Announcement of the German Medical Text Corpus Project (GeMTeX)

Frank MEINEKE, Luise MODERSOHN, Markus LOEFFLER and Martin BOEKER

University of Leipzig, IMISE, Germany
Technical University of Munich, AIIM, Germany

ORCiD ID: Frank Meineke orcid.org/0000-0002-9256-7543,
Luise Modersohn orcid.org/0000-0002-8258-4366,
Markus Loeffler orcid.org/0000-0002-0424-9933, Co-Coordinator GeMTeX,
Martin Boeker orcid.org/0000-0003-2972-2042, Coordinator GeMTeX

Abstract. The largest publicly funded project to generate a German-language medical text corpus will start in mid-2023. GeMTeX comprises clinical texts from information systems of six university hospitals, which will be made accessible for NLP by annotation of entities and relations, which will be enhanced with additional meta-information. A strong governance provides a stable legal framework for the use of the corpus. State-of-the-art NLP methods are used to build, pre-annotate and annotate the corpus and train language models. A community will be built around GeMTeX to ensure its sustainable maintenance, use, and dissemination.

Keywords. Natural Language Processing, Text Corpus, German Medical Informatics Initiative

1. Introduction

The potential of natural language processing (NLP) is quickly improving with rapid advances in machine learning esp. deep learning. The language of clinical documentation is very different in vocabulary and structure from normal written language or the language of scientific publications. Therefore, the progress of clinical NLP crucially depends on specifically trained language models, requiring authentic clinical documents. The German Medical Informatics Initiative (MII) [1] provides a unique opportunity to make clinical documents accessible on a large scale, and to enrich them with annotations. A German Medical text Corpus is expected to boost the development of NLP-resources that support German clinical text analysis [2]. GeMTeX will address two major bottlenecks that have hindered German clinical language models to date [3,4], ie. data accessibility and data annotation.
2. Main Objectives

GeMTeX will provide more than 150K clinical texts for NLP at six German university hospitals, reflecting a wide variety of regional and clinic-specific medical language styles. Annotation is performed by teams of trained student assistants and documentalists following annotation guidelines and using state-of-the-art tools. The texts represent different medical specialties and text types in a balanced way (e.g., discharge summaries, findings reports). Meta-information is added, related structured data from the data integration centers can be accessed. Besides basic annotation, in-depth annotations are performed in four medical areas (cardiology, pathology, pharmacy, and neurology), each on a sub-corpus. The annotation uses a structured annotation vocabulary based on semantic standards such as standardized terminologies (e.g., SNOMED CT, ICD-10, TNM and others), ontologies and information models (FHIR). Pre-annotations, based on existing terminologies, algorithms and models simplify and speed up the annotation process. A strong governance will adopt a scientific framework using a study protocol. Its legal regulations are based on the data protection concept and the data use regulations and contract of the MII. Access to patient data and documents is only possible where patients have agreed to the MII broad consent. Annotated documents are automatically anonymized and manually checked. Access to the document corpus for scientific projects is additionally subject to the data access regulations of the MII and must be decided by a Data Use and Access Committee at each site center. GeMTeX will provide up-to-date tools and methods for corpus generation and NLP. Depending on the research question, different integration scenarios for texts can be implemented, which allow local, distributed and central use. Distributed machine learning will be implemented within GeMTeX and in cooperation with other projects. Leading NLP companies Averbis® and ID Berlin®, participate with their respective NLP solutions in the development and use of GeMTeX.

3. Expected Results

An important advantage of GeMTeX is its open, transparent, and extensible publication, distribution, and use according to FAIR principles. Sustainability will be achieved through cooperation with the German Information Center for Life Sciences (ZB MED), which will make the generated resources (texts, methods, tools) available on a long-term basis. The corpus for German clinical texts will be accessible to scientific projects via the MII technical infrastructure and terms of use. GeMTeX will be funded from mid-2023 by [1]. Due to brevity the 17 PIs of the project are not listed in the author list.

References