Clinical Note Summarization with Quality-Aware Analysis

Group 4: Ruju Shah, Bhargav Pamidighantam

1. Description

Healthcare professionals generate vast amounts of clinical documentation daily, including discharge summaries, progress notes, and patient encounters. Reading through lengthy clinical notes is time-consuming and can delay critical decision-making in patient care. Automated summarization of clinical text can significantly reduce documentation burden while preserving essential medical information.

We aim to build models that can generate high-quality summaries of clinical notes while maintaining medical accuracy and completeness. Our project will compare extractive and abstractive summarization approaches, analyzing their performance across different medical specialties and developing confidence estimation techniques to assess summary quality automatically.

2. Dataset

We will use the **Medical Transcriptions dataset** from Kaggle (https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions), which contains approximately 5,000 medical transcription samples across various medical specialties. Each row includes:

- **transcription**: Full clinical text from medical encounters
- **medical_specialty**: The medical field (e.g., Cardiology, Surgery, Radiology)
- **description**: Brief summary/description that can serve as our target summary
- **sample_name**: Identifier for the transcription type

The dataset covers 40+ medical specialties and is already de-identified, making it suitable for academic research without privacy concerns.

3. Methodology and Expected Results

We have selected two models in order to solve this problem

3.1 Model A: BioBART (Pretrained Transformer)

- **Architecture**: Encoder-decoder Transformer (BART-based)
- **Pretraining**: Trained on biomedical corpora including PubMed abstracts and MIMIC-III clinical notes
- **Model Source**: HuggingFace (cambridgeltl/BioBART)

- **Fine-tuning**: Trained on our dataset using supervised summarization objective
- **Tokenizer**: SentencePiece (subword-level)
- **Framework**: PyTorch with HuggingFace transformers and Trainer API

3.2 Model B: Custom Seq2Seq with Attention (From Scratch)

- **Architecture**: Bi-directional LSTM encoder + LSTM decoder with Bahdanau Attention
- **Training**: From scratch on the available dataset
- **Tokenizer**: Word-level or subword (learned from training corpus)
- **Loss**: CrossEntropyLoss with teacher forcing
- **Optimizer**: Adam
- **Epochs**: 20–30 with early stopping
- **Framework**: PyTorch or TensorFlow

Expected Results:

- ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) comparing generated summaries to reference descriptions
- Confidence scores indicating model certainty in generated summaries
- Performance analysis across different medical specialties
- Visualizations showing model performance, specialty-wise analysis, and summary length distributions

4. Timeline

- **Week 9**: Data exploration, cleaning, and preprocessing. Create train/validation/test splits stratified by medical specialty.
- **Week 10 & 11**: Implement extractive and abstractive summarization models. Fine-tune models and establish baseline performance metrics.
- **Week 12**: Optimize models, implement confidence estimation, and conduct comprehensive evaluation across medical specialties.
- **Week 13**: Perform final analysis, create visualizations, write the report, and prepare the final presentation.

5. Responsibilities

There are two members in our team: Ruju Shah and Bhargav Pamidighantam. Each member will contribute equally to this project.

Ruju Shah will be mainly responsible for all model training activities, including implementing both extractive summarization models (BERT-based sentence ranking) and abstractive summarization models (T5/BART fine-tuning), hyperparameter optimization, and model performance monitoring throughout the training process.

Bhargav Pamidighantam will handle data preprocessing and cleaning, building the comprehensive evaluation framework (ROUGE metrics, medical entity preservation analysis), implementing confidence estimation techniques, conducting error analysis across medical specialties, creating all visualizations, and leading the report writing and presentation preparation.

Both members will collaborate on initial data exploration, final results interpretation, and project coordination to ensure seamless integration of training results with evaluation and analysis.

6. Conclusion

This project presents a comparative analysis between a pretrained transformer (BioBART) and a custom-built LSTM model for summarizing clinical notes. BioBART is expected to outperform across ROUGE/BLEU metrics, but the custom model offers interpretability and a valuable baseline.

By systematically comparing metrics and sample outputs, we aim to make an informed decision on which model better suits clinical summarization under different constraints (e.g., performance, deployment feasibility, or interpretability).