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

[1] [2] [3] [4] [5] 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. [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.

[7] Bidirectional Transformer encoder(BERT) 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 and clinical question answering (QA) is highly center of attention by researchers in recent days. Though there are tremendous advancements in this area, the development in Chinese medical domain is relatively rearward. It can be attributed to the difficulty of Chinese text processing and the lack of large scale data sets. To fill the gap between, this paper introduces a Chinese medical QA data set and proposes effective methods for the task [8]. Results also show that their proposed semantic clustered representation module improves the performance of models by up to 5.5% Precision at 1 and 4.9% Mean Average Precision.

Introduced a large scale Chinese medical QA data set and cast the task into a semantic matching problem. They also compare different CWS tools and input units. Among the two state-of-the-art deep matching neural networks performs much better. Results also show the effectiveness of the initiated semantic clustered representation module. 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 reach out with a regularized patient, asking questions to complete a clinical assessment of the patient's medical case. 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 [5]. 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) is a recent paper published by researchers 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 (MNLI), and others [7]. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The paper's results show that a language model which is bidirectionally trained can have a deeper 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 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 (left-to-right or right-to-left), 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).

Masked LM (MLM)

Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence 1. In technical terms, the prediction of the output words requires:

1) Adding a classification layer on top of the encoder output.
2) Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.
3) Calculating the probability of each word in the vocabulary with softmax.

The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words. As a consequence, the model converges slower than directional models, a characteristic which is offset by its increased context awareness.

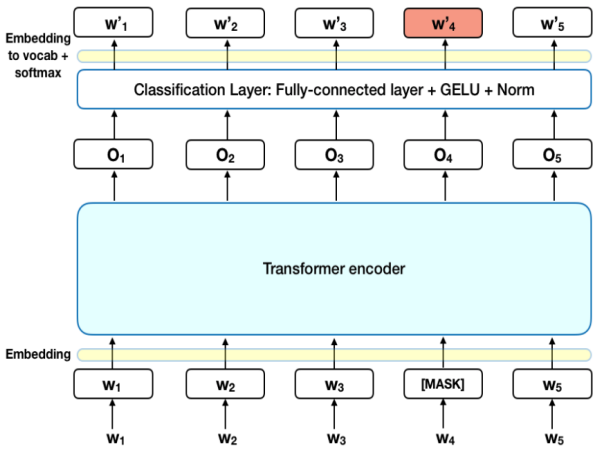

Fig. 1. Architecture of MaskedLM(MLM)

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 intent to pre-process the data by removing stop words, avoiding multiple cases within the passage.

We will be using Sentence Piece which is an unsupervised text tokenizer and de-tokenizer for neural network based natural language processing systems. Sentence Piece is an unsupervised text tokenizer and de-tokenizer mainly for Neural Network-based text generation systems where the vocabulary size is predetermined prior to the neural model training. Sentence Piece implements sub word units (e.g., byte-pair-encoding (BPE) [9] with the extension of direct training from raw sentences. Sentence Piece allows us to make a purely end-to-end system that does not depend on language-specific pre/post processing.

Data collection is the ongoing, trending, emerging and systematic process of gathering, analyzing and interpreting various types of information from various resources. In general, data collection is done for research purposes in order to understand the overview of an area of interest and to build a foundation for decision-making. In the healthcare sector, we can find the best examples of how data tracking and analysis change the world for the better. The significance of data in medicine is inspired by the necessity to solve both local organizational issues, such as reducing workloads and increasing benefits of a medical agency, and the universal issues of humanity, such as forecasting epidemics and combating existing diseases more efficiently. Breast cancer (BC)

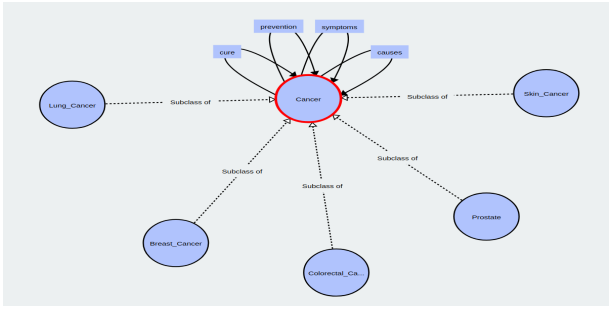

Fig. 2. Properties and Types of Cancer using OWL

```
----------------------------------------------------
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. 3. Topic Modeling on Cancer Data

is one of the most common cancers identifying among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

Initially we implemented, web crawling to collect the data as part of data sets on cancer data of size around 70GB. But our systems were unable to process such a huge data. So, we extracted the data using Web scraping also called as web data extraction, is the process of retrieving or scraping data from a different resources such as websites.

Generally, web scraping involves three steps:

- We send a GET request to the server and we will receive a response in a form of web content.
- Second, parse the HTML code of a web site following a tree structure path.
- Finally, we use the python library to search for the parse tree.

As part of data, we organized data into five categories of cancers such as breast cancer, lung cancer, colorectal cancer, skin cancer, prostate cancer 2. Unlike the traditional procedure, extraction is a process of manually extracting the data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet's seemingly endless boarders. 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 3.

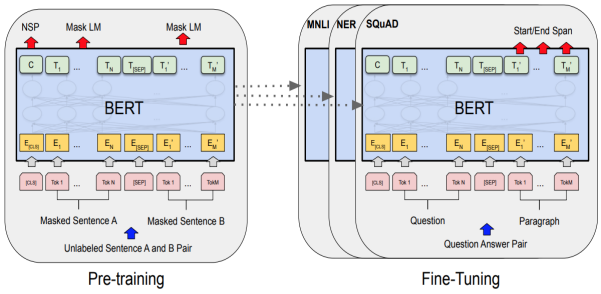

Fig. 4. Architecture of Distilling Bert

B. Proposed Models

[10] 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.

C. Architecture

This is architecture of distilling bert 4.

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. Transformers lets you quickly train and evaluate Transformer models. Only limited lines of code are needed to initialize a model, train the model, and evaluate a model. Currently supports Sequence Classification, Token Classification (NER), Question Answering, Multi-Modal Classification, and Conversational AI. 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. Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks.

In this first case, we implemented BERT model where we will calculated 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.of.overlapping words and total words in system

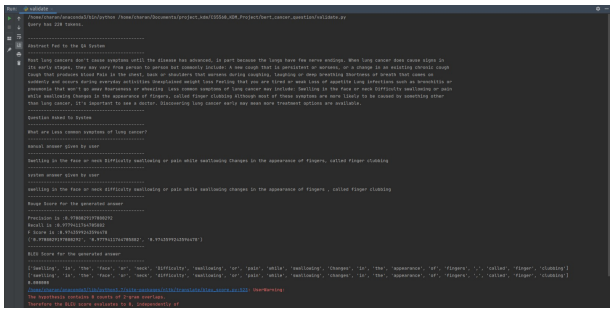

Fig. 5. BLEU and ROUGE Score for BERT fine tune models

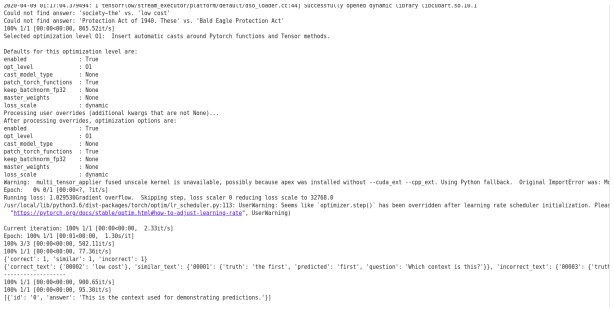

Fig. 6. Results based on Distill BERT

summary, no.of.overlapping words and total.words.in referencing summary respectively. The above figure 5 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 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 pertaining to cancer data set. we currently 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.

V. Conclusion

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] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[10] 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.