AI in Healthcare Final Project Proposal

Elrich MIRANDA **Emma SCHAPIRA** **Tonghan WEN**

1 Problem statement

The problem we are investigating is the extraction and classification of medical decisions from discharge summaries.

1.1 Relevance of this study

Physicians document a wide range of patient care decisions in clinical notes, but their unstructured nature makes analysis and retrieval challenging. AI-driven extraction of medical decisions can transform clinical practice by enabling large-scale analysis, identifying trends, ensuring guideline adherence, and enhancing BioNLP research. This contributes to explainable AI in healthcare by capturing clinical reasoning. Beyond medicine, it can also inform health policy and insurance regulations by assessing treatment alignment with guidelines, detecting inefficiencies, and evaluating regulatory impacts.

1.2 Complexities and challenges

This task is complex due to two key challenges: accurately detecting medical decision spans in clinical notes and classifying them while preserving medical terminology's meaning. Technical hurdles include class imbalance, which biases model learning, and handling long texts, as most NLP models are limited to 512 tokens.

2 Data Overview

We shall be using data sourced from MIMIC (a publicly available database containing de-identified clinical records). In particular, the MIMIC-III dataset contains 451 discharge summaries, covering more than 54,000 sentences. Each summary contains multiple medical decisions categorized into ten different decision types : e.g. drug-related decisions, therapeutic procedures, evaluation of test results, and legal/insurance-related decisions. These decisions were manually annotated by two expert annotators, with disagreements resolved by a third senior annotator.

The specific aspects of the data that we plan to utilize include: decision spans and their categories for model training and evaluation, annotated summaries to analyze decision-making patterns across different diseases and metadata on patient demographics to explore potential biases.

There are multiple potentials obstacles we foresee with the data, such as class imbalance, long clinical texts and ambiguity in decision classification.

3 Planned Methodology

We plan to use MedDec: A Dataset for Extracting Medical Decisions from Discharge Summaries (Elgaar et al., 2024) as a baseline guide to extract key medical decisions from medical reports and decision summaries via span-detection methods.

There are several approaches that are under consideration:

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- Fine-tuning BERT-based models such as BioClinicalBERT, RoBERTa and ELECTRA and implementing a multi-class classification where each token is assigned one of several predefined medical decision categories
- Evaluate zero-shot and one-shot LLM prompting methods to extract structured decisions from summaries and investigate their ability to g across different phenotypes of diseases
- Combining rule-based filtering with deep learning models and then use unsupervised clustering to identify decision-making patterns across patient groups

3.1 Leveraging existing studies

The reference paper contains annotated medical decisions across ten categories and proposes a sequence chunking method that processes lengthy discharge summaries by breaking them into manageable 512-token segments. We can leverage the annotation schema and chunking approach into our pipeline to further fine-tune various LLMs for our task.

Besides MedDec, there are other studies that have developed annotated medical datasets which we can use to test our models, such as CLIP (Mullen-bach et al., 2021) which is a dataset of MIMIC-III summaries annotated with seven types of action items and MDACE (Cheng et. al., 2023), a dataset of clinical notes annotated with ICD codes (the International Classification of Diseases). This step can be carried out provided we gain access to these annotations.

3.2 Potential improvements to existing work

Here are some potential improvements over existing work related to this task that we could study and try to implement:

- Improving the detection of span boundaries by incorporating unified medical language system (UMLS)
- Fine-tune models for specialized medical domains (e.g., cardiology, oncology)
- To reduce bias in the results towards specific disease phenotypes, apply data augmentation for low-represented categories.

4 Evaluation of the Results

In general, we plan to evaluate our model's performance across two main dimensions: token level and span level, as well as within the match to each decision category. We shall apply standard evaluation metrics for assessment: accuracy for token-level evaluation and F1 score for span-level and decision category evaluations.

- **Span Exact Match**: Measures the model's ability to predict spans with both correct boundaries and categories.
- **Token Accuracy**: Assesses the prediction of decision categories at the token level. This metric is more flexible, allowing partial overlaps with true spans.
- **Decision Categories**: Medical decisions can be grouped into distinct categories, facilitating structured organization and enhancing patient comprehension of medical transcripts. Evaluating the model's performance in each category helps determine its capability to accurately classify medical decisions.

In order to evaluate the model's capacity of generalization, we also plan to compare F1 score performance of span detection at seen phenotype with unseen phenotype. This comparison could allow us to check when a novel disease appears, whether our model could still show ideal performance in classifying their categorization with high accuracy.