CSE 842

Project Report

EMRQA: A Valuable Resource for Medical Question Answering in Electronic Health Records

Team Members: Daniyal Dar, Muhammad Talha, Souradip Biswas

Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed remarkable advancements, enabling machines to understand and interpret human language with unprecedented accuracy. This project aims to harness the power of NLP to develop an intelligent Medical Question-Answering (QA) system, facilitating seamless access to accurate and relevant medical information.

Motivation

Healthcare information is vast and constantly evolving, making it challenging for individuals to navigate and obtain accurate answers to their medical queries. The motivation behind this project lies in addressing the need for a user-friendly, intelligent system that can bridge the gap between individuals seeking medical information and the wealth of knowledge available in the field.

Problem definition

The project aims to train and evaluate machine learning and deep learning models for the task of answering questions related to electronic medical records. These questions can range from specific medical queries (e.g., "What medications was the patient prescribed?") to more general inquiries (e.g., "What is the patient's medical history?"). The input space for this project primarily consists of electronic medical records, which are composed of unstructured textual data. These records include patient histories, medical notes, diagnosis reports, treatment plans, and other clinical documents. The output space involves generating answers to questions posed about EMR. For example, if the input is a question like "What treatments were administered to the patient?" or "What is the patient's current medical condition?" the output space includes the corresponding answers extracted from the EMR.

Solution

Machine learning serves as an optimal approach to address the complexity of medical question answering within electronic health records due to its adaptability and capacity for pattern recognition within unstructured textual data. The use of machine learning models, particularly transformer models, allows for the extraction of nuanced information from the diverse and extensive EMR sources. These models excel in learning intricate relationships within the text, enabling them to comprehend the context, semantics, and syntax embedded within medical records. Additionally, the scalability of machine learning algorithms facilitates continuous learning and adaptation to the evolving nature of healthcare information. By leveraging these techniques under carefully designed settings—such as training on annotated datasets, fine-tuning models based on domain-specific knowledge, and employing attention mechanisms to focus on relevant information within records—machine learning becomes a pivotal tool in generating accurate, context-aware responses to diverse medical queries posed within electronic health records.

Machine Learning Design

The ML design is basically a pipeline for generative question-answering using a transformer-based model, specifically T5 (Text-To-Text Transfer Transformer).

a) Model Architecture: We are using a pre-trained T5 model for conditional text generation (T5ForConditionalGeneration). The model is fine-tuned on a dataset for the task of question-answering**.**

b) Data Handling: We have created a custom dataset class (QADataset) to handle the input data. The dataset consists of contexts, questions, and answers.

c) Training Loop: The training loop involves using a DataLoader to efficiently load batches of data. You are training the model using the AdamW optimizer.

Learning Examples and Features Specification:

a) Learning Examples: Learning examples consist of contexts, questions, and answers. Examples are used to train the model to generate appropriate answers given a context and question.

b) Features Specification: The input to the model includes both the context and the question. These inputs are tokenized using the T5 tokenizer, and the resulting token IDs are used as input to the model. Padding is applied to ensure uniform input size.

Vector Representations:

T5 models, including the one we're using, employ word embeddings to represent words in vector form. The transformer architecture utilizes attention mechanisms to capture relationships between words and generate contextual embeddings. The model processes input sequences and produces output sequences as a series of token IDs. During training, the model learns to adjust its parameters to minimize the difference between generated and target sequences.

Model

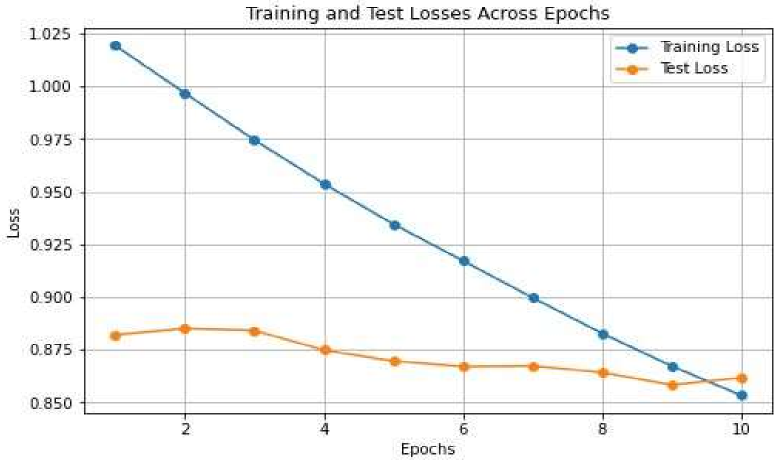
Our T5 model for QA follows the text-to-text paradigm. It frames the QA task as a text generation problem, where the model is trained to generate the answer text given the concatenated question and context as input.

The Architecture Component

Input Encoding (Embedding) Layer: The input, which consists of the question and context, undergoes tokenization and embedding. Each token is converted into a dense vector representation (embedding). These embeddings capture the semantic meaning of words and their relationships.

Transformer Encoder Layers: T5 employs transformer encoder layers for processing the input. These layers use self-attention mechanisms to capture contextual information. They allow the model to weigh the importance of different words in the input sequence, especially useful for understanding relationships between question and context.

Task-Specific Output Layer: The final layer of the model is designed for the specific QA task. It could be a linear layer for span prediction, where the model predicts the start and end positions of the answer within the context. Alternatively, it might be a classification layer if the task involves selecting an answer from predefined options.



Working of the model:

Tokenization and Embedding: The input question and context are tokenized into subword units, and these tokens are embedded into dense vectors.

Input Concatenation: The embeddings for the question and context are concatenated into a single input sequence. Special tokens are often used to indicate the separation between the question and context.

T**ransformer Processing:** The concatenated input sequence is then processed through multiple transformer encoder layers. These layers attend to different parts of the input, allowing the model to understand the context and relationship between the question and context.

Task-Specific Prediction: The output from the transformer layers is fed into the task-specific output layer. For QA, this layer might produce logits for the start and end positions of the answer span or probabilities for different answer options.

Training and Fine-Tuning: During training, the model is optimized by adjusting its parameters to minimize a loss function, which measures the dissimilarity between the predicted and actual answers. Fine-tuning can be performed on a task-specific dataset to adapt the pre-trained T5 model to the specifics of the QA task.

Data and Tools

The size of the dataset is 44771 samples. Firstly, For the training part, we used 10000 top data examples due limited computational resources.
Each example has Question, Context and Answer. We implemented in pytorch using Transformers library mainly. Apart from that Sckitlearn and Pandas were also being utilized for our project.

Results

We conducted an iterative approach to training and tes8ng. Initally, we trained our model on approximately 1000 data points to gauge its performance. The model exhibited a commendable 98% BLEU score at this stage. Subsequently, in an effort to enhance its generalizability, we extended our training set to 10,000 data points. This broader training corpus yielded an impressive 87.33% BLEU score, signifying a robust performance tailored to our specific problem while ensuring enhanced generalizability with a good score. To explore alternative approaches and compare the performance of our Medical Question-Answering system, we extended our experimentation to include the use of BERT (Bidirectional Encoder Representations from Transformers), a powerful transformer-based model known for its contextual understanding of language. With 10,000 data points, it achieved a BLEU score of approximately 62.5%.

Related work

Yes, machine learning and deep learning techniques have been employed to tackle the problem of medical question answering in EMRs. Researchers have explored the application of ClinicalBERT, a domain-specific variant of BERT, for clinical question answering. Attention mechanisms, popularized by transformer models, have been incorporated into medical QA systems to improve the model's focus on relevant parts of the input text.

Conclusion

As a conclusion, our iterative approach to model training and testing, incorporating both T5 and BERT architectures, has provided valuable insights into the development of a Medical Question-Answering system. The T5 model demonstrated high BLEU scores, indicating a strong understanding of medical language and effective response generation. It took us around five hours to train the T5 model.