NLP for Healthcare - CL3.411
Assignment 2

Medical Entity Recognition and Linking

Deadline: 14th September, 2025 11:59 PM

1 Objectives:

Develop an entity-linking pipeline that identifies medical entities in de-identified clinical discharge summaries and links them to SNOMED CT concept IDs.

2 Exploratory Analysis: [5 Marks]

Inspect samples: Examine a few discharge notes and their associated annotation spans to understand the annotation format and labeling.

3 Baseline: [20 marks]

Build a two-step system that reads free-text discharge summaries, identifies clinical entities, and then links each entity to the correct SNOMED CT concept IDs.

3.1 Stage 1 - Named Entity Recognition (NER):

Train a Named Entity Recognition (NER) model to identify and extract relevant medical entities accurately from discharge summaries.

3.2 Stage 2 - Linker:

1. **Candidate Generation:** For each extracted span, generate a list of possible SNOMED CT concept IDs.
2. **Candidate Selection:** Evaluate the generated candidates and select the single best-matching concept ID for the entity.

4 Experiments: [35 marks]

4.1 NER + Dictionary-based Linking:

Train the NER model to extract the spans and examine the OMOP vocabulary set & map the extracted spans to concept IDs using synonyms, abbreviations, and simple linguistic rules.

4.2 NER + similarity matching:

Train the NER model to extract the spans and generate the list of possible SNOMED CT concept IDs, and select the single best-matching concept ID for the entity using similarity matching.

4.3 LLM-based:

1. **Entity Extraction with LoRA-Fine-Tuned LLM:** Fine-tune a Large Language Model (LLM) using LoRA on annotated discharge notes to automatically identify and extract clinical entities (text spans, start/end indices) from raw clinical text.
2. **Candidate Concept Retrieval Using Faiss:** Build a vector database of SNOMED CT concept descriptions using an embedding model and Faiss; for each extracted entity span, retrieve the top-N most similar candidate concepts.
3. **Concept Classification with LLM:** Supply the retrieved candidates and the entity span to the fine-tuned LLM, which will then choose the most appropriate SNOMED CT concept ID, producing the final linked output in the challenge format.

5 Evaluation:

Evaluation is based on a class-level macro-averaged character Intersection-over-Union (IoU), defined for character-level predictions $\boldsymbol{P}$ and ground-truth classifications $\boldsymbol{G}$ as:

$$\text{IoU}_{\text{class}} = \frac{P_{\text{class}}^{\text{char}} \cap G_{\text{class}}^{\text{char}}}{P_{\text{class}}^{\text{char}} \cup G_{\text{class}}^{\text{char}}}$$

$$\text{macro IoU} = \frac{\sum_{\text{class} \in P \cup G} \text{IoU}_{\text{class}}}{N_{\text{class} \in P \cup G}}$$

Where $P_{\text{char,class}}$ is the set of characters in all predicted spans for that class, $G_{\text{char,class}}$ is the set of characters in the ground truth spans for that class, and classes $\in P \cup G$ represents all classes appearing in either the predictions or the ground truth.

Note: The predicted concept ID must match exactly.

6 Analysis and Report: [10 marks]

1. Analyse which of the methods gives a better score & output on the test set and why?
2. Compare 3 methods and elaborate your findings using graphs and visualizations.

7 Submission:

- Source code (Should include .py files only.)
- Report in PDF (include analysis, scores, comparison, graphs, and visualizations.)
- Output in CSV format, including each method's output under a separate header (e.g., Method 1, Method 2, ...), along with the test set concept IDs.

8 Viva: [30 Marks]

Good Luck