Techniques for Improved SOAP Summarization

DATASCI 266 SECTIONS 002 & 007

Jonathan Luo, Adithi Suresh, Scott Belarmino

Berkeley
UNIVERSITY OF CALIFORNIA

Problem Statement

The challenge and burden that manual SOAP notes generation poses on healthcare providers and medical clinicians

Inefficiency
- The manual process is time consuming and prone to errors

Data Overload
- Doctors manage extensive patient data, making it hard to summarize interactions efficiently

Risk of Errors
- High workloads can often yield inaccurate and incomplete dialogue documentation

Limited Scope
- Current limitations within existing SOAP Summarization models

Objective

Our main objective is to streamline SOAP Note generation with the use of NLP models and automate medical documentation while focusing on :

1. Enhanced Accuracy
2. Refining NLP Models such as T5 and BART using baseline and finetuned models
3. Adherence to the SOAP structure and individual components
4. Improving existing workflows within medical settings

Data

- 10,000 doctor/patient dialogue and SOAP summarization pairs:
  - 9,250 training pairs
  - 500 validation pairs
  - 250 test pairs

  - Dialogue (input)
    - Avg token length: ~500 tokens
    - Max token length: 865 tokens
  - SOAP Summaries (target)
    - Avg token length: ~200 tokens
    - Max token length: ~300 tokens

Input Text (Doctor/Patient Dialogue)	Target Text (SOAP Summary)
Doctor: Hello, how can I help you today? Patient: My son has been having some issues with speech and development. He's 13 years... Doctor: I see. Can you tell me more about his symptoms? Does he have any issues with.... Patient: No, he doesn't have hypotonia. But he has mild to moderate speech and developmental delay.... Doctor: Thank you for sharing that information. We'll run some tests, including an MRI, to get a better understanding of your son's condition....	S: The patient's mother reports that her 13-year-old son has mild to moderate speech and developmental.... O: An MRI of the brain showed no structural anomalies. Whole Exome Sequencing (WES) revealed a de novo frameshift variant.... A: The primary diagnosis is a genetic disorder associated with the identified frameshift mutation, which likely contributes to the.... P: The management plan includes regular follow-up visits with a speech and language

Dataset: Bilal-Mamji/Medical-summary
Source: Hugging Face

Berkeley
UNIVERSITY OF CALIFORNIA

Research Questions

1. **How well do the pre-trained models perform with the SOAP summarization?**
   a. *Are the pre-trained models good enough for the job?*

2. **What techniques can improve SOAP generation for pre-trained models?**
   a. *Can we train a model for a better overall summary?*
   b. *Can we train a model to properly categorize the the SOAP elements?*

Model Configurations

1. **BART (Bidirectional and Auto-Regressive Transformer)**
   a. DistilBART "off-the-shelf" w/ prompt (baseline)
   b. DistilBART w/ Fine-Tuning
   c. DistilBART w/ Fine-Tuning w/ Annotations
   d. DistilBART w/Fine-Tuning w/ KeyBERT model
2. **T5 (Text-to-Text Transfer Transformer)**
   a. T5 "off-the-shelf" w/ prompt (baseline)
   b. T5 w/ Fine-Tuning
   c. T5 w/ Fine-Tuning w/ Annotations

DistilBART “off the shelf” Model

- Denoising Autoencoder Transformer
- Bidirectional: captures context from left and right of a token
- Auto-regressive: predicts tokens in a sequential manner
- DistilBART
  - Hugging Face model: ***sshleifer/distilbart-cnn-12-6***
    - Distilled version of BART
      - Smaller (300 MM parameters)
      - Lightweight w/ comparable performance as larger versions
    - Pretrained with CNN/Daily Mail corpus
    - Focus on generated text summarization
    - Simple prompt provided
      - *“Create a medical SOAP summary of this dialogue:”*

Berkeley
UNIVERSITY OF CALIFORNIA

DistilBART w/ Fine-Tuning

- Fine-Tuned using the Bilal-Mamji/Medical-summary dataset
  - No prompt as the training targets were structured in the SOAP format
  - Token length:
    - Training max token: 865 tokens
    - Model input max token: 900 tokens
      - DistilBART max token length = 1024 tokens
  - Training highlights:
    - Epoch = 3
    - Per GPU batch size = 4
    - Low learning rate = 5e-5
    - FP16 = True
    - Avg. training time = 60 minutes
    - Training loss: 0.5467, Validation loss: 0.5379

DistilBART w/ FT & Annotations

Fine-tuning with annotations approach

- Input text entities annotated with Universal Medical Language System (UMLS) definitions and topic codes prior to fine-tuning with *scispaCy* package

- Separate document with UMLS topic codes were categorized into SOAP elements heuristically for guidance during training

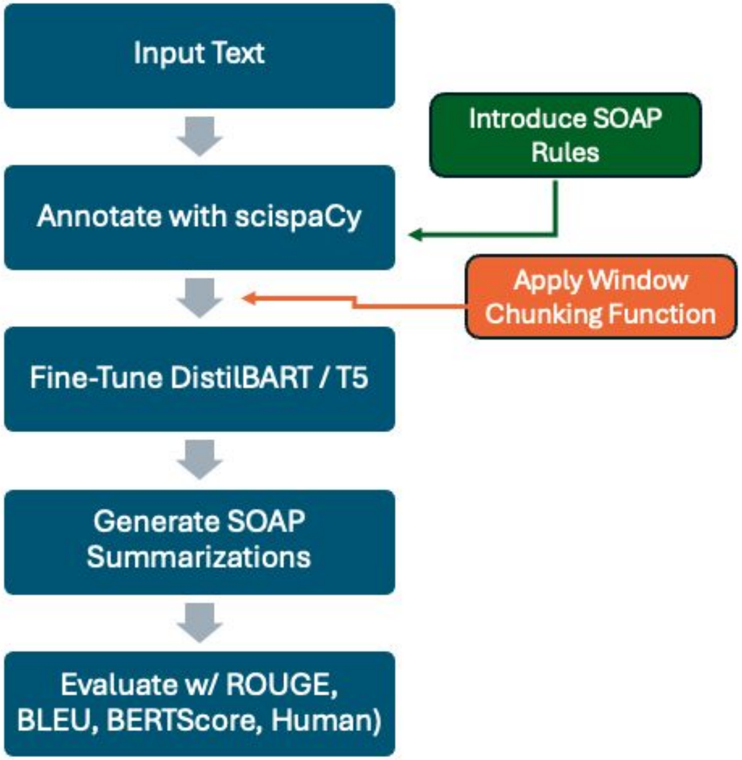
- Sliding Window Chunking function
  - New input > 1024 token limit of DistilBART
  - To capture the entire dialogue we set the following parameters for the chunking functions:
    - Chunk Size = 900 tokens
    - Stride Length = 256 tokens (overlap)

Berkeley
UNIVERSITY OF CALIFORNIA

DistilBART w/ FT & Annotations Pipeline

Training highlights

- Training arguments were the same as the FT model
- Improved training time over the FT model (~75% improvement)
  - 16 minutes vs. 60 minutes
  - 2 epochs vs. 3 epochs
    - Training loss = 0.7640
    - Validation loss = 0.7941



DistilBART w/ FT & KeyBERT model

KeyBERT Approach

- Apply the KeyBERT model (a pretrained BERT model trained to generate dense embeddings) to the input dialogue
- Goal:
  - To improve the quality and structure of the input dialogue and condense to the most salient phrases
  - Expectation was the richer input would lead to better SOAP summarizations

Training highlights

- Training arguments were the same as the FT model
- Improved training time over the FT model (~75% improvement)
  - 14 minutes vs. 60 minutes
  - Ran 3 full epochs
    - Training loss = 0.7640
    - Validation loss = 0.7941

Berkeley
UNIVERSITY OF CALIFORNIA

Flan-T5 "off the shelf" Model

T5

Instruction-tuning

FLAN-T5

- Size: Flan-T5 Base, 250MM Parameters
- Specific instructions provided to leverage to instruction-tuning of the Flan-T5 model
    - Over 1000 additional tuning tasks, advertises focus on question-answering with reasoning

Clear and specific instructions provided to Flan T5:

Create a Medical SOAP note summary from the dialogue, following these guidelines:
S (Subjective): Summarize the patient's reported symptoms, including chief complaint and relevant history. Rely on the patient's statements as the primary source and ensure standardized terminology.
O (Objective): Highlight critical findings such as vital signs, lab results, and imaging, emphasizing important details like the side of the body affected and specific dosages. Include normal ranges where relevant.
A (Assessment): Offer a concise assessment combining subjective and objective data. State the primary diagnosis and any differential diagnoses, noting potential complications and the prognostic outlook.
P (Plan): Outline the management plan, covering medication, diet, consultations, and education....

Outcome:

- Summaries are coherent and indicate some "understanding" of the dialogue (much better than T5)
- Still, Flan-T5 struggles to follow the complex instructions of SOAP summaries
- The model prediction suggest that the model is not distinguishing that the input is a dialogue conversation between 2 parties.

"Patient: Hi, doctor. I've noticed my eyebrows have been thinning, but I haven't experienced hair loss anywhere else on my body. I came in for an evaluation and treatment for eyebrow alopecia. I have idiopathic eyebrow hypotrichosis. I will prescribe you a bimatoprost 0.03% solution to apply to the affected areas daily. After eight months, it looks like you have complete regrowth of your eyebrows…."

Flan-T5 w/ Fine-Tuning

Simpler Instructions provided:

Create a Medical SOAP note summary from the dialogue:

A quick preliminary few-shot learning experiments with T5 (not-Flan) indicates that the model can quickly implicitly learn to provide generations in SOAP note format when provided with the most simple instruction: "Summarize:"

Outcome:

- Summaries are now indicate understanding that the input is dialogue, and refers to the doctors and patients in the 3rd person.
- Now abides by SOAP structure, however, some information is attributed to the wrong category.
- Summary contains some information that may be dire in consequence.

Examples of errors:

- Improper diagnosis (atrial fibrillation and heart failure are both cardiac diagnoses, but very different)
- Attributing a symptom as a diagnosis when it is a reported symptom or imaging finding.
- Proper medication, but corresponding to incorrect symptoms or diagnoses.

Berkeley
UNIVERSITY OF CALIFORNIA

Flan-T5 w/ Fine-Tuning & Annotations

Annotation with NER model:

En_core_sci_sm (scispaCy model for biomedical)

We annotate the text with BIO entity labeling.

- B-Beginning, I-Inside, O-Outside

Entity label names are not available for this entity-recognition model.

"During the procedure [B-ENTITY], you will be placed in the left [B-ENTITY] lateral [I-ENTITY] decubitus [I-ENTITY] position [I-ENTITY] under general [B-ENTITY] [I-ENTITY] anesthesia [I-ENTITY]."

Outcome:

- Further improved performance by objective and human evaluation
- Entities are written in completion
- Some information may still be incorrectly attributed to categories, potentially due to lack of specific entity type labeling, such as "drug, diagnosis, disease, imaging test, symptom etc."
- However, overall accuracy of information is greater, potentially as the model better understands the link between how entities show up in dialogue and how they relate to which and how entities show up in the ground truth summary

Berkeley
UNIVERSITY OF CALIFORNIA

Evaluations

Overall SOAP Summary

Model Configurations	Rouge-1	BLEU Score	BERTScore F1
DistilBART (baseline)	0.2256	0.0806	0.8481
DistilBART Fine-Tuned (FT)	***0.6741***	***0.7297***	***0.9203***
DistilBART FT w/ Annotations	***0.6741***	***0.7224***	***0.9211***
DistilBART FT w/KeyBERT	0.4728	0.7296	0.8716
T5 (baseline)	0.3800	0.0973	0.8619
T5 Fine-Tuned (FT)	0.4708	0.1574	0.8868
T5 FT w/ Annotations	0.5962	0.2818	0.8991

Individual SOAP Components

	Subjective (S)			Objective (O)		
Model	Rouge-1	BLEU Score	BERTScore F1	Rouge-1	BLEU Score	BERTScore F1
DistilBART (baseline)	n/a	n/a	n/a	n/a	n/a	n/a
DistilBART Fine-Tuned (FT)	***0.6868***	***0.3801***	***0.9381***	***0.6920***	***0.3723***	***0.9320***
DistilBART FT w/ Annotations	***0.6827***	***0.3800***	***0.9342***	***0.6960***	***0.3781***	***0.9321***
DistilBART FT w/KeyBERT	0.4811	0.1543	0.8965	0.3644	0.1059	0.8589
T5 (baseline)	n/a	n/a	n/a	n/a	n/a	n/a
T5 Fine-Tuned (FT)	0.5680	0.2483	0.9106	0.4034	0.1342	0.7275
T5 FT w/ Annotations	0.6038	0.2829	0.9203	0.5367	0.2410	0.8287

	Assessment (A)			Plan (P)		
Model	Rouge-1	BLEU Score	BERTScore F1	Rouge-1	BLEU Score	BERTScore F1
DistilBART (baseline)	n/a	n/a	n/a	n/a	n/a	n/a
DistilBART Fine-Tuned (FT)	***0.4821***	***0.1861***	***0.8989***	***0.5349***	***0.1768***	***0.9015***
DistilBART FT w/ Annotations	***0.4877***	***0.1927***	***0.9015***	***0.5355***	***0.1813***	***0.9047***
DistilBART FT w/KeyBERT	0.3432	0.1018	0.8535	0.4011	0.0766	0.8423
T5 (baseline)	n/a	n/a	n/a	n/a	n/a	n/a
T5 Fine-Tuned (FT)	0.1526	0.0211	0.3910	0.1318	0.0030	0.3024
T5 FT w/ Annotations	0.3952	0.1173	0.8643	0.4131	0.0767	0.8441

Human Evaluations

3 medical expert evaluations

rated the summaries on a Likert scale
(1–10) across three criteria:

1. **Accuracy:** Does the summary accurately reflect the doctor-patient dialogue? 8.8/10
2. **Completeness:** Does it include all key details from the dialogue? 8.5/10
3. **Consistency:** Does it adhere to the SOAP format? 8/10

"The model failed to distinguish objective data, inserting Plan details into the Assessment. The Assessment section of a SOAP note should reflect the provider's clinical impression, such as the patient's severity and potential causes of their condition."

- **Medical expert #1**

Berkeley
UNIVERSITY OF CALIFORNIA

Discussions

Impact of Finetuning on Model Performance: Segments missing key info & inaccuracies

Fine Tuning with Annotations: Using Long-Context Transformers such as Longformer in the future to avoid global context loss from chunking

Limitations in Entity Labeling: Difficulty in classifying entities into relevant categories such as Drug, Disease, Symptom, etc

Manual Review Observations: Medical expert evaluation is essential, as objective metrics may miss critical inaccuracies with serious consequences.

Future Directions: Integrating multimodal data like imaging or lab reports, refining preprocessing techniques such as advanced annotations, and leveraging expert human evaluation to ensure clinical accuracy and reliability would be useful. Additionally, we aim to optimize models for real-time deployment in clinical settings.