Building the Infrastructure for the German Medical Text Corpus Project (GeMTeX)

Jakob FALLER[a], Christina LOHR[b], Martin BOEKER[c] and Frank MEINEKE[b1]
[a]Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany [b]Institute for Medical Informatics, Statistics, and Epidemiology, Leipzig University, Germany, [c]Institute of Artificial Intelligence and Informatics in Medicine, Medical Center rechts der Isar, Technical University Munich, Germany
ORCiD ID: Frank Meineke https://orcid.org/0000-0002-9256-7543

Abstract. The German Medical Text Project (GeMTeX) is one of the largest infrastructure efforts for German-language clinical documents. To determine the different starting points of the participating institutions, we conducted a survey regarding standards, annotation tools, privacy and availability of clinical text documents of the participating institutions We summarize the answers and measures taken to establish the project infrastructure.

Keywords. Natural Language Processing, Health Information Interoperability

1. Introduction

The GeMTeX project aims to generate the largest annotated reference corpus of German language clinical text documents for research, e.g. to train or fine tune Large Language Models [1]. We report on the establishment of the necessary technical and organizational infrastructure for the development of a multi-center text collection process at 6 university hospitals. The management of the text documents is carried out by the data integration centers (DIC) of the German Medical Informatics Initiative (MII) [2]. At project start the requirements placed on the sites were a novelty in both technical and structural terms.

2. Interview to determine the status of the text management infrastructure

The specific technical situation at each DIC was determined in a survey[2] regarding i) **Standards**: Text documents are stored in local specific health information systems. In-

[1] Author to whom correspondence should be addressed

[2] interview guideline will be published under zenodo.org when accepted. Temporary available at https://drive.google.com/drive/folders/1DKO8EmHh4S3f0nbAv44aKRBseenCVaWe

house coding is used to categorize document types (e.g., discharge summary, findings). ii) **Annotation tools**: Various tools are used for manual annotation; tools for automatic annotation are not used. iii) **Privacy**: Documents are not anonymized, patient consent is mandatory and has been obtained for all personal data used in GeMTeX (MII Broad Consents) [2], but the required number of patient consents MII (Broad Consents) could not yet be provided at most sites. iv) **Text availability**: Export of text from HIS is proprietary or not established. Availability varies by document type and department.

3. Results and Outlook

We responded to these survey results as follows: i) We initiated a process to define a FHIR-based standard for document referencing in the national MII core data set[3] using established code systems, e.g. for categorization. Text bodies are stored in plain text, annotations in UIMA CAS format. ii) We rolled out the same annotation software (INCEpTION) in all DIC, which allows the integration of recommender systems [4]. The primary automatic recommender is the text mining and machine learning platform Health Discovery from Averbis[4]. All 6 sites recruited and trained teams of a total of 32 medical student annotators. iii) The process of obtaining patient consent has developed well at all locations (currently over 40.000 consents) **-** the basis for the provision of texts. The study protocol was submitted with a privacy policy and accepted by the IRBs at all sites necessary. All automatic annotations get controlled by at least 2 annotators with ongoing control of the inter-annotator agreement. iv) The DIC and the clinics established interfaces and ETL routes to make documents from various clinical hospital departments accessible. All sites have now implemented the presented harmonized extraction and annotation pipeline [5] and are in the process of locally de-identifying clinical documents as a prelude to the subsequent semantic annotation of medical entities during the next 18 month. [see more on GeMTeX https://www.smith.care/en/gemtex_mii]

Acknowledgement: This work was supported by BMBF within the GeMTeX project under grant 01ZZ2314B (CL, FM) and 01ZZ2314A (MB) as well as 01ZZ2314G (JF).

References

[1] Meineke F, Modersohn L, Loeffler M, Boeker M. Announcement of the German Medical Text Corpus Project (GeMTeX). Stud Health Technol Inform. 2023;302:835-836. doi:10.3233/SHTI230283

[2] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56. doi:10.3414/ME18-03-0003.

[3] Zenker S, Strech D, Jahns R, Müller G, Prasser F, Schickhardt C et al. Nationally standardized broad consent in practice: initial experiences, current developments, and critical assessment. Bundesgesundheitsbl 2024:67, 637–647. doi:10.1007/s00103-024-03878-6

[4] Klie JC, Bugert M, Boullosa B, Eckart de Castilho R and Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation], Proceedings of the 27th International Conference on Computational Linguistics, 2018:5–9, Association for Computational Linguistics.

[5] Lohr C, Matthies F, Faller J, et al. De-Identifying GraSCCo - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus. Stud Health Technol Inform. 2024;317:171-179. doi:10.3233/SHTI240853

[3] https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set

[4] https://averbis.com/health-discovery/