Question Answering System Based on BERT for Cancer Data

Charan Tej Thota
Universiyt of Missouri
Kansas City-64110
Email: ctpcc@mail.umkc.edu

Siva Kumar Buddi
University of Missouri
Kansas City-64110
Email: sbfyw@mail.umkc.edu

Elfagir Abdelmoneim
University of Missouri
Kansas City-64110
Email: aaeq85@mail.umkc.edu

***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. the 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. This technique could help patients identify diseases in preliminary stages which could help save lives of individuals.**

***Index Terms*—BERT, Natural Language Processing,**

I. Introduction

With Advancements in the Medical industry, most of the life cautionary diseases are treatable if diagnosed in early phases. It's not common for the general people to have a good understanding of possible symptoms of life cautionary diseases, apparently such problems can be easily misinterpreted for non-life threatening diseases. We aim to solve this knowledge gap of life threatening diseases to general people using advanced artificial intelligence. A Question and Answer system(QA) can mimic a doctor to help patients under their initial symptoms and possible causes which can further provoke patients to get tested with a physician. [1] [2] [3] [4] [5] [6]

Question Answering (QA) system is an information retrieval system in which a straight forward output is expected in response to a submitted query, instead of a group of references that may contain the several answers. The basic idea of QA systems in Natural Language Processing (NLP) is to supply accurate answers to the queries for the public or medical practitioners who could ask inquiries to the system.The training may require detailed information of each sort of disease in a specific domain. So, these facts reflect the emergency and genuine need of an information retrieval system that accepts the queries from medical practitioners in natural language and returns the answers quickly, effectively and efficiently. This system could help patients identify diseases in preliminary stages which could help save lives of individuals.

Bidirectional Transformer encoder(BERT)[7] has been state-of-art technique which has beaten its peers by a reasonable margin. The design of BERT lies in understanding the context from both the directions of the token forward and backward. It has proven to have more efficient technique which is a result of using transfer learning. We intend to utilize this model which is not applied so far on the cancer data for question answering based task.

II. Related Work

Medical field and clinical research domain in constant need for continuous development as it functions on the bases of saving the peoples' lives. In recent years, Question Answering (QA) systems development for medical and clinical use have received tremendous attention from the researchers. However, the development of such systems has faced challenges and language barriers. For instance, the development of such systems for Chinese medical domain is still relatively undeveloped. This could be linked to the complexity of Chinese text processing, and large dataset for the domain. In a study authored by (He et al, 2019) aimed at searching for technology applications that could solve this issue a Chinese medical QA data set was tested for QA system analysis [8]. The research results showed promising results as the proposed semantic clusters improved the module performance by 5.5% on precision at 1 and 4.9% Mean Average Precision. A comparison of other CWS tools have also been implemented, and the results showed effectiveness on the semantic cluster representation [8].

Anamnesis process is needed to get the symptoms of the disease, question and answer process between the patient and medical department whose results are stored in the Electronic Medical Record (EMR) in the form of description to suggest in the process of Clinical Decision Support (CDS). EMR is often difficult to do computing processing due to irrelevant grammar. For computers to interpret text data, Natural Language Processing (NLP) techniques is has been used. In this study, an NLP system was created that can identify symptoms of the digestive disease by using to optimize the CDS process. The method used to identify symptoms of the disease is Named Entity Recognition, which determines which tokens are included in the symptoms of the disease. The model trained with 800 epochs produces f1 score accuracy of 0.79 [6]. Medical students undergo exams, called" Objective Structured Clinical

Examinations" (OSCEs), to evaluate their medical competence in clinical tasks. In these OSCEs, a medical student reaches out with a regularized patient, asking queries to complete a clinical assessment of the patient's medical care issues. In this work, they develop a deep learning framework to enhance the virtual patient's conversational skills. First, deep neural networks learned domain specific word embedding. Then, long short-term memory networks derived sentence embedding before a convolutional neural network (CNN) model selected an answer to a given question from a script [6]. Empirical results on a homegrown corpus showed that this framework outclassed other approaches and reached an accuracy of 81%.

BERT (Bidirectional Encoder Representations from Transformers) experiments conducted by Google AI Language has resulted echoes in the Machine Learning community as it showcased state-of-the-art results in NLP applications, such as the QA (SQuAD v1.1), and Natural Language Inference (NLI) [7]. BERT's based applying a bi-directional training of Transformer. Such approach is different from previously implemented approaches that tests the text strings from a single direction, meaning either left to right approach, as it combined the L-R and R-L approaches at the same time. The researchers were able to generate results that shows the bidirectionally trained model could gain a deep sense of language context, compared to a single-direction language model. This novel technique named Masked LM (MLM), for bi-directional model training.

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by scholars at Google AI Language. It has caused a fiddling in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (NLI), and others [7]. BERT's key technical invention is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which inspect at a text sequence either from left to right or combined L-R and R-L training. The paper's produces results which shows that a language model which is bidirectionally trained can have a in depth sense of language context and flow than single-direction language models. In the paper, the researchers detail a novel technique named Masked LM (MLM) which allows bidirectional training in models in which it was previously almost impossible.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. As opposed to directional models, which read the text input sequentially (L-R or R-L), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (Left and Right of the word).

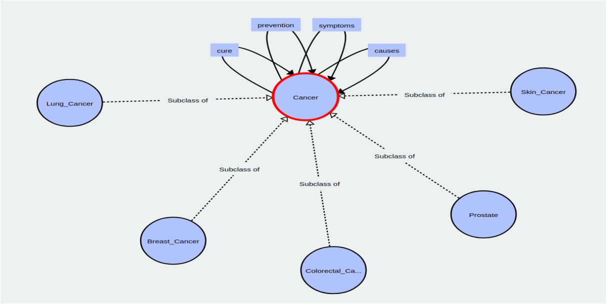


Fig. 1. Properties and Types of Cancer using OWL

III. Proposed work

A. Data Sets

The objective of our approach is to make use of the real data that's available over web, we intend to use web scrapping/web crawling techniques to collect data pertaining to various forms of cancer. However, such data would be inefficient for processing right away due to possible noise. We intend to pre-process the data without removing stop words, avoiding multiple cases within the passage.

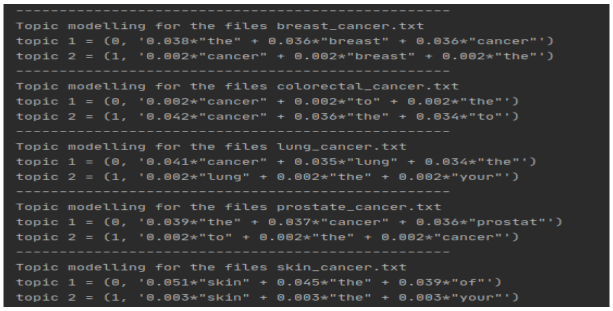
In healthcare domain, the data collection and analysis is setting the best example for data usage for improving the life of humans. It inspires using the data for the betterment of the healthcare services and health care agencies workflow. It also plays a major role in forecasting epidemics and curing existing diseases in an efficient way possible.

1) Web Crawling: As the data concerns to Medical domain and cancer being a critical data. We have handpicked few official websites which we believed to be right content for answering questions related to cancer. The initial implementation of the information extraction is the implementation of the web crawling, where in the crawling extended from originated website to flowing through the links of the other websites. This has resulted in about 70 GB worth data. However, due to the limited control on the url those are linked from the originating websites, we had to resort using Web scraping technique. As part of this experiments we have also realized that we couldn't use this amount of data on training BERT given the size of the BERT already.

2) Web Scraping: Due to limitations and the learning from the above, we have used Web scraping to collect data related to cancer from the sites. It consisted of three steps:

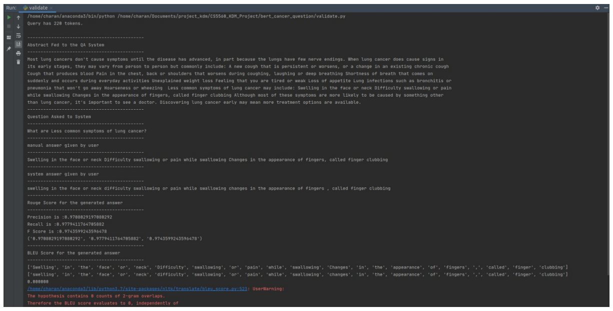
Generally, web scraping involves three steps:

- GET request is sent to the server and then receives the response for the web.
- HTML code parsing of the web site in a tree web structure.
- Python library is then used to look for tree which is parsing.



```
--------------------------------------------
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"')
```

Fig. 2. Topic Modeling on Cancer Data



```
/home/charan/anaconda3/bin/python /home/charan/Documents/project_kdm/CS5560_KDM_Project/bert_cancer_question/validate.py
Query has 220 tokens.

--------------------------------------------
Abstract Fed to the QA System
--------------------------------------------
Most lung cancers don't cause symptoms until the disease has advanced, in part because the lungs have few nerve endings. When lung cancer does cause signs in
its early stages, they may vary from person to person but commonly include: A new cough that is persistent or worsens, or a change in an existing chronic cough
Cough that produces blood Pain in the chest, back or shoulders that worsens during coughing, laughing or deep breathing Shortness of breath that comes on
suddenly and occurs during everyday activities Unexplained weight loss Feeling that you are tired or weak Loss of appetite Lung infections such as bronchitis or
pneumonia that won't go away Hoarseness or wheezing  Less common symptoms of lung cancer may include: Swelling in the face or neck Difficulty swallowing or pain
while swallowing Changes in the appearance of fingers, called finger clubbing Although most of these symptoms are more likely to be caused by something other
than lung cancer, it's important to see a doctor. Discovering lung cancer early may mean more treatment options are available.
--------------------------------------------
Question Asked to System
--------------------------------------------
What are Less common symptoms of lung cancer?
--------------------------------------------
manual answer given by user
--------------------------------------------
Swelling in the face or neck Difficulty swallowing or pain while swallowing Changes in the appearance of fingers, called finger clubbing
--------------------------------------------
system answer given by user
--------------------------------------------
swelling in the face or neck difficulty swallowing or pain while swallowing changes in the appearance of fingers , called finger clubbing
--------------------------------------------
Rouge Score for the generated answer
--------------------------------------------
Precision is :0.9708029197080292
Recall is :0.9779411764705882
F Score is :0.9743599243596478
('0.9708029197080292', '0.9779411764705882', '0.9743599243596478')
--------------------------------------------
BLEU Score for the generated answer
--------------------------------------------
['Swelling', 'in', 'the', 'face', 'or', 'neck', 'Difficulty', 'swallowing', 'or', 'pain', 'while', 'swallowing', 'Changes', 'in', 'the', 'appearance', 'of', 'fingers', ',', 'called', 'finger', 'clubbing']
['swelling', 'in', 'the', 'face', 'or', 'neck', 'difficulty', 'swallowing', 'or', 'pain', 'while', 'swallowing', 'changes', 'in', 'the', 'appearance', 'of', 'fingers', ',', 'called', 'finger', 'clubbing']
0.000000
/home/charan/anaconda3/lib/python3.7/site-packages/nltk/translate/bleu_score.py:523: UserWarning:
The hypothesis contains 0 counts of 2-gram overlaps.
Therefore the BLEU score evaluates to 0, independently of
```

Fig. 4. BLEU and ROUGE Score for BERT fine tune models

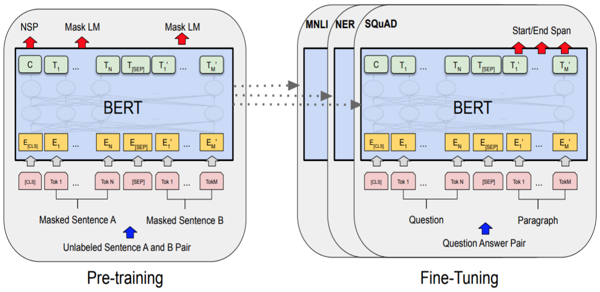


Fig. 3. Architecture of Distilling Bert

As part of data, we organized data into five categories of cancers such as breast cancer, lung cancer, colorectal cancer, skin cancer, prostate cancer 1. Unlike the traditional procedure of manually extracting the data, web scraping is automated data retrieval techniques which retrieve hundreds, millions, or even billions of data records the internet. After collecting data, we performed data cleaning on cancer data where special symbols appears and kept the data remaining same i.e., we have not removed any stop words since there is a lot of impact on the model which is suggested by BERT. In addition to that, we identified significance keywords using topic modelling 2.

B. Proposed Models

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. Distillation is done of very large batches by using gradient accumulation using dynamic masking [9].

C. Architecture

This is architecture of distilling bert in Fig. 3.

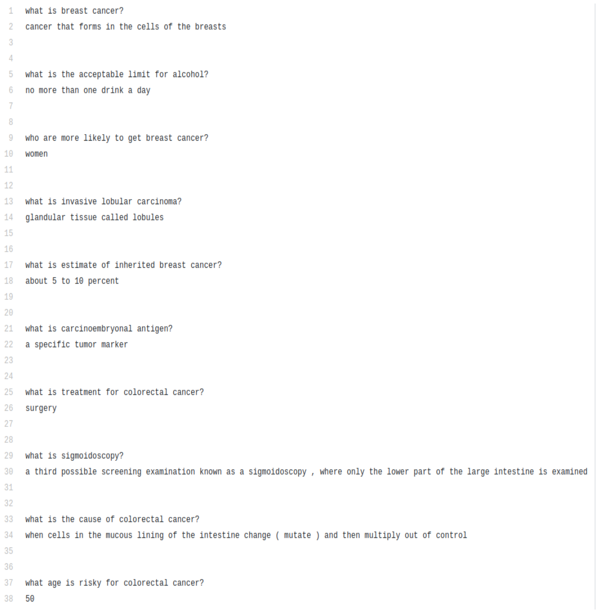
IV. Experiments

In this section, 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. Limited number of lines in the code are needed to initiate training model of the transformers. Which is also supports Sequence Classification, Token Classification (NER), Question Answering.

Among these implementations, we focused only on Question Answer System. Bilingual Evaluation Understudy (BLEU), is a score for comparing a candidate translation of text to one or more reference translations. In spite developed for translation, it can be used to evaluate text generated for a suite of NLP tasks. In this first case, we implemented BERT model where we will calculate the BLEU score for evaluating and scoring candidate text using the nltk library in Python. We consider only exact one-to-one matches between words. Precision and Recall is computed as ratio of the number of words in the translation that match words in the reference translation, and the number of words in the translation, ratio of the number of matching words and the number of words in the reference translation respectively. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is essentially of a set of metrics for evaluating automatic summarizing of texts as well as machine translation. It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced. Precision and Recall is calculated as ratio of no overlapping words and total words in system. 5

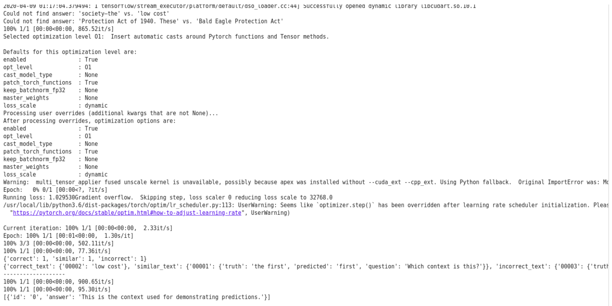
V. Summary

No.of.overlapping words and total.words.in referencing summary respectively. The above Fig. 4 shows 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. In ROUGE, we have achieved 0% results because the it is considering uni-grams instead of bi-grams. In the second case Fig. 6 , we implemented Distilled BERT which makes use of few test data-sets. Since Distil BERT has not been extended to currently available medical domain concern to cancer data set. Currently we don't have a state-of-art to make comparisons. comparison will be drawn to observe performance with previous techniques such a Word2Vec, TF-IDF based question answering. The comparison is not only done to understand the accuracy at which the BERT performs, but also to draw conclusions on how much time is perceived to build such model in comparison with peers.



```
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
```

Fig. 5. Results from Question Answering System of BERT



```
2020-04-09 01:17:04.379494: I tensorflow/stream_executor/platform/default/dso_loader.cc:44] Successfully opened dynamic library libcudart.so.10.1
Could not find answer: 'society—the' vs. 'low cost'
Could not find answer: 'Protection Act of 1940. These' vs. 'Bald Eagle Protection Act'
100% 1/1 [00:00<00:00, 865.52it/s]
Selected optimization level O1:  Insert automatic casts around Pytorch functions and Tensor methods.

Defaults for this optimization level are:
enabled                : True
opt_level              : O1
cast_model_type        : None
patch_torch_functions  : True
keep_batchnorm_fp32    : None
master_weights         : None
loss_scale             : dynamic
Processing user overrides (additional kwargs that are not None)...
After processing overrides, optimization options are:
enabled                : True
opt_level              : O1
cast_model_type        : None
patch_torch_functions  : True
keep_batchnorm_fp32    : None
master_weights         : None
loss_scale             : dynamic
Warning:  multi_tensor_applier fused unscale kernel is unavailable, possibly because apex was installed without --cuda_ext --cpp_ext. Using Python fallback.  Original ImportError was: Mo
Epoch:   0% 0/1 [00:00<?, ?it/s]
Running loss: 1.029530Gradient overflow.  Skipping step, loss scaler 0 reducing loss scale to 32768.0
/usr/local/lib/python3.6/dist-packages/torch/optim/lr_scheduler.py:113: UserWarning: Seems like `optimizer.step()` has been overridden after learning rate scheduler initialization. Pleas
  "https://pytorch.org/docs/stable/optim.html#how-to-adjust-learning-rate", UserWarning)

Current iteration: 100% 1/1 [00:00<00:00,  2.33it/s]
Epoch: 100% 1/1 [00:01<00:00,  1.30s/it]
100% 3/3 [00:00<00:00, 502.11it/s]
100% 1/1 [00:00<00:00, 77.36it/s]
{'correct': 1, 'similar': 1, 'incorrect': 1}
{'correct_text': {'00002': 'low cost'}, 'similar_text': {'00001': {'truth': 'the first', 'predicted': 'first', 'question': 'Which context is this?'}}, 'incorrect_text': {'00003': {'truth
-------------------
100% 1/1 [00:00<00:00, 900.65it/s]
100% 1/1 [00:00<00:00, 95.30it/s]
[{'id': '0', 'answer': 'This is the context used for demonstrating predictions.'}]
```

Fig. 6. Results based on Distill BERT

VI. Conclusion

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 from state-of-art models. In addition to that, the data we collected from different web sources using web scraping and the data is categorized different types of cancers using topic modelling. Finally, 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. QA systems has great potentials in the healthcare applications. Especially in the proposed 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.

VII. Future work

1) The comparison has to be drawn by comparing with other types of question answering systems using encoders as TF-IDF, Word2Vec.
2) the topic model technique used in our project is LDA, which was able to perform on accurate levels when fed the data. Research has to be extended to perform better topic modelling techniques.
3) we believe some of precision drop issues attributes to identifying a different means to pre-processing data to have context in the format of squad data set.

References

[1] S. Wang, N. Phan, Y. Wang, and Y. Zhao, "Extracting api tips from developer question and answer websites," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 321–332.

[2] H. Liang, B. Y. Tsui, H. Ni, C. C. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, *et al.*, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature medicine*, vol. 25, no. 3, pp. 433–438, 2019.

[3] N. Siangchin and T. Samanchuen, "Chatbot implementation for icd-10 recommendation system," in *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*. IEEE, 2019, pp. 1–6.

[4] R. B. Mathew, S. Varghese, S. E. Joy, and S. S. Alex, "Chatbot for disease prediction and treatment recommendation using machine learning," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2019, pp. 851–856.

[5] J. El Zini, Y. Rizk, M. Awad, and J. Antoun, "Towards a deep learning question-answering specialized chatbot for objective structured clinical examinations," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.

[6] F. B. Putra, A. A. Yusuf, H. Yulianus, Y. P. Pratama, D. S. Humairra, U. Erifani, D. K. Basuki, S. Sukaridhoto, and R. P. N. Budiarti, "Identification of symptoms based on natural language processing (nlp) for disease diagnosis based on international classification of diseases and related health problems (icd-11)," in *2019 International Electronics Symposium (IES)*. IEEE, 2019, pp. 1–5.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] J. He, M. Fu, and M. Tu, "Applying deep matching networks to chinese medical question answering: a study and a dataset," *BMC medical informatics and decision making*, vol. 19, no. 2, p. 52, 2019.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.