Master Thesis Proposal Nr:

Topic: Semantic annotation, topic, and similarity analysis of a large corpus of clinical documentation forms

Over the last 20 years, a large corpus of documentation forms has been built up in the clinical workplace system at the University Hospital Erlangen. The forms were implemented according to the specifications of the various clinical departments, resulting in many redundancies (e.g., various forms for documenting smoking status with partial overlaps, but also individual differences). There is currently no semantic annotation of the forms with standardized terminology. Semantic annotation is time-consuming but enables automated comparison between different form types, hospital sites and even languages [2]. Several comparable data elements within the analyzed forms could enable better data reuse as demonstrated in Krumm et al. [2].

In preparation for a coordinated re-implementation of the forms in a new clinical workplace system (KAS), the existing forms are to be semantically annotated and, on this basis, systematically analyzed in terms of their content and similarity to each other.
Natural Language Processing methods (NLP) such as Language Models or Large Language Models (LLMs) are to be used for semantic annotation. SNOMED CT is currently the most comprehensive health terminology in the world, a constantly growing ontology of preferred terms and their synonyms [3]. It can be combined with LLMs for medical concept normalization, entity extraction, typing, and classification [1].

To analyze the topics and similarities, the annotation will be stored in a graph database and corresponding distance measures will be determined. The results of the work represent an essential input for the form concept in the future KAS of the University Hospital Erlangen.

Questions:
- Q1: Can the KAS corpus be semantically annotated using available NLP or LLM methods?
- Q2: Which topic complexes and similarities can be derived from the annotated corpus?
- Q3: Can references to publicly available form collections (e.g., MDM portal) be established?

Tasks:
- A1: Familiarization with the available source data (KAS), literature, and methods on medical information extraction and concept normalization.
- A2: Semantic annotation of the corpus & storage of the results in a graph database using LM- or LLM-based methods for information extraction and concept normalization.
- A3: Topic and similarity analysis and visualization of the results
- A4: Comparison of the results with the forms of the MDM portal
- A5: Code and documentation: ensure the usability (train/test) of the model on the hospital servers

The thesis must contain a detailed description of all developed and used algorithms as well as a profound result evaluation and discussion. The implemented code has to be documented and provided. Extended research on literature, existing patents and related work in the corresponding areas has to be performed.

Advisors: Prof. Dr. Bjoern Eskofier, Prof. Dr. Thomas Ganslandt (Medical supervisor), Andrea Riedel M. Sc., Dr. Emmanuelle Salin, and Arijana Bohr M. Sc.
Student: Michelle Kosminski B.Sc.
Start - End: 01.01.2025—01.07.2025

References

[1] Chang, E. and Sung, S., 2024. Use of SNOMED CT in Large Language Models: Scoping Review. *JMIR Medical Informatics*, *12*(1), p.e62924.

[2] Krumm, R., Semjonow, A., Tio, J., Duhme, H., Bürkle, T., Haier, J., Dugas, M. and Breil, B., 2014. The need for harmonized structured documentation and chances of secondary use–Results of a systematic analysis with automated form comparison for prostate and breast cancer. *Journal of biomedical informatics*, *51*, pp.86-99.

[3] https://www.bfarm.de/EN/Code-systems/Terminologies/SNOMED-CT/_node.html