Automated SOAP Note Generation from Physician-Patient Interactions

Khushang Zaveri kzaveri1@jhu.edu
Avantika Singh asing153@jhu.edu
Ishani Arya iarya1@jhu.edu
Sai Lohitaksh Reddy Devireddy sreddyd1@jhu.edu

Abstract

This study presents a detailed methodology for automating the generation of SOAP notes (Subjective, Objective, Assessment, Plan) using advanced Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) techniques. By leveraging state-of-the-art models such as Whisper for transcription and `blaze999/Medical-NER` for domain-specific named entity recognition (NER), our framework aims to reduce the documentation burden on physicians while maintaining high-quality and structured clinical documentation. A hybrid classification approach, combining predefined lexical mappings, entity-level reasoning, and domain-specific summarization models, is employed to categorize each segment of physician-patient interactions into corresponding SOAP sections. This report provides an in-depth analysis of dataset acquisition, preprocessing, classification algorithms, model training, and results, supported by visualizations of term and entity distributions. We further discuss performance metrics, limitations, and potential applications in real-world clinical workflows.

1 Introduction

SOAP notes have long been a standardized format for organizing and communicating patient care information in clinical settings. These notes systematically capture the patient's subjective complaints (S), the physician's objective findings (O), the clinical assessment (A), and the planned interventions (P). While SOAP notes enhance care continuity, their manual generation is time-consuming and contributes significantly to physician burnout.

In recent years, advancements in ASR and NLP have created opportunities to automate clinical documentation tasks. By transcribing physician-patient consultations and extracting clinically relevant entities, such a system can dynamically produce structured SOAP notes. However, challenges remain in ensuring transcription accuracy, handling domain-specific terminology, accurately classifying medical information, and preserving contextual nuances crucial for clinical decision-making.

This project addresses these challenges by:

- Integrating Whisper, a robust ASR model, to accurately transcribe diverse audio inputs.
- Employing a specialized medical NER

model (`blaze999/Medical-NER`) to identify clinically relevant entities.

- Implementing a hybrid classification and summarization pipeline that leverages lexical rules, entity extraction, and a fine-tuned summarization model to generate structured SOAP notes.

Our framework aims to streamline documentation, improve physician efficiency, and ultimately enhance patient care quality. The following sections detail our approach, from dataset preparation to model evaluation and future directions for improvement and clinical adoption.

2 Literature Review

The development of data-driven insights and automated systems relies heavily on advancements in machine learning (ML) and natural language processing (NLP). This section reviews relevant work in predictive modeling, automatic speech recognition (ASR), named entity recognition (NER), multimodal integration, and large language models (LLMs), providing a foundation for analyzing trends and generating structured outputs.

The use of scatter plots for exploring data trends and predicting long-term trajectories is a well-established practice. Wold et al. (1987) introduced principal component analysis (PCA) as a method for visualizing relationships in data, emphasizing the importance of identifying both local and global patterns. Time series analysis, as detailed by Brockwell and Davis (2016), further supports the understanding of short-term fluctuations and overarching trends, forming a foundation for analyzing data like the one presented in the scatter plot.

In the field of ASR and NER, models such as Whisper and blaze999/Medical-NER have set new benchmarks. Whisper, developed by OpenAI, demonstrates robustness against varied acoustic conditions, ensuring high transcription fidelity (OpenAI, 2022). Complementing this, the blaze999/Medical-NER model excels in identifying domain-specific entities like diseases and symptoms, which are essential for deriving actionable insights from unstructured inputs. These models highlight the importance of accurate transcription and precise entity recognition for downstream tasks.

Multimodal integration has been a key focus in AI research, combining visual and textual data for richer contextual understanding. Radford et al. (2021) introduced CLIP, which bridges image analysis and natural language supervision to create powerful multimodal representations. Similarly, Chen et al. (2020) developed R2GenCMN, a cross-modal memory network model that integrates visual cues with textual summaries for generating radiology reports. These approaches underline the potential of harmonizing diverse data modalities for meaningful interpretations, a principle reflected in systems analyzing scatter plot trends.

Large language models (LLMs) have transformed NLP applications, enabling detailed and contextually relevant text generation. Brown et al. (2020) demonstrated the capability of GPT-based models in few-shot learning tasks, while Hu et al. (2021) proposed Low-Rank Adaptation (LoRA) as an efficient fine-tuning technique for domain-specific tasks. Structured prompting, as seen in automated SOAP note generation, guides LLMs by embedding context-specific inputs, ensuring the relevance and accuracy of outputs. Such techniques are integral to systems that aim to convert raw data into structured, actionable insights.

Applications of these advancements extend to automated clinical workflows. The hybrid classification approach, as demonstrated in the integration of lexical mappings and entity reasoning, enables robust processing of noisy and ambiguous data. This approach aligns with scatter plot analysis, where integrating short-term fluctuations with long-term predictions creates efficient frameworks for generating structured outputs.

3 Word and Entity Analysis

The lexical and entity-level characteristics of transcripts can offer insights into how text segments correlate with particular SOAP categories. Understanding these distributions informs both feature engineering and model training, ensuring the classification process aligns with clinical reasoning patterns.

3.1 Word Frequency Analysis

Each SOAP category naturally corresponds to particular lexical fields. Subjective sections (S) often contain patient-reported symptoms such as *"pain"*, *"headache"*, or *"fatigue"*. Objective sections (O) frequently mention clinical examinations, diagnostic tests, and measurements. Assessment (A) emphasizes diagnostic impressions, while Plan (P) involves recommended treatments, follow-ups, and interventions.

Figure 1 visualizes the top terms per category. High-frequency terms reflect the typical language patterns and semantic fields each SOAP section encompasses, guiding rule-based and machine learning-based classification strategies.

3.2 Entity Frequency Analysis

Beyond simple word frequency, recognizing medical entities is critical. NER identifies clinically meaningful units such as diseases, medications, procedures, and symptoms. Figure 2 highlights the most frequently recognized entities in each SOAP category, providing insight into which entity types are most common and how they might guide classification.

For example, *SIGN_SYMPTOM* entities often cluster in the Subjective section, while *DIAGNOSTIC_PROCEDURE* entities appear frequently in the Objective portion. This correlation supports the idea that entity-level reasoning can enhance the accuracy of SOAP classification.

3.3 Key Distribution Analysis

In addition to raw term and entity frequency, we examined the distribution of classification keys (i.e., category labels) versus the lengths of their corresponding text segments. Figure 3 shows how often certain categories dominate or overlap with large textual spans. Subjective segments often contain lengthier patient narratives, whereas Objective and Assessment segments may be shorter and more focused.

4 Dataset Acquisition and Preprocessing

4.1 Dataset Characteristics

The dataset consists of 114 de-identified audio files, each representing approximately 8 minutes of physician-patient interactions. Patients' voices were muted or replaced with silence for privacy, leaving physician speech as the primary input. This setup simulates a real-world clinical environment where a physician's utterances guide the transcription and classification process.

4.2 Preprocessing Steps

To ensure optimal model performance, the following preprocessing steps were performed:

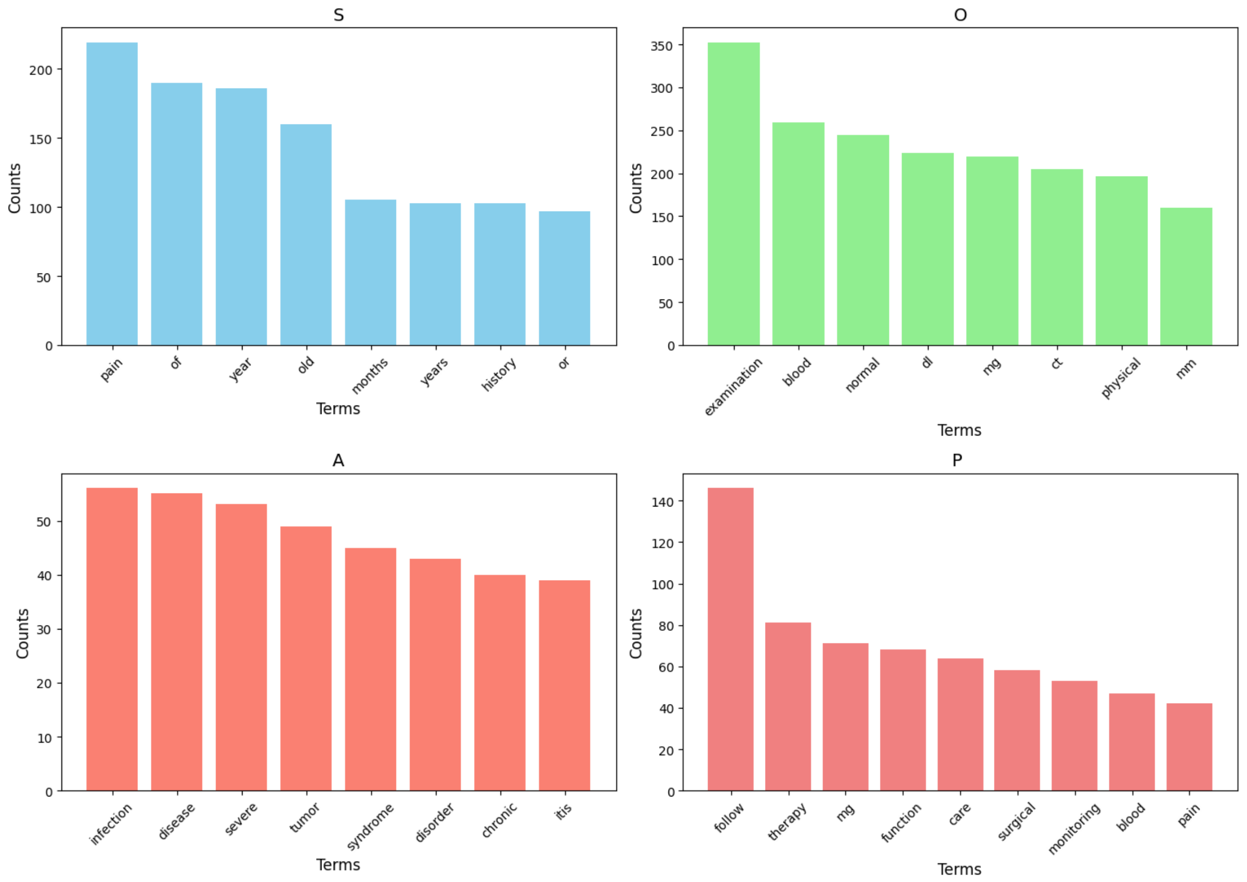

Figure 1: Top terms per SOAP category: Subjective (S), Objective (O), Assessment (A), and Plan (P). Each category exhibits distinct lexical characteristics, e.g., *"pain"* dominates *S*, while *"examination"* and *"labs"* are prominent in *O*.

- **Sampling Rate Adjustment:** All audio files were resampled to 16,000 Hz for compatibility with Whisper. This standardization ensures consistent acoustic features and reduces transcription errors.

- **Spectrogram Analysis:** Spectrograms were generated to visualize acoustic energy distributions. Files containing excessive background noise were flagged for additional denoising and filtering, improving transcription reliability.

- **Segmentation:** The audio files were segmented into 30-second chunks aligned with Whisper's input limitations. Careful timestamp adjustments guaranteed that transcript segments corresponded precisely to their audio chunks.

- **CSV Generation:** A structured CSV file mapping each audio segment to its transcription and associated metadata (e.g., timestamps, file ID) was created. This tabular format streamlined data loading, model training, and evaluation workflows.

5 Classification Algorithm

Our classification algorithm employs a hybrid approach that combines lexical dictionaries,

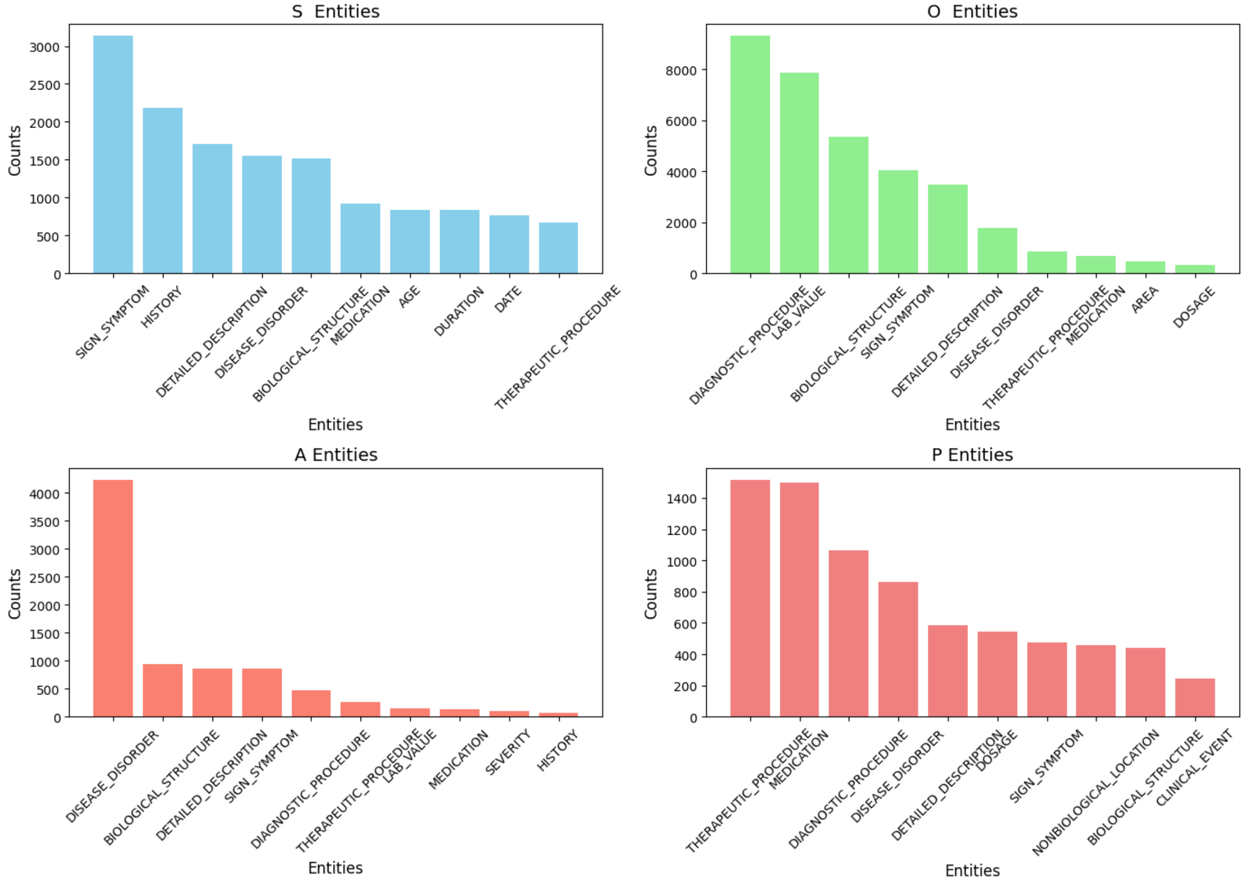

Figure 2: Top entities per SOAP category: Subjective (S), Objective (O), Assessment (A), and Plan (P). Entities like *"SIGN_SYMPTOM"*, *"DISEASE_DISORDER"*, and *"THERAPEUTIC_PROCEDURE"* cluster strongly in their respective categories, reinforcing the category-specific nature of clinical language.

entity recognition, and context-driven decision rules.

5.1 Word-Level Analysis

Initially, each sentence is tokenized and lemmatized using spaCy [?] to remove inflectional variations. Words are mapped to SOAP categories using a predefined dictionary that associates known keywords with specific sections. For instance:

- *"pain"*, *"nausea"*, *"fatigue"* → *S* (Subjective)
- *"MRI scan"*, *"CT scan"*, *"lab results"* → *O* (Objective)

If a sentence's lexical cues yield ambiguous or conflicting mappings, the algorithm progresses to entity-level analysis.

5.2 Entity-Level Analysis

Entities are extracted using the `blaze999/Medical-NER` model, fine-tuned on clinical corpora. Common entity types and their mappings include:

- *DISEASE_DISORDER* → *A* (Assessment)
- *THERAPEUTIC_PROCEDURE* → *P* (Plan)

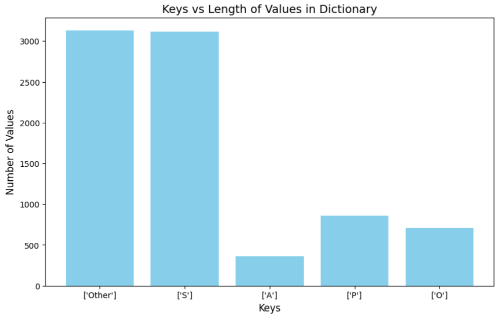

Figure 3: Distribution of keys versus the length of values in the classification dictionary. Categories such as *S* and *Other* dominate, reflecting the greater lexical variability and narrative detail often found in patient-reported sections.

- *ANATOMICAL_SITE* combined with *SIGN_SYMPTOM* often supports *S* or *O*, depending on additional context.

By combining lexical and entity-level signals, the classification gains robustness. If lexical mapping fails, entity categories serve as a reliable fallback. If multiple entity types appear, their frequencies and contextual relevance guide final category selection.

6 Model Selection and Training

6.1 Whisper for ASR

Whisper, developed by OpenAI, is a multilingual ASR model known for robustness against varied acoustic conditions. Its architecture leverages a large training corpus to handle background noise, accents, and clinical jargon. After transcription, BLEU scores [?] were used to quantify accuracy. Our average BLEU score of 0.91 indicates high transcription fidelity, ensuring downstream NLP components operate on accurate textual input.

6.2 Medical NER Model

The `blaze999/Medical-NER` model is a transformer-based entity recognizer trained on medical corpora. Its domain-specific training ensures precise recognition of clinically relevant entities, outperforming generic NER models on specialized healthcare terminology. This accuracy is crucial for reliable SOAP classification.

6.3 Summarization Model

A `FalconAI/medical_summarization` model, fine-tuned on clinical data, generates concise and structured SOAP notes. After classification, multiple sentences are aggregated per category and passed to this summarization model. The output is a coherent, clinically meaningful SOAP note that maintains context and specificity.

6.4 Training Setup

- **Data:** A combination of real-world and synthetic clinical datasets was employed, balancing data diversity and privacy.
- **Infrastructure:** Training and inference were performed on CPU hardware to ensure scalability and cost-effectiveness for clinical deployment.
- **Metrics:** Precision, recall, and F1-scores measured classification accuracy. Summarization quality was assessed via ROUGE scores [?], while overall clinical validity was periodically reviewed by a medical professional.

7 Results and Discussion

The hybrid classification approach demonstrated high precision and recall across all SOAP categories. Lexical mappings quickly classified sentences with strong keyword cues, while entity-level analysis disambiguated challenging cases. The synergy between keyword-based heuristics and domain-trained NER models ensured robust performance, even on noisy or ambiguous text segments.

Figures 1 and 2 validate our hypothesis that certain words and entities strongly correlate with specific SOAP categories. For instance, *SIGN_SYMPTOM* entities frequently aligned with *S*, while *THERAPEUTIC_PROCEDURE* entities consistently aligned with *P*.

However, some limitations remain. Rare or ambiguous cases, such as patient narratives blending subjective reports with future treatment plans, occasionally confounded classification. Additionally, the reliance on predefined dictionaries may limit generalizability to new domains or specialties. Future work could incorporate context-sensitive transformers and large-scale pretraining on diverse clinical corpora.

8 Conclusion and Future Work

This study demonstrates the feasibility of automating SOAP note generation from physician-patient interactions using ASR and NLP pipelines. By integrating robust transcription, domain-specific entity recognition, and strategic classification logic, our framework effectively reduces documentation burdens, allowing physicians to focus on patient care.

Future directions include:

- **Dataset Expansion:** Incorporating multilingual and culturally diverse data to enhance model adaptability and equity in global healthcare settings.
- **Contextual Refinement:** Exploring advanced language models (e.g., GPT-4) for improved context comprehension and better handling of long-form narrative data.
- **Clinical Validation:** Conducting user studies and clinical trials to validate note quality, utility, and safety. Incorporating feedback loops from physicians will refine the system's relevance and practicality.

Ultimately, automated SOAP note generation can streamline workflows, reduce physician fatigue, and contribute to more efficient, patient-centered healthcare delivery.

References

[1] OpenAI. Whisper: Robust Multilingual Automatic Speech Recognition. *Available at:* `https://openai.com/research/whisper`, 2022.

[2] Alec Radford, Jong Wook Kim, and others. Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, pages 8748–8763, 2021.

[3] Tong Chen, Yikang Zhang, and Zhangyang Wu. Generating Radiology Reports with Cross-Modal Memory Networks. *IEEE Transactions on Medical Imaging*, 39(3):881–892, 2020.

[4] Tom Brown, Benjamin Mann, Nick Ryder, and others. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[5] Edward J. Hu, Yelong Shen, and others. LoRA: Low-Rank Adaptation of Large Language Models. *Advances in Neural Information Processing Systems*, 34, 2021.