Graduate School of Information Science　Doctoral Thesis Abstract

Lab name (Supervisor)	Social Computing (Prof. Eiji Aramaki)		
Name (surname) (given name)	Herman Bernardim Andrade Gabriel	Date	2024/07/20
Title	Minimizing the Burden of Named Entity Annotation for Clinical Applications		

The growing adoption of Electronic Health Records (EHR) within health institutions worldwide has allowed for the easier collection of vast amounts of patient information, including details about their medical history, diagnoses, treatments, and laboratory results. While EHRs can be a primary source of data for biomedical research and the improvement of healthcare procedures, most of it is comprised of clinical narratives, which, due to their unstructured free textual format, makes it rather difficult to process for secondary use.

Natural Language Processing (NLP) methods have emerged as a promising solution to such a challenge, allowing for the development of systems that can automatically analyze, understand, and extract relevant information from clinical data. Particularly in the biomedical domain, Named Entity Recognition (NER) has been used to extract entities such as diseases, symptoms, and drugs from medical texts and is a fundamental step for the successful execution of other downstream tasks, such as relation extraction, knowledge discovery, and hypothesis generation.

Despite the ubiquity of Deep Learning (DL) techniques fostering the creation of highly accurate NER systems, their implementation in a real-world clinical setting still faces obstacles. For instance, different vocabularies and writing styles used across medical specialties and institutions hinder the generalizability of these models, limiting their effectiveness when applied to a different context, even those within the healthcare domain.

While efforts have been made to enhance the portability of NER models in the biomedical field, labeled training data remains a critical bottleneck for supervised methods. The scarcity of publicly available labeled datasets to various specific medical subdomains constrains the ability to adapt NER models effectively. Consequently,

achieving high performance in new target domains often requires the creation of a purpose-specific training dataset.

Named Entity (NE) annotation, as an inherently manual process, allied to the sheer volume of data that must be meticulously labeled to produce an accurate model, makes it an exhausting and time-consuming task. Thus, effective and efficient methods for corpora annotation are required to streamline the process.

This dissertation addresses this challenge through approaches to improve the efficiency and effectiveness of NE annotation processes by (1) investigating the impact caused by the size of the labeled dataset has on the performance of the model when adapting it to a new medical subdomain and (2) proposing a method to enhance annotation efficiency through the relaxation of the annotation precision, particularly, on entity boundary definitions, aiming to reduce the annotator workload without compromising annotation quality.

Through case studies and evaluations, we demonstrate the feasibility and effectiveness of these approaches in streamlining the NE annotation process while maintaining high-quality annotations. Our findings contribute to advancing the development of NLP systems for biomedical applications, facilitating more accurate and efficient information extraction from unstructured medical data and ultimately supporting advancements in healthcare and clinical research.

論文審査結果の要旨

申請者氏名　　HERMAN BERNARDIM ANDRADE GABRIEL

本論文は，病院に集積される大量の医療電子カルテから情報抽出する際のソリューションを提供するものであり，施設間のモデルの移動，データ構築を大幅に省コスト化する新手法について研究したものである．両方の研究とも新規性が高く，社会的にも重要な課題である．本論文は，今後の医療言語処理の基礎的な理解から実際の社会への応用をカバーした学術的にも社会的にも意義の高い内容であり，博士(工学)の学位論文としての価値があるものと認める