Gastrointestinal Cancer Related Question Answering Using BERT

S ArunaDevi, J Abirami and B Bharathi

Department of CSE
Sri Sivasubramaniya Nadar College of Engineering,
Tamil Nadu, India

Abstract

In today's world, Large Language Models (LLMs) have made significant advancements across various fields. While there are numerous models tailored for specific purposes, there remains a scarcity of models dedicated to the medical domain. This paper details our participation in the shared task "Conversational System for Differential Diagnosis of GI Cancer" at FIRE 2024, which addresses this gap. We employed a BERT model specifically trained for question answering. This task involved responding to inquiries posed by both doctors and patients.

Keywords

BERT, Gastrointestinal cancer, Large Language model, Bleu, Rouge

1. Introduction

Natural language processing (NLP) has been used to extract information from medical text for several decades [1], and a thorough review of NLP-based information extraction for cancer-related EMR notes can be found in the study by Datta et al[2]. More recently, there has been growing interest in more highly automated deep learning approaches for clinical NLP [3]. A category of tumors affecting the digestive system is called gastrointestinal cancers. These malignancies can develop in a number of organs that are involved in food absorption and digesting. Gastrointestinal (GI) complaints and symptoms account for approximately 10% of all general practice consultations and are apparently very common in the general population[4, 5] . They are common in the general population and can vary in severity from more serious illnesses like inflammatory bowel disease (IBD) or irritable bowel syndrome (IBS) to moderate problems like dyspepsia or stomach pain. The aforementioned issues involve a substantial amount of datasets.

A deep learning model that has been trained on a large volume of textual data to produce language that is similar to that of humans is called a Large Language Model (LLM). Large Language Models (LLMs) play a crucial role in clinical information processing, showcasing robust generalization across diverse language tasks[6]. LLMs can offer innovative methods in medical education[7]. Advanced neural network designs, usually based on the Transformer, are used by these models, including GPT (Generative Pre-trained Transformer) and BERT, to capture the complexity of language, such as grammar, context, and meaning. Using huge datasets to train LLM models on billions of parameters is one of their main advantages. Additionally, they are also helpful for answering questions.

Joseph Ross Mitchell[8] emphasised on the growing importance of deep learning approaches for informing the patients and doctors about the various details of gastrointestinal cancer. A model created by LLM and included in the BERT model. BERT and related architectures have facilitated significant improvements in multiple medical applications and others[9]. Deep learning models of the BERT (Bidirectional Encoder Representations from Transformers) kind are intended for natural language processing (NLP) applications. Recognize context in both directions: BERT reads full sentences at once, in contrast to previous models that only processed text in one way (either left-to-right or right-to-left).

Forum for Information Retrieval Evaluation, December 12-15, 2024, India
aruna2210499@ssn.edu.in (S. ArunaDevi); abirami2210382@ssn.edu.in (J. Abirami); bharathib@ssn.edu.in (B. Bharathi)

This enables BERT to comprehend a word's context by taking into account both its left and right surrounds.BERT can be used to solve a variety of NLP issues, including: Text categorization, answering questions, etc.

To sum up, BERT models are highly effective in deciphering the meaning and context of language, which makes them helpful for a variety of NLP applications requiring in-depth language comprehension. This paper is sectioned as follows: Section 2 describes the previous works that have been done by various authors in the field of hope speech detection. Section 3 provides a detailed explanation of the data set. Section 4 provides an overview of the work done. Section 5 deals with the development of model for interactive question answering for different types of gastrointestinal cancer. Section 6 analyses the result obtained from our system. Section 7 provides the conclusion of this paper.

2. Related Works

Joseph et al[8] developed a BERT based system to automatically extract detailed tumor site and histology information from oncological pathology reports. They first trained a base language model to comprehend the technical language in pathology reports. This involved unsupervised learning on a training corpus. Then they trained a question-and-answer (Q&A) model that connects a Q&A layer to the base pathology language model to answer pathology questions. Their final system called as CancerBERT(caBERTnet) network consisted of a network 3 BERT based model.

Qingqing Zhou et al[6]focused on the application of RAG in the field of clinical gastroenterology in China, aiming to address the issue associated with the continuous increase in the infection rate of Helicobacter pylori and the rising incidence of gastric cancer. The fine-tuned model exhibited an 18% improvement in hit rate compared to its base model, gte-base-zh. Moreover, it outperformed OpenAI's Embedding model by 20%. For fine-tuning the gte-base-zh model, we employed GPT-3.5 Turbo to aid in generating question-answer pairs.The aim of Adi Lahat et al[10] is to evaluate the performance of ChatGPT in answering patients' questions regarding gastrointestinal health. ChatGPT was able to provide accurate and clear answers to patients' questions in some cases, but not in others.

Jiajia Yuan et al[11] found out that prompt engineering affects large language models' performance in GI oncology. They designed the prompts as follows: Initially, the models are subjected to a more sophisticated introduction prompt, intricately crafted with complex semantic. Then an advanced method of in-context learning was introduced, encouraging the models to extract knowledge and patterns from various contexts rather than individual sentences, fostering a more comprehensive understanding of the text. Lastly, they have implemented an iterative feedback loop through multi-round question-and-answer sessions, reinforcing the model's ability to comprehend, retain, and apply information over successive interactions.

3. Dataset Description

In this shared task, we were not given any explicit dataset to train the model. There were no specific methodology given to access public data sources. So, we were asked to use API for data or any other public data sources. Considering this, we have collected our dataset from Wikipedia for each type of gastrointestinal cancer for general information such as symptoms, causes, diagnosis and treatments. For genetic mutations and its effect for each of the cancers, we have used both Wikipedia and various other public data sources such as journals, articles and sites.[1]

[1]Data sources: Esophageal Cancer, Genetic mutations for Esophageal Cancer, Pancreatic Cancer, SMAD4- Pancreatic cancer, p53-Pancreatic Cancer, ARID1A- Pancreatic Cancer, GNAS- Pancreatic Cancer, KRAS- Pancreatic Cancer, MEN1- Pancreatic Cancer, Gallbladder Cancer, CDKN2A- Gallbladder Cancer, Stomach Cancer, Liver Cancer, Anal Cancer, Colorectal Cancer, PIK3CA- Colorectal Cancer, Gastrointestinal Stromal Tumor

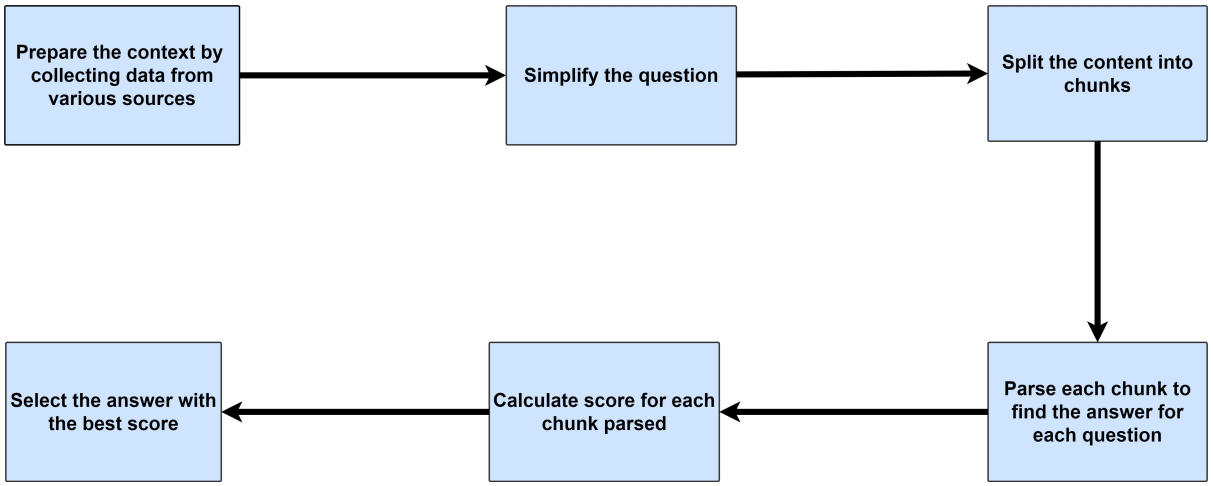

Figure 1: Proposed Architecture

In our dataset, we have collected all the above mentioned informations for 8 gastrointestinal cancers such as Gastrointestinal Stromal Tumor(GIST), esophageal, pancreatic, gall bladder, stomach, liver, anal and colorectal cancers. Each cancer's data is stored as a paragraph. Then data is labelled appropriately for distinguishing between the different factors of that cancer.

A sample of the dataset:
Gallbladder cancer: *Symptoms: Steady pain in the upper right abdomen, Indigestion (dyspepsia),Bilious vomit,Weakness, Loss of appetite, Weight loss, Jaundice and vomiting due to obstruction, Early symptoms mimic gallbladder inflammation due to gallstones. Diagnosis: Transabdominal ultrasound, CT scan, endoscopic ultrasound, MRI, and MR cholangio-pancreatography (MRCP) can be used for diagnosis.*[2]

The entire data we that we have used for the task is uploaded in our github page.[3]

4. Proposed Work

The proposed system architecture, as illustrated in Figure 1, follows a structured methodology designed for efficient question answering. The methodology is composed of the following key steps: 1. Selecting the context according to the question given to the system so that the model can extract the answer from it. 2. In the this step, the system simplifies the input question. This transformation helps the model better understand the essence of the question . 3. The selected context is then divided into multiple chunks. This segmentation allows the system to handle large contexts more efficiently. 4. Each chunk of context is parsed individually to locate potential answers. This step allows the model to consider all possible segments of the text that could contain the answer. 5. The model calculates the score for each chunk parsed. 6. Finally by comparing the scores of all the chunks, we choose the answer with the best score.

[2]Gallbladder cancer
[3]Access our work and dataset through Github

5. Implementation

LLaMA(Large Language Model Meta AI) models are large language models built to handle more general-purpose language understanding and generation. These models are usually larger in scale and require more memory and computational power. LLaMA2 has 7 billion parameters and LLaMA3 has 47 billion parameters. Whereas BERT large model has 340 million parameters. Due to large computational requirements of LLaMA models, we were not able to deploy it in our systems.

Hence we have used the method of Question Answering (QA) by utilizing a BERT model (bert-large-uncased-whole-word-masking-finetuned-squad) to find an answer inside a given text passage based on the question asked by the doctor. Tokenization, input preprocessing and output extraction are the main steps in the process. With a pre-trained tokenizer built for BERT, the question and text passage are transformed into tokens. Tokenization divides the text into more manageable units called tokens, which are then translated to integer IDs that the BERT model can comprehend.Usually, BERT models can only handle 512 tokens. So, the context is divided into 3 segment: (i) From causes to symptoms (ii) From symptoms to genetic mutations (iii) From genetic mutations to treatments

After the above division into segments, the input IDs are produced as tensors and fed into the BERT model together with the attention mask, which indicates which tokens are actual and which may be padding. Two sets of logits, called startscores and endscores, are returned by the model. These sets of logits represent the likelihood that each token marks the beginning or end of the answer, respectively.

The BERT model predicts the most likely span of tokens that answer the question by finding the token positions with the highest start and end scores. Subword tokens are merged correctly by stitching the tokens between these points back into a human-readable format (by removing continuation indicators).The model produces an invalid answer (e.g., predicting [SEP]) if the start or end scores are too low, or it returns a fallback message indicating that no answer could be found.

6. Results

Since we were not provided with an explicit training dataset. We collected data from various sites on our own and used it as context to be parsed. We did not use any dataset to train our model. Hence the results obtained might be low in their accuracy rate. We tested the outputs our model produced using the test data that FIRE 2024 provided, and the following outcomes were attained.

From Table 1, we can infer that BLEU and ROUGE-1 scores are low. BLEU (Bilingual Evaluation Understudy) score measures the quality of machine-generated answers by comparing them to a set of reference answers. It computes the overlap of n-grams (sequences of n words) between the predicted and reference answers, with higher n-gram precision indicating better alignment. The brevity penalty (BP) is a component of the BLEU score designed to penalize machine-generated translations or answers that are shorter than the reference text.

ROUGE-1 score measures the overlap of unigrams between the generated answer and the reference answer. It provides a measure of how well the predicted answer covers the important words found in the reference. If the generated answer is shorter than the reference, it typically lowers recall, potentially lowering the overall ROUGE score. ROUGE-2 score measures the overlap of bigrams (pairs of consecutive words) between the generated answer and the reference answer. Since bigrams focus on word pairs, this metric evaluates both content and some level of fluency or coherence in the generated text. Our dataset consists of only a limited information due to the token limit in BERT. This makes answer generated by our system smaller in length when compared to the original answers due to which our BLEU, ROUGE-1 and ROUGE-2 scores are low.

Table 1
Results on testing dataset

Evaluation Metric	Score
BLEU	0.008
ROUGE-1	0.19
ROUGE-2	0.05

Question: *This patient likely has pancreatic cancer.Can you provide information on the role of BRCA mutations in pancreatic cancer and potential implications for treatment?*
Answer: *The mutations BRCA1 and BRCA2 increase a person's lifetime risk of developing pancreatic cancer.Their normal function is to repair damage to DNA, but when BRCA1 or BRCA2 is mutated and doesn't work correctly, the accumulation of unrepaired DNA damage can ultimately lead to unregulated cell growth, or cancer.*

Though the model can extract answers from the given text, when the chunk size is too large for the model to handle, it is not able to extract the answer from the context. This reduces the reliability of the system.

7. Conclusions

The applications of large language models (LLMs) like BERT are vast, yet their use in the medical field remains in its developmental stages. To address this gap, we have used a BERT model specifically tailored to answer questions related to gastrointestinal cancer. Our approach involved preparing the dataset for various gastrointestinal cancers and selecting the relevant context based on the specific cancer mentioned in the query. This was followed by tokenizing the input and extracting the appropriate answers.

Looking ahead, this model can be trained on larger datasets to enhance its performance further. Additionally, it can be adapted to extract data directly from medical resources, thereby improving its accuracy and reliability.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4o in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] G. S. Dunham, M. G. Pacak, A. W. Pratt, Automatic indexing of pathology data, Journal of the American Society for Information Science 29 (1978) 81–90.
[2] S. Datta, E. V. Bernstam, K. Roberts, A frame semantic overview of nlp-based information extraction for cancer-related ehr notes, Journal of biomedical informatics 100 (2019) 103301.
[3] G. Burger, A. Abu-Hanna, N. de Keizer, R. Cornet, Natural language processing in pathology: a scoping review, Journal of clinical pathology 69 (2016) 949–955.
[4] B. Seifert, G. Rubin, N. de Wit, C. Lionis, N. Hall, P. Hungin, R. Jones, M. Palka, J. Mendive, The management of common gastrointestinal disorders in general practice: a survey by the european society for primary care gastroenterology (espcg) in six european countries, Digestive and Liver Disease 40 (2008) 659–666.

[5] K. Holtedahl, P. Vedsted, L. Borgquist, G. A. Donker, F. Buntinx, D. Weller, T. Braaten, P. Hjertholm, J. Månsson, E. L. Strandberg, et al., Abdominal symptoms in general practice: Frequency, cancer suspicions raised, and actions taken by gps in six european countries. cohort study with prospective registration of cancer, Heliyon 3 (2017).

[6] Q. Zhou, C. Liu, Y. Duan, K. Sun, Y. Li, H. Kan, Z. Gu, J. Shu, J. Hu, Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation, Frontiers in Medicine 11 (2024). URL: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1392555. doi:10.3389/fmed.2024.1392555.

[7] G. Eysenbach, et al., The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers, JMIR Medical Education 9 (2023) e46885.

[8] J. R. Mitchell, P. Szepietowski, R. Howard, P. Reisman, J. D. Jones, P. Lewis, B. L. Fridley, D. E. Rollison, A question-and-answer system to extract data from free-text oncological pathology reports (cancerbert network): Development study, J Med Internet Res 24 (2022) e27210. URL: https://www.jmir.org/2022/3/e27210. doi:10.2196/27210.

[9] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, C. Zhang, Bitenet: bidirectional temporal encoder network to predict medical outcomes, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 412–421.

[10] A. Lahat, E. Shachar, B. Avidan, B. Glicksberg, E. Klang, Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: Are we there yet?, Diagnostics 13 (2023). URL: https://www.mdpi.com/2075-4418/13/11/1950. doi:10.3390/diagnostics13111950.

[11] J. Yuan, P. Bao, Z. Chen, M. Yuan, J. Zhao, J. Pan, Y. Xie, Y. Cao, Y. Wang, Z. Wang, et al., Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers, The Innovation 521 (2023).