Question Answering System Based on BERT for Cancer Data

Instructor Mr. Syed Jawad Shah
Presenters: Charan Tej, Siva Kumar, Abdel

School of Computing Engineering, UMKC

May 1, 2020

(Knowledge Discovery Management)

Abstract

- Question Answering (QA) system is an information retrieval system during which an on the spot answer is predicted in response to a submitted query, instead of a collection of references which will contain the answers.
- he essential idea of QA systems in Language Processing (NLP) is to supply accurate answers to the questions for the general public or medical practitioners who could ask inquiries to the system. The training may require detailed information of every kind of disease during a specific domain.
- So, these facts reflect the urgent and genuine need of an information retrieval system that accepts the queries from medical practitioners in natural language and returns the answers quickly and efficiently.

Introduction I

Diseases and health condition have better outcomes if detected early on

1. Understanding of different symptoms.
2. Possible health outcomes/illnesses.
3. Life-threatening conditions

QA Systems comes into play to assist individuals navigate the literature on their specific symptoms.

1. Understanding of different symptoms.
2. Possible health outcomes/illnesses.
3. Life-threatening conditions

QA System I

1. Information Retrieval system that utilizes Natural Language Processing (NLP) to accurately respond to queries submitted by the user.
2. It is built to extract information for public and health practitioners, based on the available knowledge in the heath domains.

Related Work I

QA systems in health care are currently the center of attention
Less advanced in the Chinese language due to:

1. Difficulty of Chinese text processing
2. No large scale datasets
3. Newley introduced semantic clustered improved the model performance by 5.5% with mean precision of 4.9%

QA Systems Utilized in the Clinical Decision Support Systems, which is based on the EMRs of the patients.

1. Challenged with the grammar issue within the clinical notes
2. Uses Name Entity Recognition and tokenization to disease symptoms

Related Work II

3. Produces accuracy score of 0.79.

Objective Structured Clinical Examinations (OSCEs) for medical students

1. Evaluated the medical students competences doing clinical tasks with standardized patients.
2. Deep leaning framework is deployed to help SPs communication.
3. NN leans the domain embedded words.
4. Long short term memory utilizing CNN model work to generate answers.
5. Achieved accuracy rate of 81%.

Dataset I

1. Web scraping was used to extract the data from different web sources.
2. We intent to pre-process the data without removing stop words where BERT system identify the context from L-R and R-L.

Dataset II

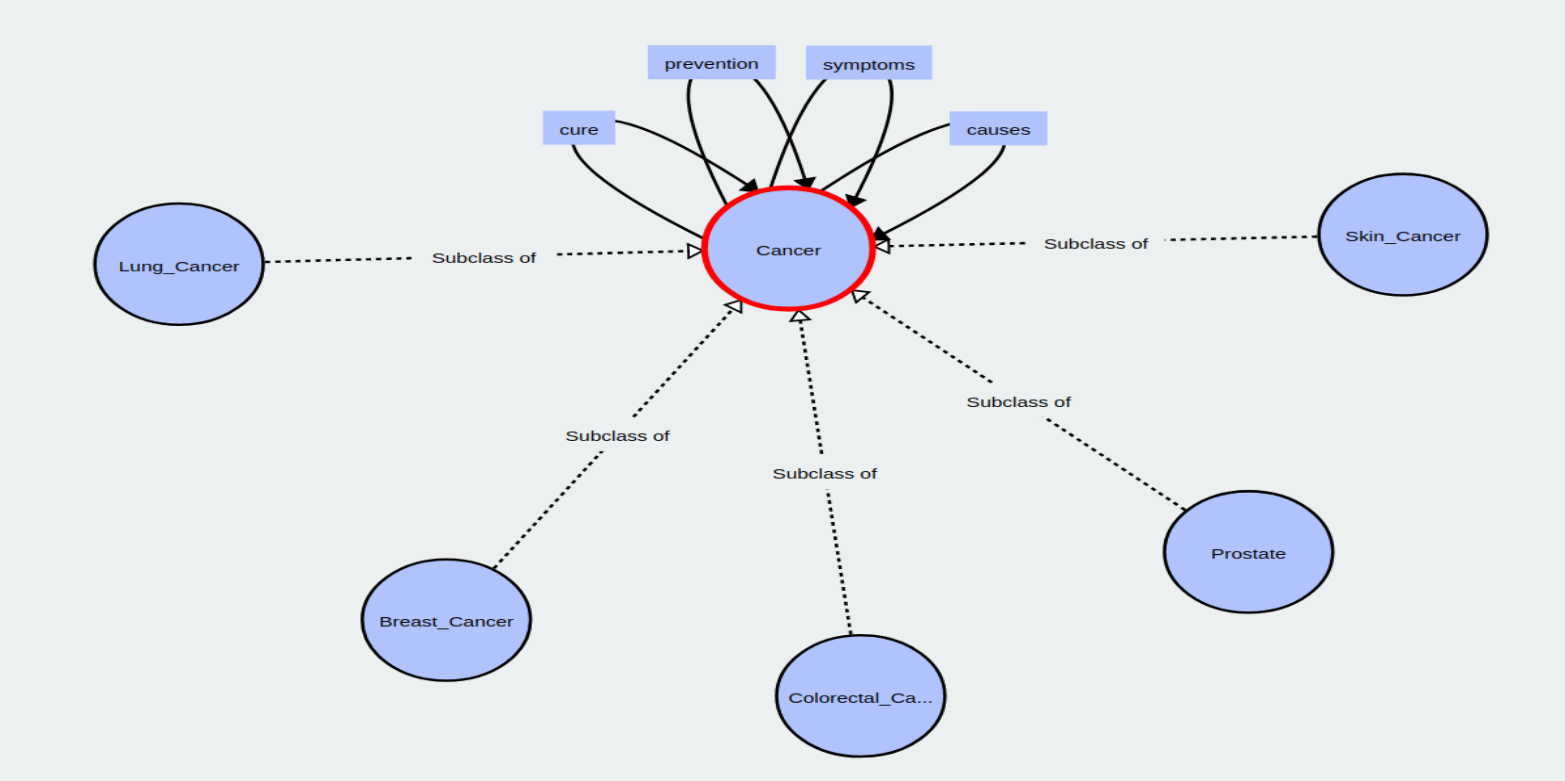

Figure: Properties and Types of Cancer Data using OWL Diagram

Proposed Models I

- BERT itself is huge model with about 340 million parameters and pre-trained model, which would required a higher level of computing power to train the model.
- Since, aim of the paper is to be reachable for the low end computing machine currently available at people disposal for the initial analysis we propose use of distilled BERT initally.
- Later we consider to use BERT architecture for Question Answer System in Fig[6].

Proposed Models II

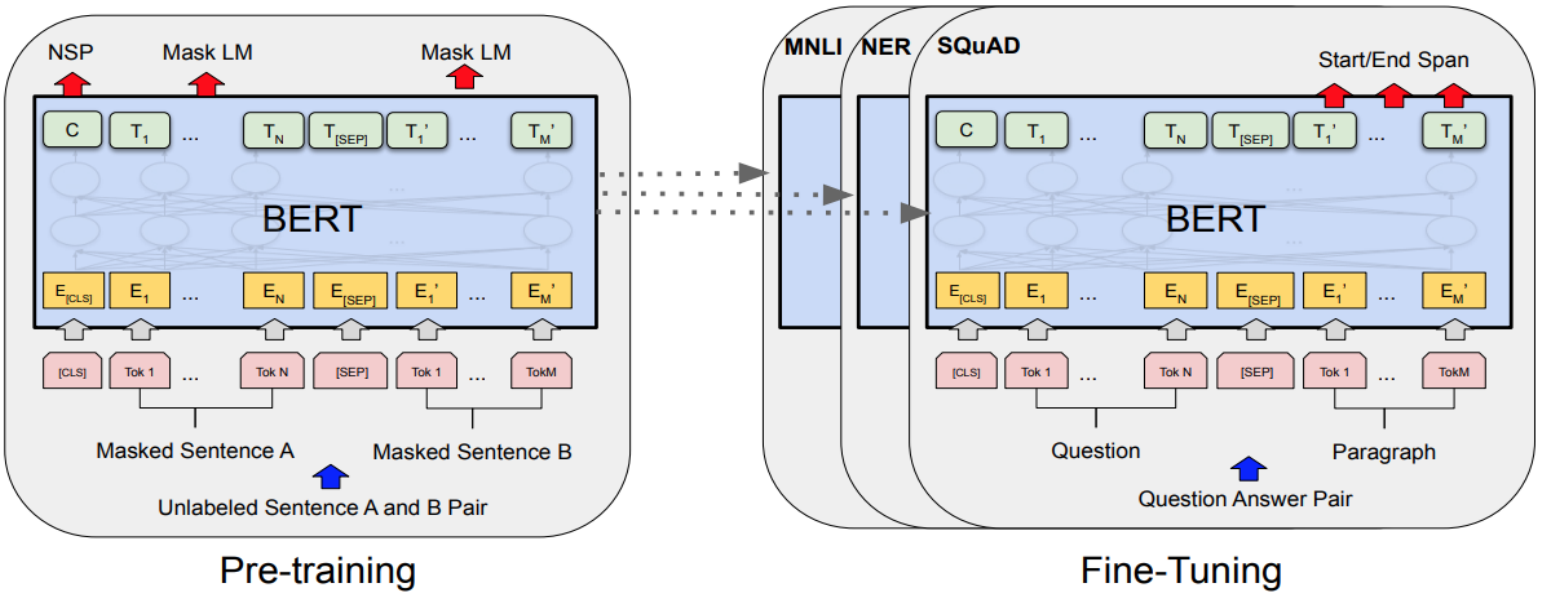

Figure: Architecture of BERT

Topic Modeling I

1. Data we collected from different web sources using web scraping and the data is categorized different types of cancers using topic modelling.
2. Topic modeling is a type of statistical modeling for discovering the abstract **topics** that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic.
3. LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. Given a dataset of documents, LDA backtracks and tries to figure out what topics would create those documents in the first place.

Topic Modeling II

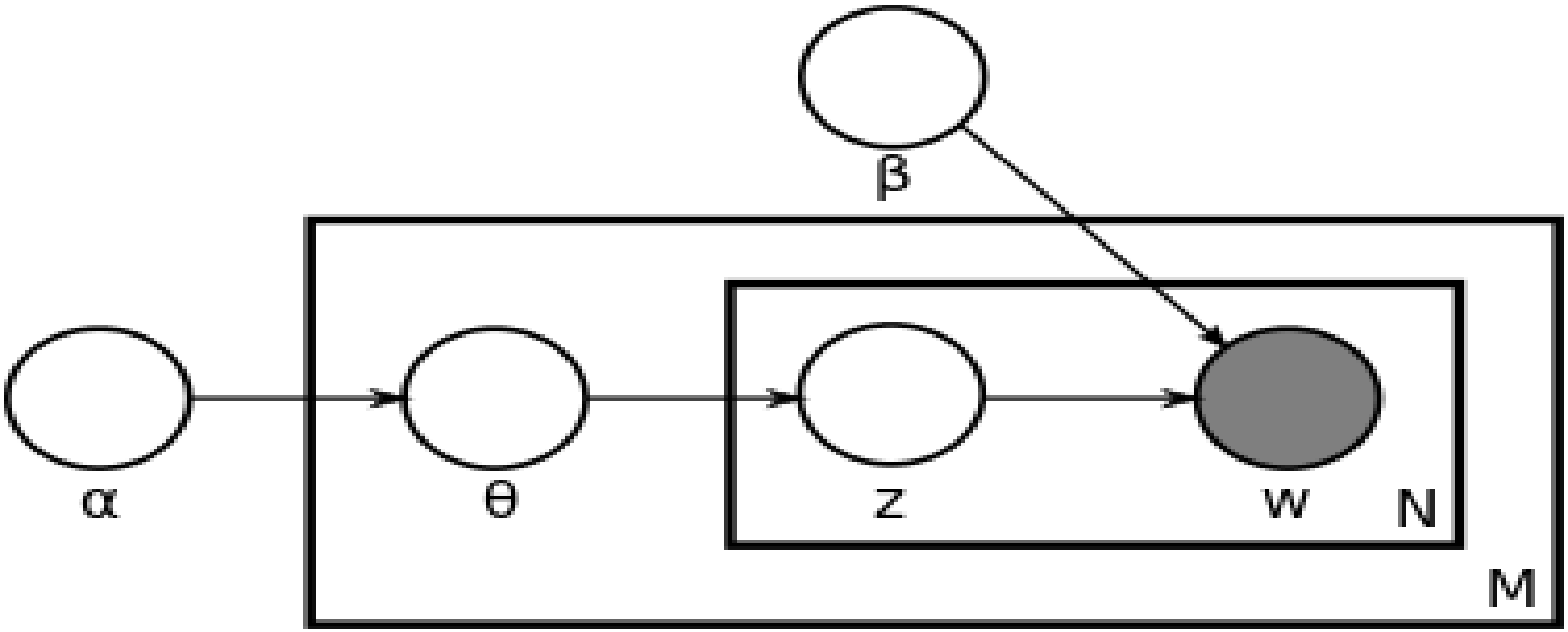

Figure: Architecture of LDA

Topic Modeling III

```
--------------------------------------------------
Topic modelling for the files breast_cancer.txt
topic 1 = (0, '0.038*"the" + 0.036*"breast" + 0.036*"cancer"')
topic 2 = (1, '0.002*"cancer" + 0.002*"breast" + 0.002*"the"')
--------------------------------------------------
Topic modelling for the files colorectal_cancer.txt
topic 1 = (0, '0.002*"cancer" + 0.002*"to" + 0.002*"the"')
topic 2 = (1, '0.042*"cancer" + 0.036*"the" + 0.034*"to"')
--------------------------------------------------
Topic modelling for the files lung_cancer.txt
topic 1 = (0, '0.041*"cancer" + 0.035*"lung" + 0.034*"the"')
topic 2 = (1, '0.002*"lung" + 0.002*"the" + 0.002*"your"')
--------------------------------------------------
Topic modelling for the files prostate_cancer.txt
topic 1 = (0, '0.039*"the" + 0.037*"cancer" + 0.036*"prostat"')
topic 2 = (1, '0.002*"to" + 0.002*"the" + 0.002*"cancer"')
--------------------------------------------------
Topic modelling for the files skin_cancer.txt
topic 1 = (0, '0.051*"skin" + 0.045*"the" + 0.039*"of"')
topic 2 = (1, '0.003*"skin" + 0.003*"the" + 0.003*"your"')
```

Figure: Topic modeling for Cancer Data

Metrics

To evaluate the performance of QA System implemented by BERT Model with our custom data, Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) which helpful to evaluate the score for comparing a candidate translation of text to one or more reference translations and evaluating automatic summarizing of texts as well as machine translation respectively.

Experiments I

1. we used BERT fine tuned model for question answer system and performed experiments on well cancer data sets which makes use of transformers that implements BERT Question Answer System.
2. The precision and recall scores of BLEU () and ROUGE models on sample cancer data. In these experiments, we achieved precision and recall scores as 97.07% and 97.79% respectively.
3. In ROUGE,we have achieved 0% results because the it is considering uni-grams instead of bi-grams.
4. In addition to that, we collected the data using web crawling to collect the data of size 70GB. It is very difficult to process such a huge data. So we proceed to implement web scraping.

Experiments II

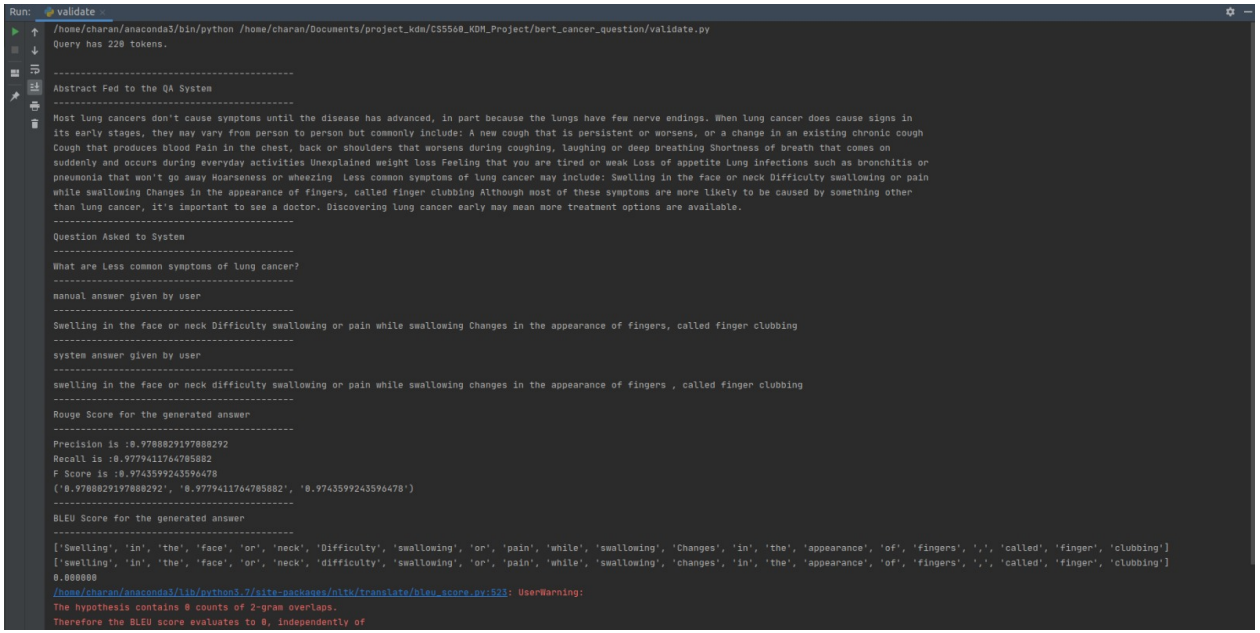

Figure: Performance of QA System using BLEU and ROUGE

Experiments III

```
what is breast cancer?
cancer that forms in the cells of the breasts


what is the acceptable limit for alcohol?
no more than one drink a day


who are more likely to get breast cancer?
women


what is invasive lobular carcinoma?
glandular tissue called lobules


what is estimate of inherited breast cancer?
about 5 to 10 percent


what is carcinoembryonal antigen?
a specific tumor marker


what is treatment for colorectal cancer?
surgery


what is sigmoidoscopy?
a third possible screening examination known as a sigmoidoscopy , where only the lower part of the large intestine is examined


what is the cause of colorectal cancer?
when cells in the mucous lining of the intestine change ( mutate ) and then multiply out of control


what age is risky for colorectal cancer?
50
```

Figure: System Generated Answers for Cancer Data using BERT

Experiments I

1. In addition to that, we collected the data using web crawling to collect the data of size 70GB. It is very difficult to process such a huge data. So we proceed to implement web scraping.

Conclusion I

1. Question answer system for cancer data, we implemented which makes use of a pre trained model called BERT that inferred good performance compare to any other models exists.
2. Question Answer system we verified using BERT and manually as well. It shows the results almost same for most of the questions we posted and accuracy levels that we verified using ROUGE and BLEU metrics.
3. QA systems has great potentials in the healthcare applications. BERT has proven to be able to achieve a higher score of precision allowing for achieving NLP tasks in a better an more efficient ways.
4. For Execution https://www.youtube.com/watch?v=IIF37oJzplg&feature=youtu.be
5. For Github Link
https://github.com/charantej224/CS5560_KDM_Project/tree/master/bert_cancer_question