Medical Domain
Question Answering System

NLP Project Report

FALL 2022

By
Meghana
Sravani
12/12/2022

Abstract

While question-answering systems have gained popularity across many industries, the medical field has proven difficult due to the wealth of specialized knowledge. As more and more sizable pretrained language models emerge, retrieval-based approaches have shown promise. The goal of this project is to develop a retrieval-based medical question answering system. To do this, we leverage huge language models and graphs to extend our knowledge. First, we effectively use Elasticsearch to get a large yet coarse set of responses. Then, using knowledge graphs and named entity recognition to take advantage of the relationship between the entities in the query and answer, we combine semantic matching with pretrained language models to get a fine-grained ranking. We offer a thorough study of both this dataset and the medical reading comprehension test in this project. Our qualitative research reveals that we QA answers are frequently insufficient and QA questions can frequently be satisfactorily answered without the use of domain knowledge. We compared to the model trained on the entire dataset. the performance of our model is close to human expert's performance, and BERT models do not beat the best performing base model.

Introduction

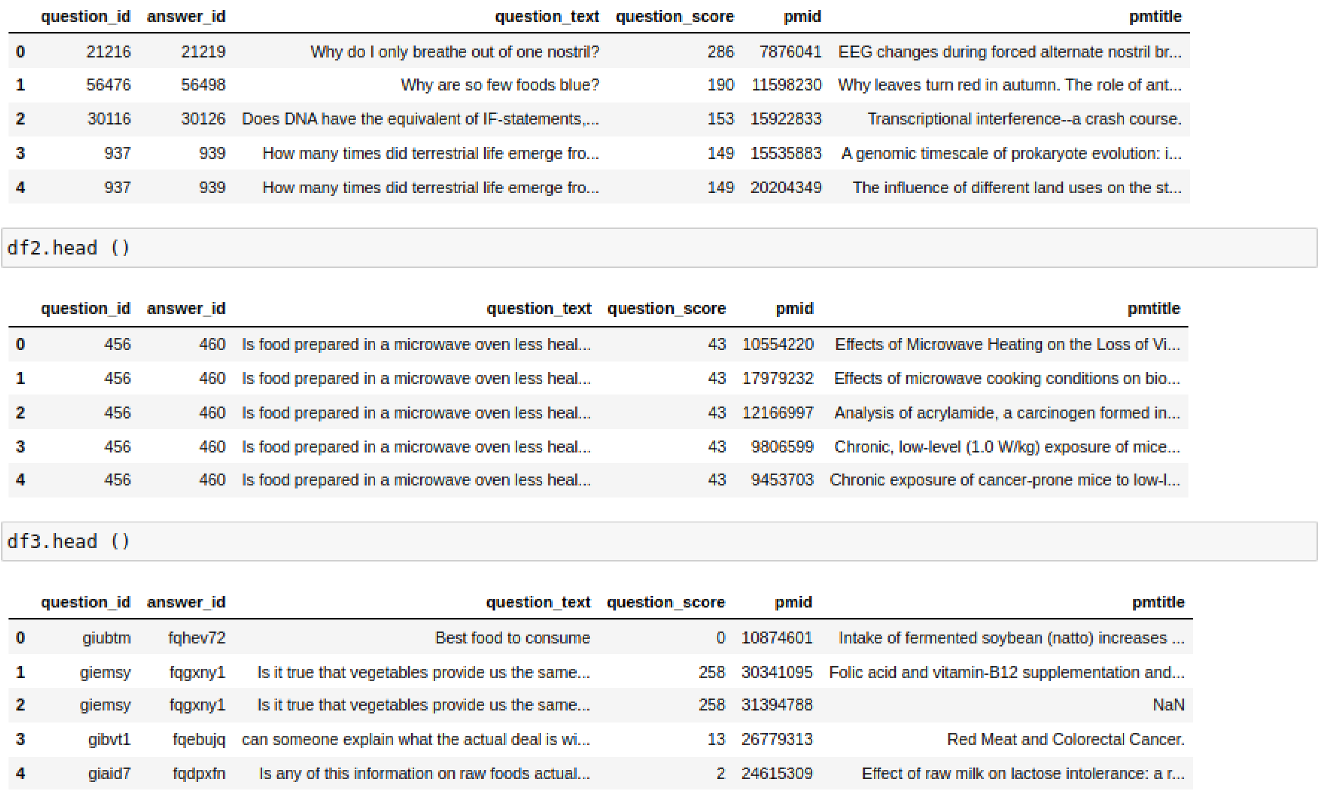
Question Answering (QA) aims to automatically answer questions asked by humans based on external sources, such as Web, knowledge base and free text. As an important type of QA, reading comprehension intends to answer a question after reading the passage. Recently, the release of large-scale RC datasets, such as CNN & Daily Mail, Stanford QuestionAnswering Dataset (SQuAD) makes it possible to solve RC tasks by building deep neural models. More recently, contextualized word representations and pretrained language models, such as ELMo, GPT, BERT, have been demonstrated to be very useful in various NLP tasks including RC. By seeing diverse contexts in large corpora, these pretrained language models can capture the rich semantic meaning and produce more accurate and precise representations for words given different contexts. Even a simple classifier or score function built upon these pretrained contextualized word representations perform well in extracting answer spans. Biomedical and Clinical QA. Due to the lack of large-scale annotated biomedical or clinical data, QA and RC systems in these domains are often rule-based and heuristic feature-based.

In recent years, BioASQ challenges proposed the Biomedical Semantic QA task, where the participants need to respond to each test question with relevant articles, snippets and exact answers. Suster and Daelemans use summary points of clinical case reports to build a large-scale cloze-style dataset (CliCR), which is similar to the style of CNN & Daily Mail dataset. PubMedQA, which extracts question-style titles and their corresponding abstracts as the questions and contexts respectively. A few QA pairs are annotated by human experts and most of them are annotated based a simple heuristic rule with "yes/no/maybe". Due to the great power of contextualized word representations, pretrained language models also have been introduced to biomedical and clinical domain, e.g., BioELMo, BioBERT, and ClinicalBERT. They adopt similar architectures of the original models but pretrained on the medical and clinical corpus, such as PubMed articles and MIMIC-III clinical notes.

Method

This project aim is to develop a question answeting system in medical domain. For the the implementation of a BERT-based model which returns "an answer", given a user question and a passage which includes the answer of the question. For this question answering task, we used three differents medical domain datasets, MedQuAD, MEDIQA2019, BiQA and Combined them for fine tuning a model that is trained on SQUAD 2.0. Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. BiQA is a dataset that has data from Biology, medical scinces, nutrition domains with number of samples, 6681, 3014, 4099 respectively. MedQuAD – 47k. MEDIQA2019 - 10k BiQA – 13k samples. I started with the BERT-base pretrained model “bert-base-uncased” and fine-tune it to have a question answering task. For Question Answering we use the BertForQuestion Answering class from the transformers library. BERT is a Bidirectional Encoder Representations from Transformers. It is one of the most popular and widely used NLP models. BERT models can consider the full context of a word by looking at the words that come before and after it, which is particularly useful for understanding the intent behind the query asked. Because of its bidirectionality, it has a deeper sense of language context and flow and hence, is used in a lot of NLP tasks nowadays. This class supports fine-tuning, but for this example we will keep things simpler and load a BERT model that has already been fine-tuned for the SQuAD benchmark. The transformers library has a large collection of pre-trained models which we can reference by name and load easily. Apparently the vocabulary of this model is identicaly to the one in bert-base-uncased. we can load the tokenizer from bert-base-uncased and that works just as well. After trying various bert architectures, bert, distilbert, albert, distilroberta and distilroberta_extra_linear, we found distilroberta_extra_linear gave better performance over the given task in medical domain dataset.

The following is the data samples from BioQA dataset



	question_id	answer_id	question_text	question_score	pmid	pmtitle
0	21216	21219	Why do I only breathe out of one nostril?	286	7876041	EEG changes during forced alternate nostril br...
1	56476	56498	Why are so few foods blue?	190	11598230	Why leaves turn red in autumn. The role of ant...
2	30116	30126	Does DNA have the equivalent of IF-statements,...	153	15922833	Transcriptional interference--a crash course.
3	937	939	How many times did terrestrial life emerge fro...	149	15535883	A genomic timescale of prokaryote evolution: i...
4	937	939	How many times did terrestrial life emerge fro...	149	20204349	The influence of different land uses on the st...

```
df2.head ()
```

	question_id	answer_id	question_text	question_score	pmid	pmtitle
0	456	460	Is food prepared in a microwave oven less heal...	43	10554220	Effects of Microwave Heating on the Loss of Vi...
1	456	460	Is food prepared in a microwave oven less heal...	43	17979232	Effects of microwave cooking conditions on bio...
2	456	460	Is food prepared in a microwave oven less heal...	43	12166997	Analysis of acrylamide, a carcinogen formed in...
3	456	460	Is food prepared in a microwave oven less heal...	43	9806599	Chronic, low-level (1.0 W/kg) exposure of mice...
4	456	460	Is food prepared in a microwave oven less heal...	43	9453703	Chronic exposure of cancer-prone mice to low-l...

```
df3.head ()
```

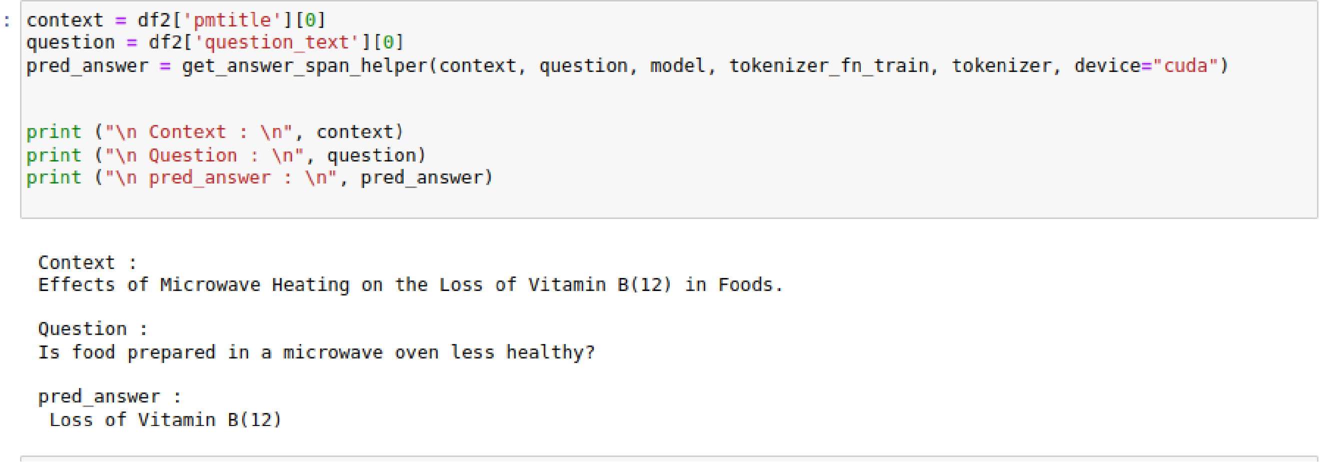
	question_id	answer_id	question_text	question_score	pmid	pmtitle
0	giubtm	fqhev72	Best food to consume	0	10874601	Intake of fermented soybean (natto) increases ...
1	giemsy	fqgxny1	Is it true that vegetables provide us the same...	258	30341095	Folic acid and vitamin-B12 supplementation and...
2	giemsy	fqgxny1	Is it true that vegetables provide us the same...	258	31394788	NaN
3	gibvt1	fqebujq	can someone explain what the actual deal is wi...	13	26779313	Red Meat and Colorectal Cancer.
4	giaid7	fqdpxfn	Is any of this information on raw foods actual...	2	24615309	Effect of raw milk on lactose intolerance: a r...

Our Implementation

- First Step - Evaluating performance of the SQUAD finetuned BERT-QA model on medicalQA.
- Second Step - SQUAD finetuned BERT-QA model on medicalQA.
- Evaluating model performance on Fine-tuned model with medicalQA dataset

Training

The trained model is saved as pytorch checkpoint file, this model can be trained further with additional dataset for a better performance or can be used for testing. After the training we tested the dataset further BERT uses wordpiece tokenization. In BERT, rare words get broken down into subwords/pieces. We use pretrained tokenizer to clean and tokenize the sentences and create word embeddings. We trained our model using pre-trained distilbert model for 10 epochs and evaluated the and measured the performance on validation and test dataset. The evaluation metrics can be found in results.



```
: context = df2['pmtitle'][0]
question = df2['question_text'][0]
pred_answer = get_answer_span_helper(context, question, model, tokenizer_fn_train, tokenizer, device="cuda")


print ("\n Context : \n", context)
print ("\n Question : \n", question)
print ("\n pred_answer : \n", pred_answer)
```

```
Context :
Effects of Microwave Heating on the Loss of Vitamin B(12) in Foods.

Question :
Is food prepared in a microwave oven less healthy?

pred_answer :
 Loss of Vitamin B(12)
```

Evaluation

We adopt our model two metrics including Exact Match (EM) and F1 scores to evaluate our model. The EM score determines the percentage of predictions that perfectly match the ground truth answer, and the F1 score demonstrates the average overlap between the prediction and the ground truth answer.

Measure	Test Scores on MedQA
Exact Match Score	61.20
F1	67.39

Conclusion

Models trained on the general domain dataset do not perform well on the domain-specific datasets. To adapt to the medical domain, task-driven fine-tuning with medical domain-specific QA dataset is one of the most important steps. Medical Domain adaptation by Language Model training with limited data (with only available paragraphs or clinical notes from QA dataset) gives a marginal improvement in performance. Fine-tuning the BERT-QA model with a large Medical domain QA dataset before fine-tuning on domain-specific QA dataset can prove helpful when the domain-specific dataset is limited.