Investigating Approaches to Minimize the Burden of Named Entity Annotation in the Clinical Context

Name: Gabriel Herman Bernardim Andrade

Laboratory's name: Social Computing

Supervisor's name: Professor Eiji Aramaki

Abstract

The adoption of Electronic Health Records (EHR) worldwide has allowed for the collection of vast amounts of patient data. While such data contains useful information for biomedical research, its unstructured textual format makes it notoriously difficult to process automatically.

Among many Natural Language Processing (NLP) techniques, Named Entity Recognition (NER) has been used to develop systems that can extract information such as diseases, symptoms, and drugs from clinical narratives. While great advancements have been made in the accuracy of such systems, different vocabularies and writing styles used across medical specialties and institutions hinder the effectiveness of the models when implemented in a real-world scenario.

Efforts to improve the portability of biomedical models are hampered by the scarcity of publicly available labeled data, which is critical for training supervised methods. Thus, adapting NLP models to new domains often requires crafting domain-specific training datasets. Due to the labor-intensive nature of the text annotation task, such a process can become rather costly and time-consuming.

This dissertation addresses this challenge through approaches to improve the efficiency and effectiveness of NE annotation processes by (1) exploring the impact of labeled dataset size on model performance in new medical subdomains and (2) proposing methods to enhance annotation efficiency by relaxing precision on entity boundaries. Through case studies, we demonstrate that these approaches can streamline the text annotation process while creating high-quality datasets. Such findings can help advance NLP systems in biomedical applications and support healthcare and clinical research advancements.