Clinical Note Generation from Doctor-Patient Encounters and Sentiment Analysis

Bishwashri Roy
CS23BTNSL11001

Ujjwal Kumar Singh
CS23BTNSK11002

Kethavath Praneeth Nayak
CS20BTECH11025

***Abstract*—Medical doctors spend on average 52 to 102 minutes per day writing clinical notes from their conversation with patients. Reducing this workload demands relevant and efficient summarization methods for which we have used the resources from this reference paper.In our model we have also tried to integrate sentiment analysis.**

***Index Terms*—summarization methods,sentiment analysis**

I. Introduction

Summarization of doctor patient conversations is highly influenced by large transformer based models and availability of large-scale datasets. It comes with its own set of challenges in addition to the natural language understanding and generation,most important among which is omission of critical medical facts which might amend patient outcomes and hence it plays a major role in choosing one summarization method over the other.

II. Dataset Description

A. MTS-DIALOG dataset

This dataset was created by generating doctor-patient conversions from the public Mtsamples collection, which provides de-identified clinical notes. The MTS-Dialog dataset is a new collection of 1.7k short doctor-patient conversations and corresponding summaries (section headers and contents).

- The training set consists of 1,201 pairs of conversations and associated summaries.
- The validation set consists of 100 pairs of conversations and their summaries.

1) Doctor: My chart here says that you're eighty three years old, is that correct, ma'am?
2) Patient: Yes doctor, that's correct, I just had my birthday.
3) Doctor: Happy belated birthday! How have you been doing since your last visit?
4) Patient: Well, my cancer hasn't needed phlebotomies for several months now, which is good.
5) Doctor: That's great, you have been treated for polycythemia vera, correct?
6) Patient: Yes, that's the one.
7) Doctor: I also see you're unassisted today, which is also great.
8) Patient: Yeah, having some independence is nice.

Section Header: History of Present Illness

Section text:The patient is an 83-year-old female with a history of polycythemia vera. She comes in to clinic today for follow-up. She has not required phlebotomies for several months. The patient comes to clinic unaccompanied.

Table 1: Example of a doctor-patient conversation and associated note/summary from the MTS-DIALOG dataset

B. Sentiment Database

We have a sentiment dataset of tweeter content , consisting 40000 entries of dialogue and their corresponding emotions. This dataset has 4 columns:

- tweetId
- sentiment
- author
- content

Out of these sentiment and content is useful for us.

III. Models used

Our reference paper has demonstrated the best results when employing BART (Bidirectional Auto-Regressive Transformer) combined with Prefinetuning (PFT) and Guided Summarization (GS). [utf8]inputenc

A. *Guided Summarization(GS)*

The dataset consists of 20 first level headers(section headers): article [utf8]inputenc

- **Family/Social History (fam/sochx)**
- **History of Present Illness (genhx)**
- **Past Medical History (pastmedicalhx)**
- **chief complaint (cc)**
- **Past Surgical History (pastsurgical)**
- **allergy**
- **review of systems (ros)**
- **medications**
- **assessment**
- **exam**
- **diagnosis**
- **disposition**

- **plan:**
- **Emergency Department Course (edcourse)**
- **immunizations**
- **imaging**
- **Gynecologic History (gynhx)**
- **Procedures**
- **Other History (other_history)**
- **Laboratory Results (labs)**

In GS,we use the section headers as a prefix in the training data to guide the summarization.

B. Pre Fine Tuning

Pre-finetuning involves training a pretrained model on a related task before fine-tuning it, for the target task.In our case it is text summarization. This additional step provides the model with relevant domain-specific knowledge, potentially improving its performance and adaptability for the specific task at hand.

In our code we are performing PFT by using : generate_data_vanilla: Generates data for single-stage Bart training. Here, preprocessing functions collectively handle tasks such as cleaning text, formatting conversation transcripts, loading data from files, and preparing data for model training by saving it in a format suitable for Bart-based summarization tasks. They are designed to be modular and adaptable, allowing for customization and integration into a data processing pipeline for single-stage Bart training.

IV. Our Input

After training the above mentioned model we have tried to implement sentiment analysis over the summaries generated.

- Example 1:"The patient is detected with brain tumour and has very less chances of survival"
  This summary generated would reflect very sad situation.The patient might need mental strength to fight back from the situation.
- Example 2:"The patient looks very fit today. Seems like she has been following the diet well."
  This summary generated would reflect a happy situation.

We have tried analyzing the sentiments across all the summaries generated

V. Instructing model to perform required task

1) Pre-trained BART Model is downloaded from this `Link`.
2) Single Stage Fine-tuning has been done.
3) preparing the data into lines of text format.
4) bpe tokenization and binarization.
5) model training.
6) BART model selection and inference.
7) evaluation.

VI. Results

ROUGE-N(Recall-Oriented Understudy for Gisting Evaluation): It measures the number of matching n-grams between the model-generated summary and a reference summary.

TABLE I
OUR MODEL

Metric	Precision	Recall	F1-Score
ROUGE-1	0.4265	0.4640	0.4066
ROUGE-2	0.2032	0.2151	0.1910
ROUGE-L	0.3315	0.3650	0.3173

These are the ROUGE scores generated for our BART-PFT-GS model. These are the ROUGE scores(F1-Score) as

TABLE II
PAPER MODEL

Metric	ROUGE-1	ROUGE-2	ROUGE-L
Score	0.4204	0.1759	0.3485

given in the paper.

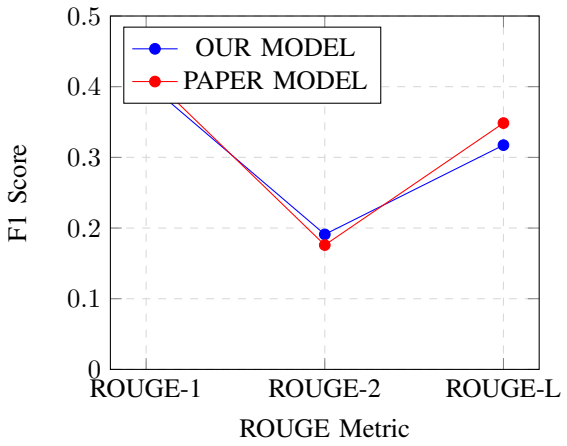

Fig. 1. **Comparison of F1 Scores for ROUGE Metrics**

42	The patient is not on any blood thinners.	neutral
43	The patient is a 26-year-old African-American male who presented to ABC orthopedics with left kne	sadness
44	The patient has had a lot of problems at school. He is having a lot of problems at school. He has b	worry
45	1. Cardiovascular status post bypass surgery.	worry
46	She is not on any medications.	neutral

Fig. 2. **Results of Sentiment Analysis**

VII. Conclusion

- The graph clearly illustrates that our model closely aligns with the model described in our reference paper.
- We can observe that the sentiments predictions is not as expected as for example,

VIII. Future Work

- Given the unavailability of a suitable dataset for sentiment analysis on doctor-patient conversations, there's potential for such datasets to emerge in the future.

- Alternatively, we could explore generating synthetic data to fulfill this need.

IX. References

1) An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters
2) MTS-Dialog Dataset
3) Code for our model
4) inproceedingszhang-etal-2021-leveraging-pretrained, title = "Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations", author = "Zhang, Longxiang and Negrinho, Renato and Ghosh, Arindam and Jagannathan, Vasudevan and Hassanzadeh, Hamid Reza and Schaaf, Thomas and Gormley, Matthew R.", booktitle = "Findings of the Association for Computational Linguistics: EMNLP 2021", year = "2021",