NAMED ENTITY RECOGNITION IN ELECTRONIC HEALTH RECORDS FOR ALPORT SYNDROME

Gabrielė Skirmantaitė[1], Gražina Korvel[1], Rimantė Čerkauskienė[2], Agnė Čerkauskaitė-Kerpauskienė[2]

[1]Vilnius university, Faculty of Mathematics and Informatics, Vilnius, Lithuania
[2]Vilnius university, Faculty of Medicine, Vilnius, Lithuania
gabriele.skirmantaite@mif.vu.lt

Rare diseases pose considerable diagnostic difficulties due to their infrequent occurrence, heterogeneous clinical presentation, and overlap with more prevalent conditions, often resulting in delayed or incorrect diagnoses that negatively affect patient management. Alport syndrome, a hereditary disorder affecting renal, auditory, and ocular systems, illustrates the clinical importance of the timely identification of disease-related information in patient records.

The increasing adoption of electronic health record (EHR) systems has led to the accumulation of large volumes of clinical data, a substantial proportion of which is recorded as unstructured free-text. Although structured EHR components can be readily processed computationally, a significant amount of clinically relevant information is documented in narrative form and is not directly accessible for automated analysis. Named entity recognition (NER), a core task in natural language processing, addresses this limitation by enabling the identification of clinically meaningful entities, such as diseases, symptoms, treatments, and medications, from unstructured clinical text and serves as a foundational step in clinical information extraction.

Against this background, the present study examines transformer-based NER applied to unstructured clinical documentation in Lithuanian electronic health records. Lithuanian constitutes a low-resource language in the clinical NLP domain, characterized by limited availability of annotated data and linguistic properties including rich morphology, extensive inflection, and high lexical variability. These features, combined with domain-specific medical terminology and the frequent use of abbreviations in clinical documentation, introduce additional challenges that prevent direct application of methods developed for high-resource languages.

The task is formulated as token-level sequence labeling using transformer-based architectures. Model performance is evaluated using standard entity-level precision, recall, and F1 metrics. The findings provide preliminary insights into the potential application of transformer-based models for named entity recognition in Lithuanian clinical documentation and indicate the relevance of language-specific methodological adaptation in low-resource clinical settings.

Keywords: Named Entity Recognition, Electronic Health Records, Natural Language Processing, Rare Diseases, Alport Syndrome, Low-Resource Languages