Application of Artificial Intelligence to the Analysis of Unstructured Clinical Text in Estonian

Summary

The purpose of this doctoral research is to develop novel methods based on artificial intelligence and natural language processing that will enable the automatic analysis and structuring of unstructured clinical texts in Estonian. In healthcare settings, vast amounts of text—such as patient histories, discharge summaries, physician's notes, and test reports—are generated every day, but their full potential remains untapped due to the absence of reliable language-processing models for Estonian. Existing NLP tools neither support Estonian nor provide the contextual sensitivity required for clinical content, which makes it essential to create an innovative solution tailored to Estonian clinical text analysis. This project will explore how AI tools can be adapted to the specific characteristics of Estonian medical language, evaluate the accuracy of different NLP when handling clinical information in Estonian, investigate ways to integrate automatically generated structured data into Estonia's healthcare information systems, and identify the practical and ethical challenges involved in developing and deploying such systems.

Research field:	Information and communication technology
Supervisor:	Gunnar Piho
Availability:	This position is available.
Offered by:	School of Information Technologies Department of Software Science
Application deadline:	Applications are accepted between June 01, 2025 00:00 and June 30, 2025 23:59 (Europe/Zurich)

Description

The research

The volume of unstructured, text-based data in healthcare is growing rapidly, and processing it effectively demands new technological approaches. As a low-resource language, Estonian currently lacks robust natural language processing (NLP) and medical terminology tools, which prevents the automated structuring of health data and limits its full use in decision support and research.

This doctoral project aims to investigate and develop experimental AI-based methods for analyzing free-form clinical text in Estonian. The work will address both text-based NLP and will generate structured outputs compliant with international standards such as SNOMED CT and FHIR. By highlighting the need for reliable solutions tailored to a low-resource language—where machine translation is insufficient and both contextual understanding and semantic precision are critical—the project seeks to fill a major gap in Estonian clinical data processing.

Key questions include how to adapt existing NLP tools to handle the structural peculiarities, abbreviations, and contextual references of Estonian clinical language; and how to ensure that identified clinical concepts are accurately linked to the Estonian version of SNOMED CT to enable semantic mapping and interoperable data exchange. The research will also examine the ethical, technical, and practical challenges of deploying these solutions in the Estonian healthcare system—addressing issues such as data protection, model transparency, and clinician trust and engagement.

Building on a master's thesis prototype that used Azure Text Analytics for Health, the Snowstorm terminology server, and FHIR formatting, this project will extend and refine that work. Initial tests of the prototype revealed significant shortcomings—misclassification of terms and incorrect contextual interpretation—underscoring the need for Estonian-tailored solutions. One line of inquiry will explore whether and how to train NLP models suitable for Estonian, including the evaluation of necessary corpora, annotation tools, and training methodologies. Integration with Estonian health information systems will be assessed, along with the potential impacts on clinician workflow, data quality, and decision support.

The goal is to deliver a development framework and a practical demonstration that can reliably structure Estonian clinical free text in accordance with international health data standards, thereby advancing both research and development efforts and supporting the digital evolution of Estonia's healthcare system..

Responsibilities and (foreseen) tasks

- Develop an analytical framework for the automated processing of Estonian clinical free text, encompassing NLP and SNOMED CT mapping components
- Map and select appropriate case studies representing different types of clinical documentation in Estonian healthcare (e.g., general practitioners, radiologists, specialists)
- Collect and anonymize text corpora in collaboration with healthcare institutions and clinicians, and perform data annotation and validation
- Develop and test experimental NLP models for Estonian medical language
- Conduct comparative evaluations between customized systems using metrics based on precision, recall, and F1-score
- Create and test a prototype capable of structuring free text and outputting results in FHIR- or SNOMED CT-based data formats
- Collaborate with clinicians and healthcare IT experts to gather qualitative feedback and assess the solution's applicability in clinical workflows
- Participate in research and development seminars
- Publish scientific articles in international journals and present at conferences on NLP and medical AI topics.

Applicants should fulfil the following requirements:

- Master's degree in information and communications technology
- A clear interest in the position's topic
- Excellent command of English
- Strong, demonstrable writing and analytical skills
- Ability to work both independently and as part of an international team
- Willingness and capacity to assist with organisational tasks relevant to the project

The following experience is beneficial:

- Programming
- Working knowledge of SQL
- Working knowledge of software engineering
- Working knowledge of healthcare standards
- Working knowledge in AI and NLP

The candidate should submit a research plan on the chosen topic, detailing the overall studies, research and publication strategy. They may expand on the listed research questions and tasks and propose the theoretical foundations to be employed.

We offer:

- A fully funded, 4-year PhD position at eMedLab (https://taltech.ee/en/emedlab), TalTech's interdisciplinary centre for digital health innovation.
- Access to state-of-the-art research infrastructure and collaborative opportunities within the European Federation of Medical Informatics (https://efmi.org/).
- Opportunities for conference travel, international research stays, and networking with leading universities.

About eMedLab

eMedLab is a research group in digital health and medical informatics at Tallinn University of Technology (TalTech). It brings together researchers from the Centre for Digital Health and Business Information Technology research groups of the School of Information Technologies. The eMedLab research group, led by Professor Peeter Ross and Dr Gunnar Piho, consists of around twenty researchers exploring the opportunities and challenges of using human-owned and -controlled health and medical data and developing new methods and technologies to address these challenges.

Additional information

For further information, please contact Dr Gunnar Piho gunnar.piho@taltech.ee or visit https://taltech.ee/en/emedlab

To get more information or to apply online, visit https://taltech.glowbase.com/positions/1035 or scan the the code on the left with your smartphone.