Medical Domain Question Answering System

Meghana

Sravani

12/05/2022

Project Description

This project presents our experiments in developing a question-answering (QA) system which aims at replying to questions on services offered by a website.

Our system is an instance of closed-domain QA, which works on a document collection restricted in subject and volume.

In closed-domain QA, correct answers to a question may often be found in only very few documents; the system does not have a large retrieval set abundant of good candidates for selection.

Moreover, if the QA system is to be used for answering questions from a company's clients, it should accept complex questions, of various forms and styles.

The system should then return a complete answer, which can be long and complex, because it has to, e.g., clarify the context of the problem posed in the question, explain the options of service, give instructions or procedures, etc.

Question Answering?

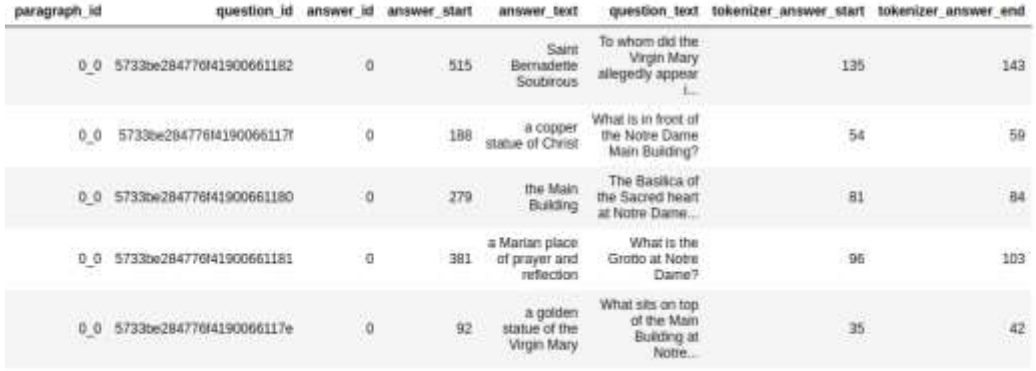
- An automated question-answering system can replace human efforts.
- Building a reliable and efficient QA system is a challenging task. Its performances are directly related to the quality of integrated tools for finding the answers and the depth of all involved NLP resources.
- When dealing with such a system and developing it, many arising questions should be answered, and multiple objectives should be fulfilled.
- In this project we build a flexible QA model that can adapt to different closed domains and train on their corpora.

Goal

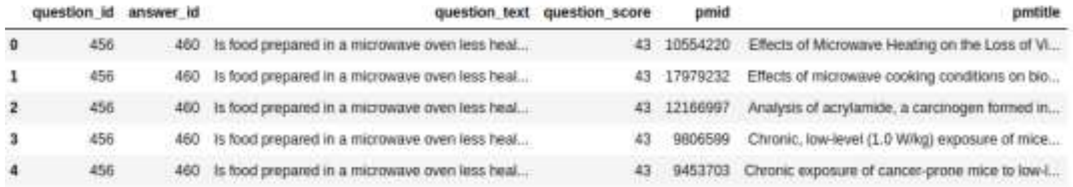
- What would be the approach to Question Answering task where input context or paragraph is n-times bigger or smaller than 512?
- How to adapt the BERT model for domain-specific QA dataset with a limited amount of domain-specific corpus (only product documents or only clinical notes)?
- Does replacing the placeholders in BERT’s vocabulary with the frequent domain-specific words help?
- How much training data is needed to achieve decent accuracy?

Dataset

- SQUAD-2
- Combined multiple Medical QA datasets for fine tuning
  - MedQuAD – 47 k
  - MEDIQA2019 - 10k
  - BiQA – 13k samples



paragraph_id	question_id	answer_id	answer_start	answer_text	question_text	tokenizer_answer_start	tokenizer_answer_end
0_0	5733be284776f41900661182	0	515	Saint Bernadette Soubirous	To whom did the Virgin Mary allegedly appear i...	135	143
0_0	5733be284776f4190066117f	0	188	a copper statue of Christ	What is in front of the Notre Dame Main Building?	54	59
0_0	5733be284776f41900661180	0	279	the Main Building	The Basilica of the Sacred heart at Notre Dame...	81	84
0_0	5733be284776f41900661181	0	381	a Marian place of prayer and reflection	What is the Grotto at Notre Dame?	96	103
0_0	5733be284776f4190066117e	0	92	a golden statue of the Virgin Mary	What sits on top of the Main Building at Notre...	35	42



	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...

Implementation

- BERT architecture
- 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

Evaluation Methodology

- 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.

Results

- 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	61.20
f1	67.39
total	5000

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.

References

- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760988.pdf
- https://towardsdatascience.com/the-definitive-guide-to-bi-directional-attention-flowd0e96e9e666b
- https://web.stanford.edu/class/cs224n/materials/CS224N_PyTorch_Tutorial.html
- https://arxiv.org/pdf/2005.00574.pdf
- https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
- https://ieeexplore.ieee.org/document/9762943