Clinical-Trial-Matching

Devi Phani Sravanthi Nittala and Pramod Kumar Undrakonda

Abstract

In the field of healthcare, a critical challenge exists in efficiently connecting patients with appropriate clinical trials that align with their unique medical conditions, demographic factors, and specific criteria. Clinical trials represent invaluable opportunities for patients to access cutting-edge treatments and therapies, contribute to medical research, and potentially find solutions to their health concerns. However, the process of matching patients with suitable clinical trials is often complex, time-consuming, and can lead to missed opportunities for both patients and medical researchers. Our project is an implementation of a model that aims to address this issue and ease the process of patient to trial matching. We have implemented a document-ranking approach to our project, where we use pretrained models finetuned for the purpose of document ranking and then use this model to rank the clincal trials to find the most relevant for a particular clincal case.

Introduction

The landscape of healthcare is continuously evolving, marked by breakthroughs in medical science, technological advancements, and a growing emphasis on personalized patient care. At the heart of this dynamic ecosystem lies the crucial domain of clinical trials, playing an instrumental role in advancing medical knowledge, testing innovative treatments, and shaping the future of healthcare. Clinical trials, by design, aim to investigate the safety and efficacy of new drugs, therapies, or medical interventions. However, a persistent challenge in the realm of clinical research is the efficient and timely recruitment of eligible participants.

The traditional methods of patient recruitment for clinical trials often involve manual screening of medical records, a process susceptible to delays, inefficiencies, and missed opportunities for patients to access potentially life-changing treatments. In response to these challenges, the intersection of healthcare and cutting-edge technologies has given rise to innovative solutions. One such transformative solution is the integration of Natural Language Processing (NLP) and machine learning into the clinical trial recruitment process. This project explores the development and implementation of a sophisticated "Clinical Trial Matching" system, leveraging these advanced technologies to revolutionize the patient recruitment landscape.

Digital approaches such as keyword-based approaches, data analytics, etc. have been approaching the forefront of replacing traditional methods and the latest among these is the usage of Large Language Models. We have proposed one such system.

Method

In our project, we propose to make use of use of bi-encoder model to rank the clincal trials based on relevance of inclusion criteria. A bi-encoder model is one where two encoders are used to encode two sentences, and then the embeddings obtained are compared using a metric such as the cosine similarity or dot score. Our model, based on the one highlighted in [2], consists of Sentence-BERT models that have been finetuned on MSMARCO dataset for document ranking. In additon to these pretrined models, we train our own model by finetuning a ClincalBERT model for the MSMARCO data based on the MultipleNegativesRankingLoss approach, in which a model is trained using triplets consisting of a query, a positive document and a negative document, such that the model creates encodings for all three. The encodings must be such that there must be small distances between the positive document/passage and the query, and large distances between the query and the negative examples.

The authors proposed a method of better performance by ensuring good negative example by using a Cross Encoder to find score for such negative examples and then using these for training the bi-encoder. We have used their training code to train a ClinicalBERT model in order to account for the medical and clincial jargon present in the proposed clinical trial reports.

There have been systems which have used the aforementioned approach of document ranking for Clincal Trial Matching, including IBM Watson, however, we aim to specifically test the appicability of a BERT-based model for the same.

Limitations

The biggest challenge we faced during this project is the lack of public-domain medical data. Due to concern about patient privacy and rights, clincal reports are not made available in the public domain and this forced us to rely on synthetically generated data for our purpose, which may not accrately reflect the real-world.

Experimental Setup

Dataset

We made use of pretrained models finetuned on the MSMARCO dataset for the purpose of our project. The MSMARCO[3] dataset is a benchmarking dataset used for passage and document ranking. It consists of a corpus of queries and corresponding relevant documents which are to be ranked. There are 3.2 million documents in it with a goal of ranking them.

For the purpose of the clincal trials, we used the API provided by clincaltrials.gov to fetch relevant clincal trials based on synthetically created patient cases.

Embarking on the integration of advanced technologies into healthcare demands a conscientious consideration of ethical dimensions, with paramount attention to patient privacy and data security. The "Clinical Trial Matching" system places ethical principles at its core, ensuring strict compliance with guidelines and regulations governing patient data.

Implementation

We made use of the sentence-tranformers library offered on HuggingFace to finetune as well as to test our and other pre-trained models. For the purpose of training our model, we trained it according to the MultipleNegativesRankingLoss given in [2] for the ClinicalBERT model instead of the DistilBERT used by the authors. We ran it on a TPU provided on Google Colab and the training took apporximately 20 hours to complete.

The main pipeline of out project is the bi-encoder system, which we have demonstrated in the ClincalTrials_Demo.ipnyb file. The idea behind a bi-encoder is that we have two encoder models, here the various pretrained models as well as our model, which take into one sentence each. In this case, it would be the query and the inclusion criteria of all the clinical trials. Then we take the cosine similarity of each criterion against the query and use the result to sort the clincal trials in order.

We have mainly run our program using Cloud GPUs and TPUs offered by Google Colab. We made extensive use of the HuggingFace library.

Results

The results can be divided into two parts: first we test the performance of a ClincalBERT model on the MSMARCO dev dataset and second, we view the performance of the model for the purpose of ranking clinical trials.

It is to be noted that since the ClincalBERT model is not directly supported by sentence-tranformers some of the weights are initialized using Mean Pooling during training.

MSMARCO - Test

The model which had been trained[2] as described above was tested using the evaluation script provided in [2] and also separetely to obtain the various metrics for the model.

Accuracy	Precision	Recall	Mean Reciprocal Rank	Mean Average Precision	NDCG
0.9290830945558739	0.09802292263610315	0.9240210124164278	0.815079763041796	0.815079763041796	0.8381790579726461

In the above table, the mean reciprocal rank refers to the metric that can evaluate a ranking system by taking the recirpocal of the rank at which the first relevant document is found. The higher the MRR is, the better is the model performance. The next is the mean average precision which is essentially the mean of the average precisions for each query, over the entire corpus. The accuracy of our system is fairly high showing that the model performs well over the test set as well.

Clincal Trial - Test

We test the same query against the first 20 clincal trials that have been fetched from the clincaltrials.gov for a given condition. For the case shown below, the condition taken was 'heart attack'. We encoded the query and then comapred against the encoding for the Inclusion criteria which dictates a set of conditions that a patient must match in order to be eligible for a clinical study. The query was given as follows:

```
A 45 year old with a clinical diagnosis of ST-segment elevation acute myocardial infarction.
```

We considered the following models and mentioned below each model is the NCT ID, i.e. the unique ID given to each clinical study, and the cosine similarity obtained w.r.t. the above query.

Model	Our Model (sravn/msmarco-clincalbert)	sentence-transformers/msmarco-bert-base-dot-v5	Capreolus/bert-base-msmarco	sentence-transformers/msmarco-MiniLM-L6-cos-v5
NCT ID of 1st ranked trial	NCT01484158	NCT01484158	NCT01109225	NCT01484158
---	--	--	--	--
Cosine Similarity	0.6799761056900024	0.9552577137947083	0.9649959206581116	0.717819094657898

As we can see in the above results, our model as well as the models trained by the sentence-transformers return the same (and correct) most relevant clincal trial, however, unlike our model, the other had a higher cosine similarity. The model ' Capreolus/bert-base-msmarco' gave another trial as the most relevant trial with a lower cosine similarity. We conclude that our model performs well for the ranking of the clinical trials, however, we must further test it with more detailed patients descriptions.

(The most relevant clinical trial was: Gait Speed for Predicting Cardiovascular Events After Myocardial Infarction.)

Discussion

Our project aimed to implement a system that assists healthcare professionals and patients to find the relevant clinical trials. While we have succeded in defining such a system, which takes the patient condition as input, fetches the clinical trials, and then based on the inclusion criteria for the trial suggests the most relevant ones, we are yet to ascertain the accuracy of this system in a real-world scenario. A major hurdle as previously discussed, is the, rightful, lack of access to actual medical data due to HIPAA rules which would have let us test the model more thoroughly. As it stands, our proposed system of using a pretrained bi-encoder model finetuned on the MSMARCO dataset for document ranking, proves to be promising.

Relevant Links:

HuggingFace model: sravn/msmarco-clincalbert

Clincal Trials Repository: https://clinicaltrials.gov

References

[1] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[2] MS MARCO https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/ms_marco/

[3] Nguyen, Tri, et al. "MS MARCO: A human generated machine reading comprehension dataset." choice 2640 (2016): 660.

[4] Henderson, Matthew, et al. "Efficient natural language response suggestion for smart reply." arXiv preprint arXiv:1705.00652 (2017).

[5] Gao, Junyi, et al. "COMPOSE: Cross-modal pseudo-siamese network for patient trial matching." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020. https://github.com/v1xerunt/COMPOSE/tree/master