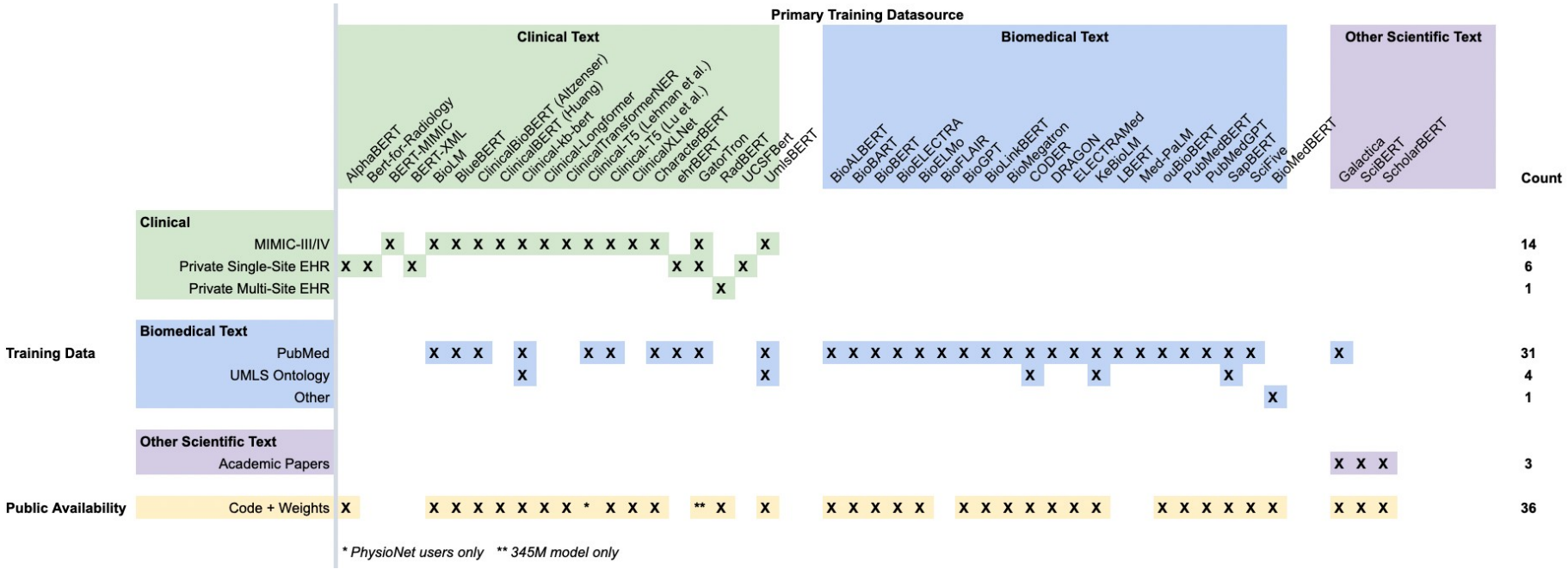
Landscape of language models

BERT

GPT-2

Discriminative Models

Generative Models



Stanford Institute for Human-Centered Artificial Intelligence. (2022, March 30). Shaky Foundations: The Foundation of Models in Healthcare. Stanford University. https://hai.stanford.edu/news/shaky-foundations-foundation-models-healthcare

Introduction

Methods

Experimentation

Results

Discussion + Conclusion

Step-I Data-centric Knowledge Injection

GPT-2

General domain LLMs for text generation

PubMed

15M Academic Papers

Knowledge Injection

Coronaviruses are a family of viruses…

BioGPT$_K$

Step-II Task Specific Fine-Tuning

BioGPT$_K$

LLM with Medical Knowledge

Relation Extraction

Question Answering

Document Classification

BioGPT

LLM finetuned for downstream tasks

PMC-LLaMA: Towards Building Open-source Language Models for Medicine, Wu et al, arXiv 2023

Preparing Label Space for Generative Models

Relation Extraction

Let us use a ⟨drug, target, interaction⟩ triplet as example. Suppose we would like to extract triplet ⟨dextropropoxyphene (drug name), mu-type opioid receptor (target name), inhibitor (relation)⟩ from an input document. Then the `svo` representation is:

dextropropoxyphene inhibits mu-type opioid receptor.

The `is-of` form is:

dextropropoxyphene is the inhibitor of mu-type opioid receptor.

The `rel-is` form is:

the relation between dextropropoxyphene and mu-type opioid receptor is inhibitor.

Question Answering

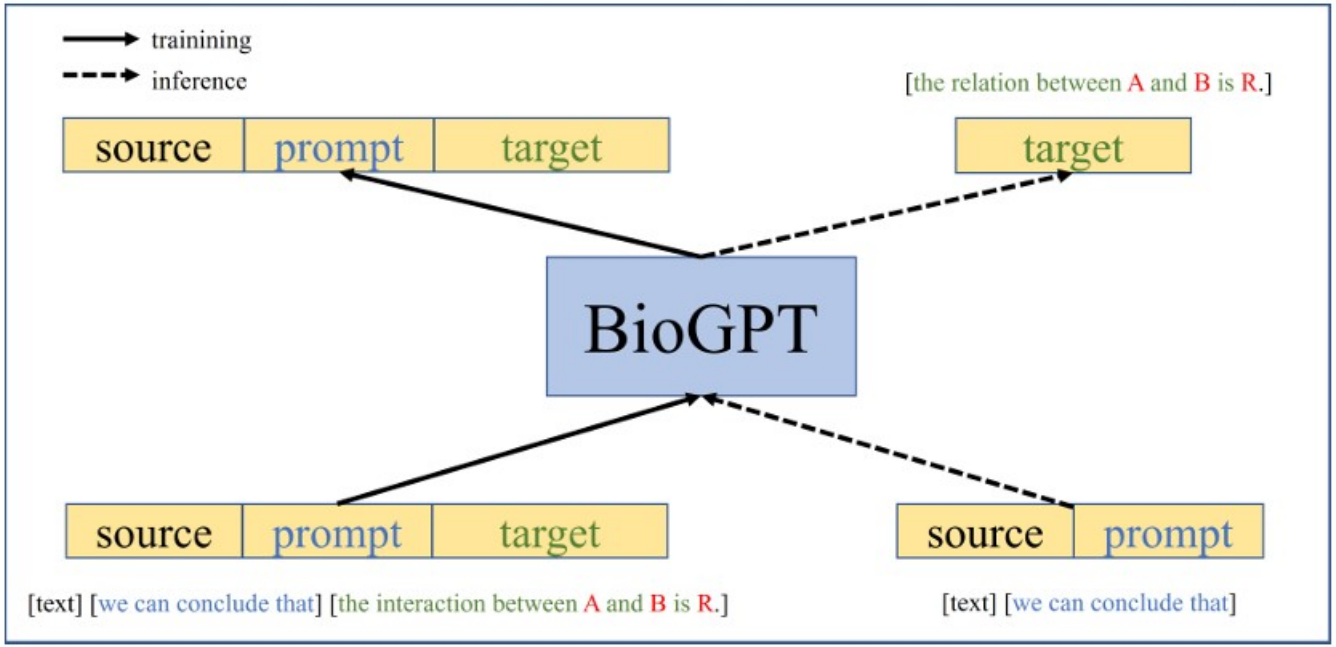
`source`: question: `question text.` context: `context text.`
`target`: the answer to the question given the context is yes.

Document Classification

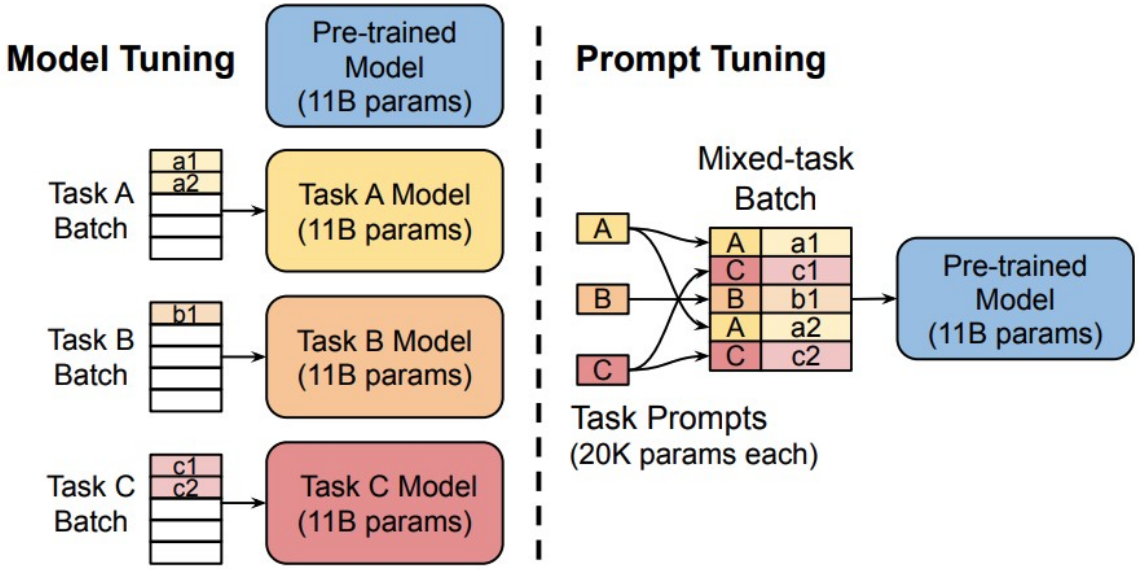
Task description Given a document text, the goal is to classify the type of the document.
Method: We generate the target sequence using the format "the type of this document is `label`". For example:

the type of this document is genomic instability and mutation.

Prompt Finetune



Prompt Finetune – Soft vs Hard Prompt



Hugging Face. "Prompting." Hugging Face Documentation, n.d., https://huggingface.co/docs/peft/conceptual_guides/prompting.

Introduction

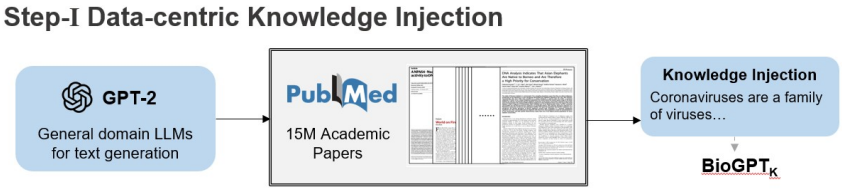
Methods

Experimentation

Results

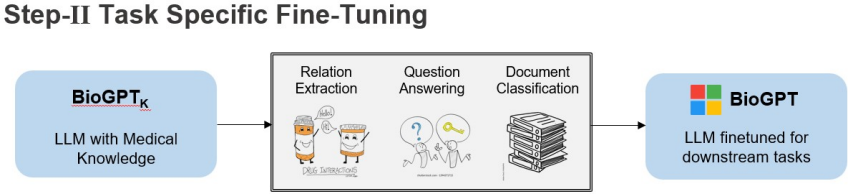
Discussion + Conclusion

Implementation details



Key Parameter	Value
Dataset	PubMed Abstracts
Backbone	GPT-2$_{medium}$
Vocab Builder	BPE
Optimizer	Adam
Peak Learning Rate	
Pre-train Steps	
Schedule	Inverse square root decay with warmup

Implementation details



Key Parameter	Value
Dataset	*Task Specific Dataset* *(BC5CDR, KD-TI, DDI, PubMedQA, HOC)*
Backbone	BioGPT$_K$
Batch Size	1024 tokens
Accumulated Steps	32

Introduction
Methods
Experimentation
Results
Discussion + Conclusion

End-to-end Relation Extraction

Chemical-disease-relation

Dataset: BC5CDR

Model	Precision	Recall	F1
GLRE (gt+pred)	34.82	18.29	23.99
GLRE (pred+pred)	23.00	4.88	8.05
GPT-2 [6]	43.92	32.55	37.39
REBEL [24]	34.28	39.49	36.70
REBEL_{pt} [24]	40.94	21.20	27.94
seq2rel [25]†	43.5	37.5	40.2
BioGPT	49.44	41.28	**44.98**
BioGPT†	49.52	43.25	**46.17**

Drug-target-interaction

Dataset: KD-DTI

Model	Precision	Recall	F1
Transformer + PubMedBERT -attn [14]	25.35	24.14	24.19
$\text{GPT-2}_{\text{medium}}$	30.53	27.87	28.45
REBEL	32.36	29.58	30.39
REBEL_{pt}	35.73	32.61	33.32
BioGPT	40.00	39.72	**38.42**

Drug-drug-interaction

Dataset: DDI

Model	Precision	Recall	F1
$\text{GPT-2}_{\text{medium}}$	23.39	31.93	24.68
REBEL	35.36	28.64	28.27
REBEL_{pt}	46.59	39.60	**40.56**
BioGPT	41.70	44.75	**40.76**

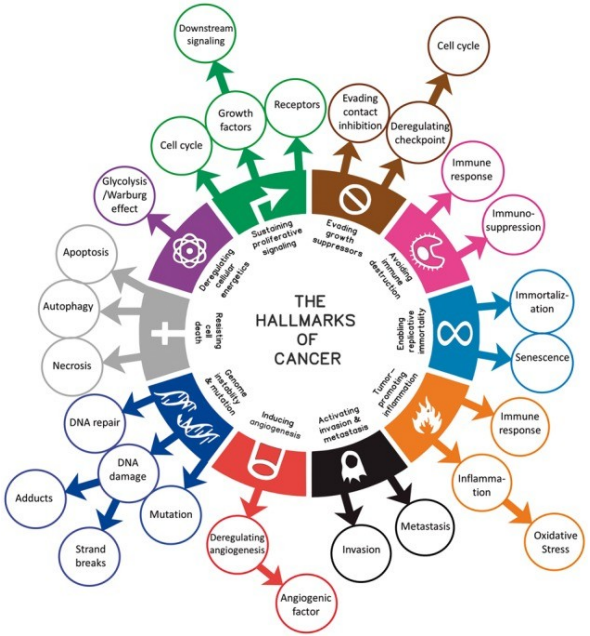
Question Answering – PubMedQA

Question:
Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?
Context:
(Objective) Recent studies have demonstrated that statins have pleiotropic effects, including anti-inflammatory effects and atrial fibrillation (AF) preventive effects [...]
(Methods) 221 patients underwent CABG in our hospital from 2004 to 2007. 14 patients with preoperative AF and 4 patients with concomitant valve surgery [...]
(Results) The overall incidence of postoperative AF was 26%. ***Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, p=0.005).*** Multivariate analysis demonstrated that independent predictors of AF [...]
Long Answer:
(Conclusion) Our study indicated that preoperative statin therapy seems to reduce AF development after CABG.
Answer: yes

Model	Accuracy
PubMedBERT [9]	55.8
BioELECTRa [28]	64.2
$\text{BioLinkBERT}_{\text{base}}$ [29]	70.2
$\text{BioLinkBERT}_{\text{large}}$ [29]	72.2
BioGPT	**78.2**

Jin, Qiao, et al. "Pubmedqa: A dataset for biomedical research question answering." arXiv preprint arXiv:1909.06146 (2019).

Document Classification – HOC



Model	F1
BioBERT [10]	81.54
PubMedBERT [9]	82.32
$\text{PubMedBERT}_{\text{large}}$	82.70
$\text{BioLinkBERT}_{\text{base}}$	84.35
$\text{GPT-2}_{\text{medium}}$	81.84
BioGPT	**85.12**

Baker, Simon, et al. "Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer." Bioinformatics 33.24 (2017): 3973-3981.

GPT-2 vs Bio-GPT

Input	Model	Text
Bicalutamide	GPT-2	Bicalutamide, a synthetic bicalutamide derivative, was found to increase the cellular proliferation of C. elegans cells. This effect was not observed with the other synthetic bicalutamide derivatives. This is in line with our previous work which showed that bicalutamide increases the proliferation of C. elegans cells but not in the presence of synthetic bicalutamide.
	BioGPT	Bicalutamide (Casodex) is an androgen receptor (AR) antagonist approved for the treatment of metastatic castration-resistant prostate cancer (mCRPC) in patients who have progressed on or are ineligible for docetaxel chemotherapy, as well as for the treatment of early-stage prostate cancer in men who have not progressed on or are ineligible for docetaxel chemotherapy, as well as for the treatment of metastatic castration-sensitive prostate cancer (mCSPC) in men who have not progressed on or are ineligible for docetaxel chemotherapy.

GPT-2 vs Bio-GPT

Input	Model	Text
Apricitabine	GPT-2	Apricitabine, duloxetine, and citalopram) were investigated in patients with a history of treatment-resistant depression (n = 10). We found that treatment with the three drugs significantly decreased depressive symptoms, but not depressive symptoms over the next 12 weeks. We conclude that antidepressant drugs that reduce depressive symptoms and reduce depressive symptoms over time may be beneficial in patients with recurrent depression.
	BioGPT	Apricitabine is an oral prodrug of 5-aza-2 '-deoxycytidine (5-aza-CdR), a DNA methyltransferase (DNMT) inhibitor, which has been approved by the US Food and Drug Administration (FDA) for the treatment of myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) in combination with low-dose cytarabine (Ara-C) and granulocyte colony-stimulating factor (G-CSF) for patients with intermediate-2 or high-risk MDS or AML.

GPT-2 vs Bio-GPT

Input	Model	Text
CP-673451	GPT-2	CP-673451) and 4-5-6-3-1-2-3-4-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22-23-24-25-26-27-28-29-30-31-32-33-34-35-36-37-38-39-40-41-42-43-44-45-46-47-48-49-50...
	BioGPT	CP-673451 is a potent, selective, and orally active inhibitor of human neutrophil elastase (HNE) and human cathepsin G (CatG) with in vitro and in vivo anti-inflammatory activity in a variety of animal models of inflammation and in a model of acute lung injury (ALI) in the rat induced by intratracheal instillation of lipopolysaccharide (LPS) and tumor necrosis factor-alpha (TNF-alpha), a model of acute lung injury (ALI) in which neutrophils play an important role.

Introduction
Methods
Experimentation
Results
Discussion + Conclusion

Ablation Study: Target Sequence Format

Target format	Precision	Recall	F1
<head> head_entity <tail> tail_entity <relation> relation	38.21	40.21	37.32
svo (head_entity relation tail_entity)	37.95	37.77	36.57
is-of (head_entity is the relation of tail_entity)	39.37	39.11	37.77
rel-is (the relation between head_entity and tail_entity is relation)	38.93	40.70	38.38

Ablation Study: Prompt Design

Prompts	Precision	Recall	F1
we have that	38.55	38.37	36.95
in conclusion,	39.03	39.45	37.76
we can conclude that	39.56	39.88	38.16
continuous embeddings (length=1)	39.50	39.71	38.06
continuous embeddings (length=5)	39.57	39.63	38.09
continuous embeddings (length=9)	38.93	40.70	38.38
continuous embeddings (length=13)	39.48	39.17	38.60
continuous embeddings (length=17)	39.82	39.60	38.28

Key Contributions

Generative Language Model for Biomedical Domain

State-of-the-art Performance Prowess

Prompt and Sequence Design Insights


UNIVERSITY OF
MICHIGAN

(T/F) In the knowledge injection stage, it's advisable to start with a Language Model (LLM) pre-trained on the general domain and subsequently fine-tune it using biomedical literature.

Target format
<head> head_entity <tail> tail_entity <relation> relation
svo (head_entity relation tail_entity)
is-of (head_entity is the relation of tail_entity)
rel-is (the relation between head_entity and tail_entity is relation)

What key finding does the paper highlight regarding the effectiveness of target sequences compared to structured prompts in downstream tasks?