

Cancer ASK

FINAL PROJECT

Student:

MIRZA TANZIM SAMI¹, mtsami@uab.edu

NILESH KUMAR², nileshkr@uab.edu

TRUPESH R. PATEL¹, tr27p@uab.edu

Supervisor:

DR. JOHN DAVID OSBORNE¹

josborne@uabmc.edu

¹Department of Computer Science

²Department of Biology

NOVEMBER 30, 2020

Cancer ASK

MIRZA TANZIM SAMI* and TRUPESH R. PATEL*, University of Alabama at Birmingham
NILESH KUMAR*, University of Alabama at Birmingham

With the advent of new sophisticated experiment and analysis technology in medicine, the data generation in medicine has accelerated by several folds since the last decades. A big part of all gathered knowledge is the collection of text documents such as research articles. Keeping abreast of biomedical research developments, several efficient natural language processing (NLP) text mining models have also developed. Though, often it is difficult to apply those text mining models directly to domain-specific biomedical corpora. In this project, we combined two different types of NLP models so that the combined model can give a comprehensive answer to cancer-specific questions. At first, we fine-tuned BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [3] model for short question types over cancer-related question answers and context texts, we called Cancer Ask. Further, we fine-tuned the Generative Pre-trained Transformer 2 (GPT2) [4, 5], models over cancer-related text, we named it as GPT2 Cancer. The fundamental objective of this project is to use answers from Cancer Ask, feed it to GPT2 Cancer and provide a comprehensive answer to cancer related queries.

ACM Reference Format:

Mirza Tanzim Sami, Trupesh R. Patel, and Nilesh Kumar. 2020. Cancer ASK. 1, 1 (November 2020), 9 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

DATASET

We are using Cancer dataset, which are published as MedQuAD dataset [1]. This dataset is specifically created for medical question-answering. The dataset created from 12 NIH websites (e.g. cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). The Cancer dataset are focused on 98 types of cancers, and each type of cancer can have maximum 12 types of questions. In total we have 729 questions, and their questions types are distributed as Table 1 and Figure 1:

Data format

MedQuAD's cancer dataset are in Extensible Markup Language (XML) [2], with specific tags as shown in Table 2:

For example:

```
<Document id="0000001_1" source="CancerGov" url="https://www.cancer.gov/types/leukemia/patient/adult-all-treatment-pdq">
  <Focus>Adult Acute Lymphoblastic Leukemia</Focus>
  <FocusAnnotations>
```

*All authors contributed equally to this project.

Authors' addresses: Mirza Tanzim Sami, mtsami@uab.edu; Trupesh R. Patel, tr27p@uab.edu, University of Alabama at Birmingham, Department of Computer Science, University Hall 4105, 1402 10th Ave. S., Birmingham, Alabama, 35294-1241; Nilesh Kumar, nileshkr@uab.edu, University of Alabama at Birmingham, Department of Biology, University Hall ???, 1402 10th Ave. S., Birmingham, Alabama, 35294-1241.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

XXXX-XXXX/2020/11-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

```

<UMLS>
  <CUIs>
    <CUI>C0751606</CUI>
  </CUIs>
  <SemanticTypes>
    <SemanticType>T191</SemanticType>
  </SemanticTypes>
  <SemanticGroup>Disorders</SemanticGroup>
</UMLS>
</FocusAnnotations>
<QAPairs>
  <QAPair pid="1">
    <Question qid="0000001_1-1" qtype="information">What is (are) Adult Acute
      Lymphoblastic Leukemia ?</Question>
    <Answer>Key Points – Adult acute lymphoblastic leukemia (ALL) is a type of
      cancer in which the bone marrow makes too many lymphocytes (a type of
      white blood cell ). – Leukemia may affect red blood cells , white blood
      cells , and platelets . – Previous chemotherapy and exposure to radiation
      may increase the risk of developing ALL. – Signs and symptoms of adult
      ALL include fever , feeling tired , and easy bruising or bleeding. –
      Tests that examine the blood and bone marrow are used to detect (find)
      and diagnose adult ALL. ...
    </Answer>
  </QAPair>
</QAPairs>
</Document>

```

Now, by the look the data looks easy to feed into BioBERT model, however, BioBERT model requier four major components “context”, “question”, “answer”, and “start_answer”. And the given data are missing two main parts “context” and “start_answer”. So, we have to take few manual following steps to make dataset compatible to BioBERT model.

Table 1. List of the question type and their count

Question type	Count
Information	112
Treatment	95
Susceptibility	88
Research	86
Symptoms	82
Exams and tests	82
Outlook	82
Stages	77
Prevention	12
Causes	7
Inheritance	5
Genetic changes	1

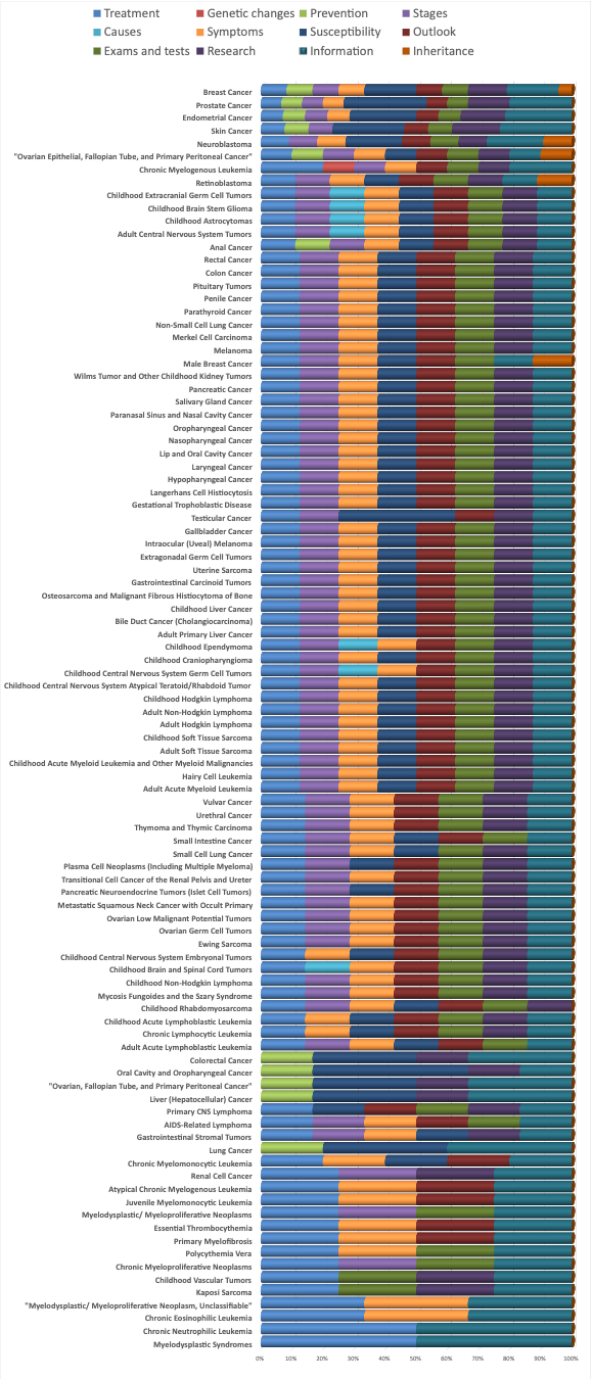


Fig. 1. List of the Cancer type and question type and their count

Table 2. List of XML tags in raw dataset

XML tags	Description
Focus	Focus of the question
UMLS	A standardized semantic knowledge source
CUI	Concept Unique Identifier
SemanticType	Semantic Features of questions
SemanticGroup	Semantic Group of questions
Question	The question text
Answer	The answer text

- The give “answer” tags contains very big string, but the correct answer is just few lines of this text. So, we made give “answer” tag as “context” tag.
- Manually find the correct answer and made it as “answer” tag.
- Calculate “start_answer” tag from new generated “context” tag
- finally, stored every data in JavaScript Object Notation (JSON) format.

Example of new data:

```
{
  "data": [
    {
      "title": "information",
      "paragraphs": [
        {
          "context": "Key Points Adult acute lymphoblastic leukemia (ALL) is a type of cancer in which the bone marrow makes too many lymphocytes (a type of white blood cell). Leukemia may affect red blood cells, white blood cells, and platelets. Previous chemotherapy and exposure to radiation may increase the risk of developing ALL. Signs and symptoms of adult ALL include fever, feeling tired, and easy bruising or bleeding. Tests that examine the blood and bone marrow are used to detect (find) and diagnose adult ALL. ...",
          "qas": [
            {
              "answers": [
                {
                  "text": "Adult acute lymphoblastic leukemia (ALL) is a type of cancer in which the bone marrow makes too many lymphocytes (a type of white blood cell).",
                  "answer_start": 11
                }
              ],
              "question": "What is (are) Adult Acute Lymphoblastic Leukemia ?",
              "id": "1"
            }
          ]
        },
        ...
      ]
    },
    ...
  ],
  "version": "1.1",
  "team": "nlp-group-project-fall-2020-deepbiocomp",
  "Disease": "Cancer"
}
```

METHOD

The following sections are categorized into three main sections, first a brief explanation on the data processing and implementation methods used to fine-tune the BioBERT model for regular question answering. Second, the data processing and fine-tuning of the GPT2 model for text-generation which is the basis for creation of long comprehensive answers. Finally, the report discusses the implementation methods for a composite model that runs the fine-tuned BioBERT for Question-Answering and uses the output of the BioBERT model as the input prompt for the GPT2 text generator for a more verbose, comprehensive answer.

Fine-tuning BioBERT

The biomedical domain texts contain a vast number of domain-specific proper nouns (e.g. BRCA1, Leukemia) which are understood mostly by biomedical researchers. In this context the BioBERT is already fine-tuned on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). Further, we fine-tuned again using Cancer QA dataset, the MedQuAD. The total Cancer and question type is summarized in Figure 1.

Fine-tuning GPT2

This section will briefly explain the data used to pretrain the GPT2 model and how the script that was implemented to use the GPT2 model for text generation. The GPT2 model was chosen for this task because of the robustness of the model and its ability to generate long sentences while maintaining relatively good semantic sense. However, the language model is much too general and requires fine-tuning to work effectively and generative texts pertinent to cancer queries. Thus, the model was fine-tuned on the same MedQuAD dataset that was used to fine-tune the aforementioned BioBERT model.

However, unlike the BioBERT model which is used for question-answering, the data must be processed differently for the GPT2 model. In the case of the question-answering models, the dataset usually has three major components, the 'question', 'context' and the 'answer'. All of which are important for training a question-answering model, but fine-tuning a language model (GPT2) does not require all three components. Concretely, the GPT2 model was trained only using the 'context' of the dataset. Furthermore, two special tokens were added, the '<BOS>' signifying the beginning of a sentence and a '<EOS>' token, signifying the end of a sentence. The code for fine-tuning the GPT2 model is based on an older (not currently available) script from the huggingface transformer git repo called 'run_language_modelling.py'. The script was modified and reimplemented as a jupyter notebook. Furthermore, it was also modified to let GPT2 accept the special tokens as mentioned above. Once the model was fine-tuned with our desired dataset, it was saved locally to be used for can related text-generation.

Composite Model for Comprehensive Question Answering

The final section will discuss the composite model that stacks the fine-tuned GPT2 based text-generation model on top of the fine-tuned BioBERT model for question-answering. The question-answering code is based on the 'run_squad.py' script from the Huggingface's Transformer git repository. The fine-tuned model takes as input the question provided by the user and tries to give an answer that is correct and contextually relevant to the question asked. Once, the BioBERT model returns an answer for the given query, the output is used as the input for the GPT2 model. The text-generation code is based on the 'run_generation.py' script from the Huggingface's Transformer git repository. The script has been heavily modified and rewritten as a python function in a jupyter notebook. It has also been modified to accept our dataset which contains the '<BOS>' and '<EOS>'

special tokens. The text-generation model returns two suitable answers which are generated based on the prompt provided from the question-answering model. The resulting final answer is not only verbose and comprehensive but semantically and contextually relevant to the question asked. Thus, providing a much better experience to the user submitting the queries to the model.

RESULTS

This section of the report will discuss the results obtained from the two models used in this project, the metrics used to measure the accuracy of the model and the qualitative measure of the generated answers.

Training and Accuracy Metrics

First, the BioBERT model was fine-tuned on our dataset. The model was trained for 20 epoch running on 4 GPU cores. The model reduced the loss value with each epoch during training. Then, the following accuracy metrics scores were achieved during evaluation on test data.

- 'exact': 70.83333333333333
- 'f1': 77.78311271345326,
- 'total': 72,
- 'HasAns_exact': 70.83333333333333,
- 'HasAns_f1': 77.78311271345326,
- 'HasAns_total': 72,
- 'best_exact': 70.83333333333333,
- 'best_exact_thresh': 0.0,
- 'best_f1': 77.78311271345326,
- 'best_f1_thresh': 0.0

The accuracy metrics were monitored carefully to ensure that the model generalizes well and not overfit to the training data.

Next, the GPT2 model was fine-tuned on the data processed for it. This model was tuned for 19 epochs and the loss value decreased with each epoch. Once the model was tuned, the accuracy of the model was evaluated on the text data. The primary metric used was the perplexity score of the model which is as follows,

- 'perplexity': 2.5901412982272793.

Again, special attention is given to ensure that the model remains general and not overfit to the data.

Qualitative Results

Given the nature of the project, numeric accuracy scores may not be an intuitive indicator of the performance of the model. Thus, a few examples are given below that shows the question asked to the model, the answer given by the BioBERT model and the verbose, comprehensive answer given by the GPT2 model stacked on top.

Set 1:

Question: "What are the symptoms of ovarian rian Germ Cell Tumors ?"

Answer: signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause.

Comprehensive:

- **Answer 1:** signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. These and other signs and symptoms may be caused by ovarian germ cell tumor or by other conditions. Check with your doctor if you have any of the following:

Weakness or feeling tired. Weight loss with little or no effect on menstrual periods. Vaginal bleeding after menopause. Sometimes pain or swelling in a women's vagina. Fever or night sweats.

- **Answer 2:** signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. These and other signs and symptoms may be caused by ovarian germ cell tumor or by other conditions. Check with your doctor if you have any of the following: Swelling of the abdomen or vaginal bleeding after menopause. Trouble starting the flow of urine. Weight loss for no known reason. Trouble emptying the bladder completely after menopause. Pain or feeling of fullness below the ribs on the left side. Fever for no known reason. Easy bruising or bleeding. Pain or feeling of fullness below the ribs on the right side.
- **Answer 3:** signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. Other conditions can increase the risk of hairy cell leukemia. Check with your doctor if you have any of the following: Pain or swelling in the abdomen. Weakness or feeling tired. Weight loss for no known reason. A menstrual period that does not go away.
- **Answer 4:** signs of ovarian germ cell tumor are swelling of the abdomen or vaginal bleeding after menopause. These and other signs may be caused by ovarian germ cell tumor or by other conditions. Check with your doctor if you have any of the following: Pain or swelling in the abdomen. A lump in the abdomen, vagina, or rectum. Weight loss for no known reason. Pain or a feeling of fullness below the ribs on the left side. Weight loss with no known reason. A dark urine like color that does not go away.

Set 2:

Question: "What are the stages of Vulvar Cancer ?"

Answer: stage i stage ii stage iii stage iv after vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body.

Comprehensive:

- **Answer 1:** stage i stage ii stage iii stage iv after vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body. Cancer that has spread within the vulva to other parts of the body is called seminoma. A biopsy is done to diagnose seminoma. A biopsy is done during surgery to remove the tumor. Sometimes, a small incision (cut) is made in the vulva and a laparoscope (a thin, lighted tube) is inserted into it. The doctor will usually remove the tumor and a small amount of tissue, after the doctor has removed all the cancer. The tissue sample is checked under a microscope for signs of cancer. A pathologist views the tissue under a microscope to look for cancer cells. If cancer cells are not found, it may not be necessary to remove more tissue. There are three ways that cancer spreads in the body. Cancer can spread through tissue, the lymph system, and the blood: Tissue. The cancer spreads from where it began by growing into nearby areas. Lymph system. The cancer spreads from where it began by getting into the lymph system. The cancer travels through the lymph vessels to other parts of the body. Blood. The cancer spreads from where it began by getting into the blood. The cancer travels through the blood vessels to other parts of the body. Cancer may spread from where it began to other parts of the body through organs and tissues. When cancer spreads to another part of the body, it is called metastasis. Cancer cells break away from where they began (the primary tumor) and travel through the lymph system or blood.
- **Answer 2:** stage i stage ii stage iii stage iv after vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body. The following stages are used for vulvar cancer: Stage I Stage II Stage III Stage IV

Stage V After vulvar cancer has been diagnosed, tests are done to find out if cancer cells have spread within the vulva or to other parts of the body. The process used to find out if cancer has spread within the vulva or to other parts of the body is called staging. The information gathered from the staging process determines the stage of the disease. It is important to know the stage in order to plan treatment. Vulvar cancer in men may affect the risk of developing rectal cancer. Having rectal cancer during surgery can increase the risk of rectal cancer and is called a risk factor. Having rectal cancer in a first degree relative (parent, brother, or sister) increases the risk even more.

CONCLUSION

The answers or responses produced by our fine-tuned model support the relevance of the question in logic with cancer QA. Our model also highlights the effectiveness of combining text generative and long-former models for further model improvement. Our findings also show that relying on a restricted set of reliable answer sources can bring a plentiful improvement in domain-specific QA.

Future Plans

The dataset used in this project has on average a very long context paragraph. For the ease of fine-tune on available resources, the only initial part of the context was used. Further, the transformer-based models like BioBERT are inefficient to handle long sequences due to expensive self-attention operations. To address this limitation, we will fine-tune the Longformer. As compared to the BioBERT (transformer-based), the attention mechanism in the Longformer scales linearly with sequence length. This feature helps it to process medical documents of thousands of tokens or longer. Additionally, the chosen dataset (transformer-based) is too small and we were unable to procure the dataset from BoiASK. When available we will finetune over this big data set to add more diversity to the model.

REFERENCES

- [1] Asma Ben Abacha and Dina Demner-Fushman. 2019. A Question-Entailment Approach to Question Answering. *CoRR* abs/1901.08079 (2019). arXiv:1901.08079 <http://arxiv.org/abs/1901.08079>
- [2] Asma Ben Abacha and Dina Demner-Fushman. 2019. A Question-Entailment Approach to Question Answering. (January 2019). https://github.com/abachaa/MedQuAD/tree/master/1_CancerGov_QA
- [3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (09 2019), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682> arXiv:<https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf>
- [4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>