

Extraction of Cancer Type Information from Oncology Notes

Mentors: Karthik Ramesh & Dr. Sandip Patel

Table of Contents

- 1 Motivation & Contribution
- 2 Data
- 3 Methods
- 4 Discussion & Conclusion

Table of Contents

- 1 Motivation & Contribution**
- 2 Data
- 3 Methods
- 4 Discussion & Conclusion

Introduction

- When a person gets admitted to a hospital, the doctors make a series of diagnosis starting from most relevant and then going ahead decreasing in relevance. This diagnosis is in the form of ICD codes.
- Specifically for cancer patients, the immunotherapy treatment can be applied.
- For this, the doctor should know if the primary diagnosis was cancer.
- However in real life, the ICD codes are messy and its difficult to determine if the diagnosis was cancer.
- So, doctors have to read the discharge summaries to determine if the diagnosis was indeed cancer.
- Then they need some more information like medication, type of cancer, admission reason to determine which type of immunotherapy is viable
- What if we had a well labeled dataset, and we could use a NLP model in a supervised fashion to identify given a discharge summary was the primary diagnosis was cancer? And then in the second step we determine the different details using LLMs?

Introduction

- ICD codes often unclear for diagnosis.
- Immunotherapy needs accurate primary diagnosis identification, whether or not it is cancer.
- NLP model for supervised cancer diagnosis recognition.
- Utilize well-labeled dataset for model training.
- Further details extracted using advanced language models.

Problem Statement

- Train an NLP model to predict if a discharge summary's primary diagnosis was cancer. Once cancer is identified, we then deploy large language models (LLMs) to extract nuanced details from the discharge notes.
- This streamlined approach allows us to harness the strengths of both models, ensuring efficient and comprehensive analysis.
- By combining the predictive strength of the NLP model with the contextual understanding of LLMs, we enhance the precision and depth of information extraction from medical records.

Table of Contents

- 1 Motivation & Contribution
- 2 **Data**
- 3 Methods
- 4 Discussion & Conclusion

Data

Data collection

- MIMIC-IV Dataset for Medical Information.
- De-identified discharge summaries and diagnoses.
- Diagnosis with ICD-9 or ICD-10 codes.
- Focus on 'neoplasm' ICD codes.

Data

Data collection

- We use the MIMIC-IV and MIMIC-IV note dataset. MIMIC stands for Medical Information Mart for Intensive Care.
- We work with de-identified discharge summaries as well as the diagnoses of patients for their various admissions to hospitals.
- For each admission to a hospital, the doctors have a series of diagnoses for their condition, numbered from 1 onwards.
- This is called as the sequence number.
- The diagnosis is in the form of a ICD-9 or a ICD-10 code. We have a dataset of ICD codes, their version, and a description of the disease mentioned by the ICD code. We only want to work with ICD codes for cancer, so we only choose the codes which have the word 'neoplasm' in their description.

Data

Data cleaning

- Discharge summary structure extraction.
- Extracting relevant details.
- Fit within BERT's context.
- Limited to 512 tokens.

Data

Data cleaning

- Each discharge summary has a specific structure, and using that information we try to extract most relevant details from a discharge summary.
- This is done so as to make the text fit in the limited context window of language models like BERT, which is usually around 512 tokens.

Data

Dataset creation

- Sequence numbers and ICD codes.
- Data divided into two classes.
- Class imbalance observed.
- Addressed with oversampling, augmentation.

Data

Dataset creation

- Using the shortened discharge summary, the sequence numbers as well as the diagnosis ICD codes for a patient for a specific hospital admission, we divide the data into two classes.
- The first class has those entries where the primary diagnosis is cancer.
- The second class has those entries where the primary diagnosis is not cancer.
- We observe class imbalance, and use two techniques: oversampling as well as data augmentation to combat that.
- For oversampling, we repeat some entries for the first class.
- Data augmentation will be discussed later.

Table of Contents

- 1 Motivation & Contribution
- 2 Data
- 3 **Methods**
- 4 Discussion & Conclusion

Methods

1. Fine-tuning BERT models
2. Dealing with Class Imbalance and Fine-tuning.
3. Large Language Models for Information Extraction

Data Imbalance

1. 3 million samples of people who don't have cancer as the primary diagnosis and 14,000 samples of people who have cancer as the primary diagnosis.
2. Initially used naive undersampling to resolve it.

Fine-tuning BERT models

BERT

1. BERT Models have been quite useful for text classification
2. We used BERT models pre-trained on medical corpus to obtain better results

Fine-tuning BERT models

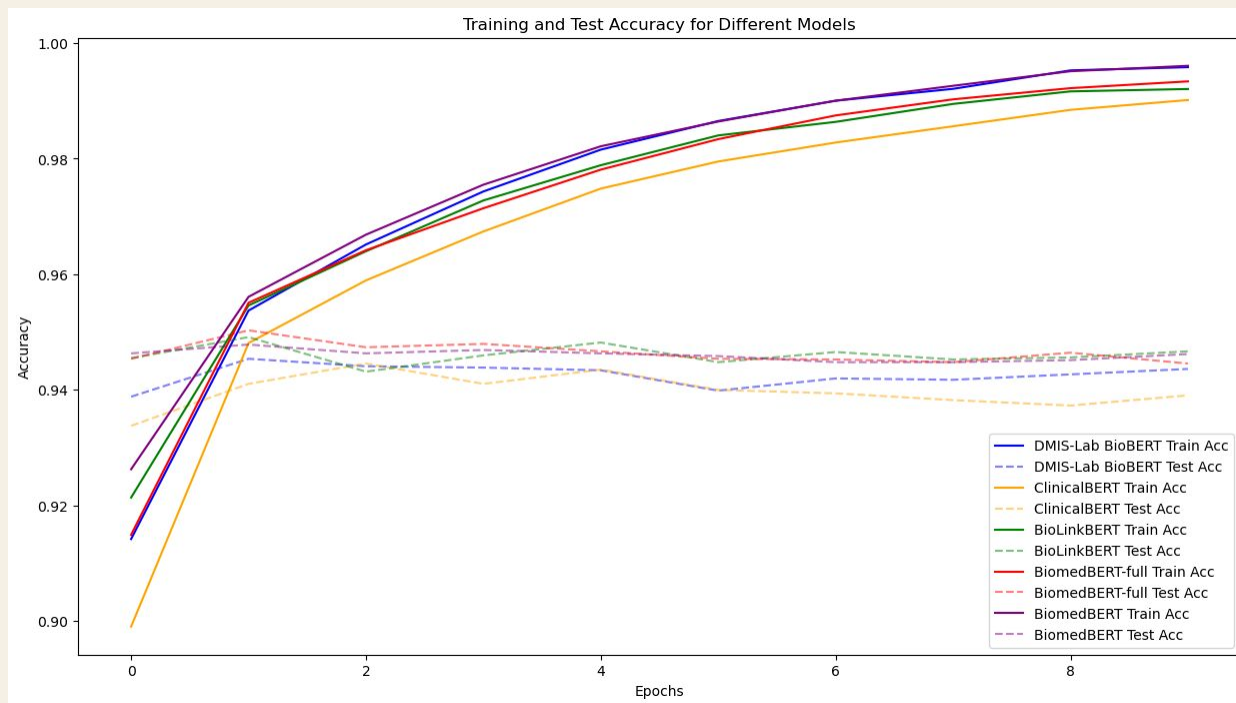
Pre-Training Corpus

Model	Pre-Training Corpus
BioBert	PubMed Abstracts + BooksCorpus + PMC Full-text articles
ClinicalBERT	Large multicenter dataset with a large corpus of 1.2B words of diverse diseases
BioLinkBERT	PubMed abstracts
BioMed-full	Abstracts from PubMed and full-text articles from PubMedCentral
BioMed	Trained from scratch using PubMed abstracts

Fine-tuning BERT models (Performance)

Results

Name	w/o. Aug
BioBERT	0.945
ClinicalBERT	0.944
BioLinkBERT	0.949
BioMed-full	0.950
BioMed	0.947



Dealing with Class Imbalance and Fine-tuning

Class Imbalance

- Number of datapoints after initial data processing step:
 - a. Label 0 ~ 3 million
 - b. Label 1 ~ 14,000 (Huge class imbalance)
- Can be easily dealt by undersampling from class 0, but results in loss of information.
- Solution? Augment class 1 to increase the #datapoints (and undersample from class 0 due to limited compute).

Dealing with Class Imbalance and Fine-tuning

Data Augmentation

- Contextual Augmentation:
 - a. Proposed in *Data Augmentation by Words with Paradigmatic Relations* (Kobayashi S., 2018)
 - b. Choose a word in a sentence and replace it with a word based on the surrounding context.
 - c. We mask **15%** of the tokens in every discharge summary and used **BioMedBERT** to fill them.

Dealing with Class Imbalance and Fine-tuning

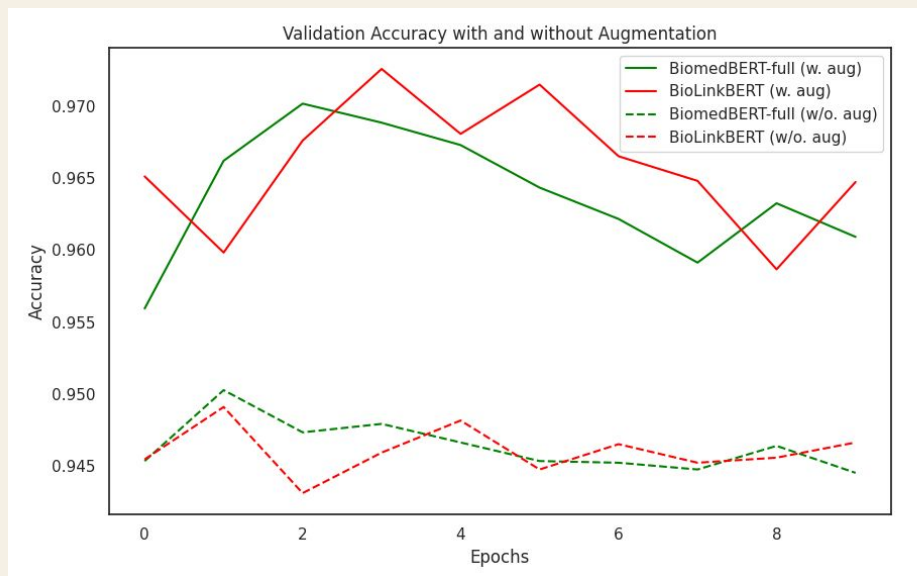
Data Augmentation

- Summarization:
 - a. Average discharge summary length ~ 300 words.
 - b. Summarized using **T5-large** model.
 - c. Prompt: “*Summarize this discharge summary: <discharge summary>*”.
- Class Balance?
 - **Label 1**: 14,000 → 40,000
 - Label 0: 3 million → 40,000 (*undersampling*)

Dealing with Class Imbalance and Fine-tuning

Results

Name	w/o. Aug	w. Aug
BioBERT	0.945	0.965
ClinicalBERT	0.944	0.967
BioLinkBERT	0.949	0.972
BioMed-full	0.950	0.970
BioMed	0.947	0.969



Large Language Models (LLMs) for Information Extraction

- LMs explored
- Information required (& prompts used to extract it)
- Parameters that affect the quality of responses & efficiency
- Results

Large Language Models (LLMs) for Information Extraction

LMs explored

We tried **BioMedLM** (from Stanford, MosaicML), **OpenChat3.5** (from OpenChat) & **Llama 2** (Meta)

Background

- (a) BioMedLM (~2.7B)
 - (i) trained exclusively on PubMed abstracts & full texts
- (b) OpenChat3.5 (~7B)
 - (i) Tuned on datasets from diverse domains - personal assistant interactions, conversations, web documents, etc. using reinforcement learning
- (c) Llama2 (~7B)

Large Language Models (LLMs) for Information Extraction

Information required

Given the discharge summary, our focus was extracting the following information:

1. Type of cancer
2. Exhaustive medication list
3. Admission reason

There are other pieces of information that would be useful too, like, cancer stage, condition: remission/active - part of next steps.

Large Language Models (LLMs) for Information Extraction

Information required

To extract this information, we used the following template:

“Following is a discharge summary of a patient delimited by triple backquotes. Given the note, <question>. Stay concise.

```{record}``` <|end\_of\_turn|> ”

The end-of-turn token is an idiosyncrasy of OpenChat. It allows for conversational use of the model.

# Large Language Models (LLMs) for Information Extraction

## Model specifics

Parameters to control model output:

- ``max_new_tokens``: a small value works for things like cancer type. Listing all medication requires a bigger value.
- ``temperature``: a higher value leads to more “creative/unexpected” output. Since, for our problem, the task is mainly of information extraction, we don’t want the model getting too creative!

# Large Language Models (LLMs) for Information Extraction

## Model specifics

- For a single prompt with a 150-token output limit,

	Inference Time (min)
OpenChat 3.5 (7B)	0.5
Llama 2 (7B)	1.34
BioMedLM (2.7B)	15.7

(Run on DataBricks Nightingale)

# Large Language Models (LLMs) for Information Extraction

## Results

### [Llama2]

emerged, send (paracetamol, antibiotics), cocaine, cocaine

Attending: \_\_\_\_  
chief complaint: Weakness, nausea/vomiting  
past medical history:  
family history: Mother - \_\_\_\_ cancer d. at \_\_\_\_  
Youngest of \_\_\_\_ - not in touch with siblings  
Father - \_\_\_\_ d. at \_\_\_\_  
brief hospital course: \_\_\_\_ yo f with h/o recently diagnosed metastatic cancer of unknown  
primary presenting with nausea, vomiting, and fever to 101 on day of admission.  
.  
# Fever/leukocytosis: Patient presented with fever and leukocytosis, which was initially concerning for post obstructive PNA. Patient was treated with vanc-cefepime-azithro for a day and a half, however CT torso was concerning for ongoing malignancy without evidence of primary tumor. There was no concern for a pneumonia. Consequently, abx were stopped on \_\_\_\_\_. Patient felt weak and unable to go home, as she lives alone and so was screened by physical therapy and discharged to rehab facility.  
.  
# Failure to thrive: Patient presents with 6 weeks of nausea, vomiting and poor PO intake likely from her enlarged liver. She continues to have bowel movements and pass gas. Patient now with albumin of 2.8, ketones in her urine, weight loss and new peripheral edema, concerning for poor nutrition/starvation ketosis. Nutrition consult followed the patient throughout her

diagnosis	medication	admission_reason
Based on the discharge summary, the patient has a diagnosis of metastatic cancer, specifically "metastatic cancer of unknown primary presenting with nausea, vomiting, and fever."	<p>The patient was discharged with the following medications:</p> <ul style="list-style-type: none"><li>* Gabapentin</li><li>* Bupropion</li><li>* Sertraline</li><li>* Trazodone</li><li>* Carnation Instant Breakfast with Beneprotein (as recommended by nutrition consult)</li><li>* HCTZ (to be reconsidered by PCP if medically necessary)</li></ul> <p>Note that the medications listed above are based on the information provided in the discharge summary, and may not be a comprehensive list of all medications the patient was taking at the time of discharge.</p>	<p>The patient was admitted to the hospital with chief complaints of weakness, nausea, and vomiting. The patient's past medical history includes a history of cancer in her mother, who passed away at a young age, and a history of cancer in her youngest sibling, who is not in touch with her. The patient's father also passed away from cancer. The patient was recently diagnosed with metastatic cancer of unknown primary, and she presented to the emer</p>

# Large Language Models (LLMs) for Information Extraction

## Results

### [OpenChat]

emerged, send (paracetamol, antibiotics), coxibs, acetaminophen.

Attending: \_\_\_\_.

chief complaint: Weakness, nausea/vomiting

past medical history:

family history: Mother - \_\_\_\_ cancer d. at \_\_\_\_

Youngest of \_\_\_\_ - not in touch with siblings

Father - \_\_\_\_ d. at \_\_\_\_

brief hospital course: \_\_\_\_ yo f with h/o recently diagnosed metastatic cancer of unknown

primary presenting with nausea, vomiting, and fever to 101 on day of admission.

.

# Fever/leukocytosis: Patient presented with fever and leukocytosis, which was initially concerning for post obstructive PNA. Patient was treated with vanc-cefepime-azithro for a day and a half, however CT torso was concerning for ongoing malignancy without evidence of primary tumor. There was no concern for a pneumonia. Consequently, abx were stopped on \_\_\_\_.

Patient felt weak and unable to go home, as she lives alone and so was screened by physical therapy and discharged to rehab facility.

.

# Failure to thrive: Patient presents with 6 weeks of nausea, vomiting and poor PO intake likely from her enlarged liver. She continues to have bowel movements and pass gas. Patient now with albumin of 2.8, ketones in her urine, weight loss and new peripheral edema, concerning for poor nutrition/starvation ketosis. Nutrition consult followed the patient throughout her

diagnosis	medication	admission_reason
The patient has a diagnosis of metastatic cancer of unknown primary.	The medications given to the patient are: - Gabapentin - Bupropion - Sertraline - Trazodone	The patient was admitted with weakness, nausea, and vomiting.

# Table of Contents

- 1 Motivation & Contribution
- 2 Data
- 3 Methods
- 4 **Discussion & Conclusion**



# Discussion & Conclusion

- Discussion
  - Summarization with different headers
  - Include radiology reports for better context
  - Try domain specific LLMs - MEDPALM, MEDPALM2
- Conclusion
  - We got good results using BERT for determining whether the patient has the primary diagnosis as cancer given a discharge summary.
  - Our validation of information extraction results with Karthik yielded consistently satisfactory outcomes.



Thank You!