Techniques for Improved SOAP Summarization

Scott Belarmino
School of Information
University of California, Berkeley
scottbelarmino@berkeley.edu

Jonathan Luo
School of Information
University of California, Berkeley
jonnyluo@berkeley.edu

Adithi Suresh
School of Information
University of California, Berkeley
adithi_suresh@berkeley.edu

***Abstract*— In this paper, we propose methods to generate SOAP (Subjective, Objective, Assessment, and Plan) notes from doctor patient conversations using state-of-the-art natural language processing models. SOAP notes are very important in the clinical documentation since they provide a structured way of recording a patient's encounter. We also show that fine-tuned architectures such as T5 and BART are capable of producing coherent summaries that follow the SOAP format. In a second phase, we assess the performance of the models in terms of the ability of the models to summarize text in the right categories following the SOAP framework. In order to improve the quality and relevance of the output notes, we performed several trials with medical annotations and keyword identification.**

Our findings highlight the importance of fine-tuning pre-trained models to improve performance and streamline clinical workflows.

I. Introduction

The SOAP (Subjective, Objective, Assessment, and Plan) summarization technique has long been a healthcare standard. The SOAP summary is a structured form of medical documentation that provides a structured format for the key components of doctor-patient encounters. This captures key details regarding a symptoms, clinical findings, diagnoses, and treatment plans. While summaries have been manually generated in the past, this process can be time-consuming and prone to error.

Automated SOAP note generation, driven by advancements in Natural Language Processing (NLP) and pretrained models, has revolutionized the conversion of unstructured dialogues into clear medical summaries, easing doctors' workloads. However, it is important to critique these systems' abilities to maintain accuracy of medical terminology, completeness of the summarization, and adherence to the SOAP structure.

This paper explores the automation of SOAP notes using NLP models, including encoder-decoder architectures like T5 and BART. We deploy and evaluate five distinct models, analyzing their effectiveness. Using carefully selected evaluation metrics, we evaluate fine-tuned and enhanced versions of our baseline models. We employ a two-step approach: first, identifying top-performing models based on summarization abilities, and then refining them to complete each SOAP component while maintaining structural integrity. This involves preprocessing techniques such as medical annotations and keyword extraction.

We compare model-generated summaries with test summaries to ensure alignment with the SOAP framework. Our evaluation process focuses on the model's ability to categorize outputs into the distinct components of the SOAP framework: Subjective (S), Objective (O), Assessment (A), and Plan (P), to ensure both structure and meaning in our outputs. The following sections detail our approach, models, results, and insights from our findings.

II. Related Works

(Goo and Chen, 2018) were one of the early researches to review dialogue summarization by reviewing dialogue acts or interactive cues between individuals during a meeting. The process of dialogue summarization has evolved to other domains such as medicine. (Joshi et al., 2020) approached medical dialogue summarization, finding a balance between copying their source text and generating novel content. The authors explore capturing elements of medical domain in open-ended conversation. In Narnoune et al., 2021, researchers use a BERT clinical knowledge graph to extract and analyze information from clinical notes.

Inspired by these past studies, especially in the medical domain, we wanted to experiment with developing a SOAP summarization model using existing pretrained models and applying novel techniques. While many of these models have been trained very well with many large corpusus, we set out to improve them prior to fine-tuning by influencing their input dialogue by discovering entity relationships (NER, KeyBERT) and complementary annotation. Like some of these prior works, we implement strategies to identify key medical extracts, those relevant to useful, effective SOAP notes.

III. Dataset

The dataset we use for the purposes of our study is the Bilal-Mamji/Medical-summary dataset, sourced from HuggingFace. This dataset contains doctor-patient dialogue input along with the ground truth summary for each interaction. In addition, SOAP summary instructions are added. In the dataset, these are the Input, Output, and Instruction, respectively.

The Input section entails a dialogue between a doctor and a patient displayed as a script. The output section showcases the SOAP notes created from the corresponding patient-doctor dialogue. The instructions section has a specific prompt to assist the model with the task of summarization.

The dataset is pre-split into training (9250), validation (500), and test (250) sets. These sets are crucial in conducting apt model training and evaluation.

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

Fig. 1. Sample of training input text (doctor/patient dialogue) and ground truth target text (SOAP summary).

IV. Models

A. DistilBART Models

1) ***DistilBART Baseline****:* The DistilBART model, specifically sshleifer/distilbart-cnn-12-6 from HuggingFace, was used as a baseline to compare against three fine-tuned versions: DistilBART w/ Fine-Tuning (FT), DistilBART w/ FT + Annotations, and DistilBART w/ FT + KeyBERT. This model, a distilled version of the BART (Bidirectional and Auto-Regressive Transformer) architecture, was selected for its balance of efficiency and performance. With 300 million parameters (compared to BART-large's 406 million), it is lightweight yet excels in summarization tasks. The model's pretraining on the CNN/Daily Mail dataset makes it well suited for generating detailed, multi-sentence summaries—ideal for summarizing doctor-patient dialogues into SOAP notes. Models trained on datasets like XSum, which emphasize single-sentence abstractive summaries, were not considered appropriate for this task.

In the baseline configuration, the off-the-shelf DistilBART model used a simple prompt to guide the generation process into the SOAP format, structuring the output into Subjective, Objective, Assessment, and Plan sections. For the fine-tuned versions, this explicit prompt was excluded because the input dataset already included targets formatted in SOAP style, allowing the models to learn the structure directly from the data.

2) ***DistilBART w/ Fine-Tuning****:* This approach fine-tunes the DistilBART baseline model using the Bilal-Mamji/Medical-summary dataset. Data preprocessing was streamlined with HuggingFace's BartTokenizer, setting the input maximum token length to 900 to accommodate the longest doctor-patient dialogue in the dataset. This ensured no context was lost during training. The target token length was capped at 600, although most ground truth SOAP summaries were well below this limit. Training arguments were carefully selected to balance memory efficiency and effective learning. Key parameters included a per-GPU batch size of 4, three training epochs, a low learning rate of 5e-5 to prevent overshooting, and mixed-precision (16-bit floats) for faster training and reduced memory usage.

For consistency across experiments, these training configurations were applied to all fine-tuned models. Similarly, inference parameters remained constant. Generated summaries were limited to a maximum of 500 tokens, with beam search (num beams=4) used to enhance output quality.

3) ***DistilBART w/ FT + Annotations****:* Building on the DistilBART w/ FT model, this approach incorporates medical annotations from the Universal Medical Language System (UMLS), a comprehensive biomedical knowledge base developed by the National Library of Medicine (Neumman et al., 2019). UMLS annotations identify and categorize medical entities within the dialogue, enriching the input and reducing ambiguity. SOAP-specific rules were applied by manually mapping UMLS entity groups to SOAP components. For example, "short of breath" was categorized as Symptom (S), while "albuterol" was categorized as Plan (P). With over 50 UMLS entity groups covering a wide range of medical terminology, these rule based annotations are designed to encourage the model to properly categorize SOAP elements along with the target summarization.

The annotations increased input token lengths beyond DistilBART's 1,024 token limit. To address this, a sliding window chunking function divided inputs into overlapping chunks with a 256-token stride to maintain context. This preprocessing pipeline (see Fig. 2) ensured the model could process long dialogues while preserving critical relationships and entity continuity, aiming to surpass baseline performance in generating high-quality SOAP notes.

4) ***DistilBART w/ FT + KeyBERT****:* The DistilBART w/ FT + KeyBERT model was designed to enhance input text for improved SOAP summarization. KeyBERT, a BERT-based keyword extraction model, was used to distill dialogues into their most relevant details, reducing noise and simplifying the input for DistilBART.

This approach hypothesizes that preprocessing the dialogue to highlight key points before DistilBART's own denoising process enhances the accuracy and relevance of the generated summaries.

B. T5 Models

1) ***T5 Baseline****:* We utilize the FLAN-T5 base model, an enhanced version of the T5 model

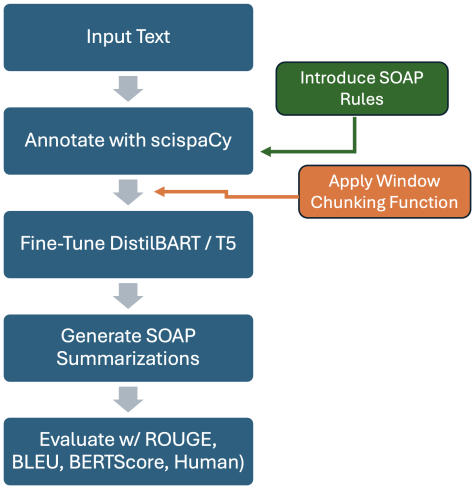

Fig. 2. Encoder-Decoder Models w/ FT + Annotations. Annotation process incorporates user SOAP based rules for input text entities. The Sliding Window Chunking function is applied during input preprocessing.

instruction-tuned to improve its ability to follow various instructions. Leveraging T5 architecture allows another experiment with architecture to test our hypotheses for later fine-tuning. As a text-to-text transformer model with 250 million parameters, we hypothesize that providing the input dialogue along with SOAP summary instructions will allow the model to produce strong baseline results. SOAP has a relatively simple but specific structure, which further increases the confidence of achieving effective understanding of the structured summary task through few-shot learning.

In the baseline configuration, the off-the-shelf FLAN-T5 model, like the DistilBart baseline, used a simple prompt to guide the generation process into the SOAP format, structuring the output into Subjective, Objective, Assessment, and Plan sections. As stated earlier, for the fine-tuned versions, this explicit prompt was excluded because the input dataset already included targets formatted in SOAP style, allowing the models to learn the structure directly from the data.

T5 is generally capable of handling longer sequences seen in the input dialogue data, so we set the token maximum length to 1024. We use the T5 tokenizer for the baseline and fine-tuned T5 models to ensure compatibility.

2) ***T5 w/ Fine-Tuning****:* We fine-tune the baseline model using the Bilal-Mamji/Medical-

summary dataset. Because of memory constraints and the lower limits of input sequence length for training vs inference, we set the token length limit to 900, and the label limit to 600 (same as with DistilBART). As we standardize hyperparameters across all experiments, the training configurations are also the same as with DistilBART (batch size of 4, three training epochs, learning rate of 5e-5 etc.)

Ultimately, we provide the model the entire length of patient-dialogue conversations as well as the ground truth SOAP summaries, which should leverage few-shot learning capabilities to not only provide accurate summaries, but also determine how to effectively structure the summary into SOAP categories.

3) ***T5 w/ FT + Annotations****:* In the annotation approach, we enhance the input data by adding entity labels using a pre-trained named entity recongition (NER) model, *en_core_sci_sm*, which is specialized for the biomedical context. We disable tagging and parsing features, decreasing computational cost and focusing on the benefit of labeling important medical concepts. The labels are appended to the pre-tokenized dialogue immediately after the corresponding words using Begining, Inside, Outside (BIO) structure. This input is then tokenized.

Because of the significantly increased length of the sequence, we chunk the training data to remain within the constraints of memory and T5 model limitations. Each chunk is capped at 900 tokens, and the sliding windows overlap over 256 tokens to increase context preservation over chunks of the same patient-doctor encounter dialogue.

To accurately infer with the model, we also apply NER labeling to the inference (test and validation) data. However, we do not chunk the data used for inference, as feeding incomplete dialogue input into the model is likely to cause incomplete summaries that miss key information. Fortunately, T5 can sufficietly intake the longer, annotated sequences for summary generation.

V. Results and Learnings

Models	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

Fig. 3. Evaluation metrics for all models

A. Overall SOAP Summarization Evaluation Results

The performance of baseline "out-of-the-box" models, using a simple prompt of "Summarize the following dialogue using the SOAP framework:" (which included descriptions of each SOAP element to guide the model), was underwhelming, yielding low evaluation scores across all metrics (Fig. 3). We hypothesize that this subpar performance is attributable to DistilBART's pretraining, which is specifically optimized for generic summarization tasks and not well-suited for handling detailed, task-specific instructions. However, a relatively higher T5 baseline indicates its few-shot learning ability with SOAP instructions, and some of the generated summaries did sometimes indicate SOAP structure resemblance. Overall, the long prompt may have overwhelmed the model, although even simpler prompts did not improve results. Generally, it appears that the baseline models are not well-trained to handle longer sequences, and fall extremely short of long enough summaries to capture all the key extracts that a SOAP note requires. As such, BERTScore may appear high, but may fail to penalize the ommision of some key important information.

In contrast, the best-performing model, DistilBART FT (fine-tuned) and DistilBART FT with Annotation, achieved evaluation scores significantly than the baseline. Both models produced summaries that matched 67% of the words in the ground truth summaries (ROUGE-1 score), indicating that they effectively captured the same information while leaving room for some variability

in phrasing. Similary, the fine-tuning and annoted version of the T5 models also indicate an improvement in ROUGE scores. These models demonstrate notably higher semantic alignment with the ground truth, as evidenced by their BERTScore F1 values of 92.03% (DistilBART FT), 92.11% (DistilBART FT + Annotation), 88.68% (T5 FT), 89.91% (T5 FT + Annotation) both significantly outperforming the baseline.

With DistilBART, despite their near-identical performance, the additional effort required to incorporate SOAP-specific rules during the annotation process did not deliver the anticipated benefits. While annotations may have helped the model better understand entity relationships and reduce ambiguity, they may also have introduced noise or been deprioritize in favor of reference data during training. Additionally, the process of adding a sliding window chunking function does have a drawback in that it may have trouble handling non-adjacent chunks, thus losing global context. Based on these findings, DistilBART FT without annotations appears to be the more efficient option for implementation.

However, with T5, a similar annotation process was applied. However chunking was not applied to T5 for inference. As such, global context is not lost and al evaluation metrics improved. It suggests that entity labeling may allow the model to identify important key medical concepts.

Although fine-tuning significantly improved performance over baselines, not all fine-tuned variants performed equally well. For example, the DistilBART FT with KeyBERT preprocessing achieved better results than the baseline but underperformed compared to the other fine-tuned models, with a lower ROUGE-1 score (47.28%) and BERTScore F1 (87.16%). We hypothesize that this drop in performance is due to over-distillation of the input dialogue by KeyBERT, which reduced the input to its most relevant components. While the intent was to denoise and structure the input prior to fine-tuning, this double denoising process likely diminished the model's ability to fully understand the dialogue during training.

B. Human Evaluation

In addition to evaluating NLP metrics, we conducted a qualitative assessment by having three medical experts review the generated summaries from our fine-tuned models. The experts analyzed a sample of 5 doctor-patient dialogues from the test set and rated the summaries on a Likert scale (1–10) across three criteria: 1.Accuracy: Does the summary accurately reflect the doctor-patient dialogue? 2.Completeness: Does it include all key details from the dialogue? 3. Consistency: Does it adhere to the SOAP format (Subjective, Objective, Assessment, Plan)?

The models achieved strong overall scores: Accuracy (8.8/10), Completeness (8.5/10), and Consistency (8/10). However, the evaluators highlighted some issues with mislabeling elements from the dialogues. For instance, Dr. James Belarmino noted, "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."

Despite these issues, the evaluators were impressed with the model's performance and agreed that SOAP summarization would benefit their practice.

C. Individual SOAP Evaluation Results

Focusing on the individual elements of SOAP, we observed notable differences between the models. The baseline models did not effectively produce distinct SOAP structure, and are excluded from this discussion. Among the five fine-tuned models, were notable differences in Rouge-1 and F1 scores across the Subjective/Objective and Assessment/Plan categories.

For DistilBart's Subjective and Objective elements, both models captured 68–70% of the ground truth words and achieved a strong semantic alignment with a 93% F1 score. However, performance was weaker for Assessment and Plan, with only 48% and 53% word matching, respectively, and a 90% F1 score. (Fig 4) The lower Rouge-1 scores may stem from the inherent overlap between the Assessment and Plan categories. In medical SOAP notes, the Assessment summarizes Subjective and Objective elements to arrive at a diagnosis, while the Plan outlines the tests and treatments related to that diagnosis (Socrates et al.,

Model	Subjective (S)			Objective (O)		
	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

Model	Assessment (A)			Plan (P)		
	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

Fig. 4. Individual SOAP components evaluation.

2023). Because both categories share diagnostic context, misclassification between them is likely.

In the T5 models, there is a significant boost in performance in the Assessment and Plan section, across ROUGE, BLEU, and BERTScore. For Assessment, ROUGE1 and BERTScore jump from 0.1526 to 0.3952 and 0.3910 to 0.8643, respectively. Likewise, for Plan, ROUGE1 and BERTScore jump from 0.1318 to 0.4131 and 0.3024 to 0.8441, respectively. The recognition of entities such as diagnoses/diseases for Assessment and medications/treatments for Plan may be explaining this significant performance boost.

VI. Discussions

It is abundantly clear that fine-tuning the models contributes significantly to model performance. The baseline models are unlikely to generate summaries long enough or with enough key information to be considered effective for clinical use. While fine-tuning does generate summaries with more key information, there are still segments of missing key information and innaccuracies.

A. Fine-Tuning with Annotation

We believe there is more to explore by developing a hybrid model to improve summarization. In our case we used a SciSpacy model for NER (Neumann et al., 2019) to modify the training set input dialogue prior to training with BART and T5. As mentioned earlier, we increased the input token sequence length due to the annotations, and to accommodate our models' input token limit, it was necessary to split the data into chunks. As an alternative we could try using Long-Context Transformers, such as Longformer which can mange 16K+ tokens. This would avoid the issue of losing global context from chunking and may yield better results.

In addition, the entity labeling annotation used for T5, while great at identifying key medical concepts as entities from manual review, did not specifically classify these entities in relevant categories such as: Drug, Disease, Imaging Test, Blood Test, Symptom etc. Such classifications are extremely relevant to the structure of SOAP. During manual review of generated summaries, both with and without annotation, we find that even when key information is included, it is sometimes not included in the correct category. Furthermore, the model may sometimes even confuse the category of an entity, even when it is labeled and applied in the summary. For example, some summaries diagnose patients with the name of a medication rather than a disease, or name a symptom as the final medical diagnosis. In other cases, medications are correctly named but attributed to the incorrect symptom or diagnosis.

VII. Conclusion

Structured summaries, such as SOAP notes, can be quite simply learned by fine-tuning widely-used architectures, making them a powerful tool for the task. However, inaccuracies and even completely hallucinated lab results and diagnoses may not only be incredibly frustrating, but cause dire consequences for patients and those involved in their care. The implementation of additional tools, such as labeling medical terms and pre-classifying them into SOAP categories can assist significantly in the performance of these models, in identifying key extracts as well as distinct SOAP structure.

Additionally, we may overcome some global context problems arising from chunking by leveraging models designed for longer sequences, which conversations tend to be. Another factor to consider is acquisition of dialogue data itself. The provided dialogue in the data is rather clean and concise, and likely not typical of the standard patient-doctor encounter. Future implementation may require greater preprocessing, and likelier,

models that can take longer input while improving summary performance.

As stated in previous section, and from what we know about summary tasks in general, purely numerical evaluation metrics are not sufficient to fully evaluate the performance of a model. The consequences of inaccurate medical information are highly consequential. Supplementing proper evaluation of summaries with medical domain knowledge is paramount.

ACKNOWLEDGMENTS

We thank the three medical experts who participated in the survey for their valuable insights and feedback. Their expertise greatly informed the evaluation of the generated SOAP summaries.

A. References

Goo, C., and Chen, Y. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *arXiv preprint arXiv:1809.05715*. Joshi, A., Katariya, N., Amatriain, X., and Kannan, A. (2020). Dr. Summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*. Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, Bouchra El Asri, BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis, Computer Methods and Programs in Biomedicine Update, Volume 1, 2021, 100042, ISSN 2666-9900, https://doi.org/10.1016/j.cmpbup.2021.100042.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159–165. https://doi.org/10.1147/rd.22.0159. Nallapati, R., Zhou, B., dos Santos, C., Gülçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *arXiv preprint arXiv:1602.06023*. Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Socrates, V., Gilson, A., Lopez, K., Chi, L., Taylor, R. A., and Chartash, D. (2023). Predicting relations between SOAP note sections: The value of incorporating a clinical information model. *Journal of Biomedical Informatics, 141*, 104360. https://doi.org/10.1016/j.jbi.2023.104360. Author manuscript; available in PMC 2024 May 01.