GraSCCo

A Fully Shareable, Multiply-Alienated German Clinical Text Corpus

Luise Modersohn*[A,B], Stefan Schulz*[C],
Christina Lohr[B] und Udo Hahn[B]

[A] AIIM, Technical University of Munich
[B] JULIE Lab, Friedrich Schiller University Jena
[C] Institute for Medical Informatics, Statistics and Documentation, Med Uni Graz

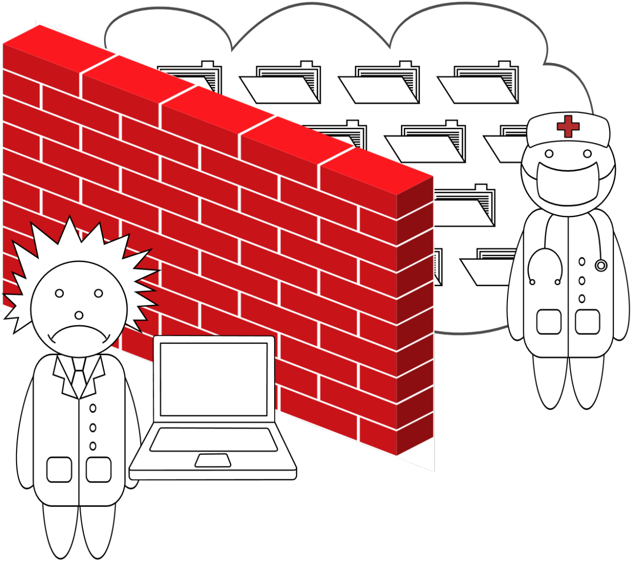
* These authors contributed equally

MEDIZIN INFORMATIK INITIATIVE

SMITH
Smart Medical Information
Technology for Healthcare

DIFUTURE
Data Integration for Future Medicine

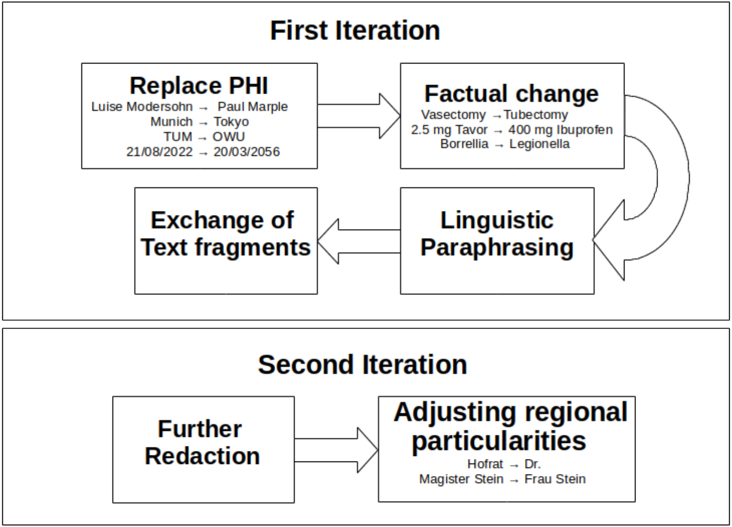
Barriers for Clinical Natural Language Processing (cNLP)



German Clinical and Medical Text Corpora

Corpus	Text genre	Original	# Docs	Shareability
3000PA$_J$ [4]	Discharge summaries	True	1,100	Non-Shareable
BRONCO150 [5]	Discharge summaries	Shuffled sentences	150	DUA
JSynCC [6]	Medical Textbooks	True	399	Code for re-creation
GGPOnc 1.0 [3]	Clinical practice guidelines	True	8,420	DUA
GGPOnc 2.0 [2]	Clinical practice guidelines	True	10,190	DUA
This work	**Case reports**	**Redacted**	**63**	**Fully Shareable**

Corpus construction



First Iteration

Replace PHI
Luise Modersohn → Paul Marple
Munich → Tokyo
TUM → OWU
21/08/2022 → 20/03/2056

Factual change
Vasectomy → Tubectomy
2.5 mg Tavor → 400 mg Ibuprofen
Borrellia → Legionella

Linguistic Paraphrasing

Exchange of Text fragments

Second Iteration

Further Redaction

Adjusting regional particularities
Hofrat → Dr.
Magister Stein → Frau Stein

Experimental Setup

Datasets

Clinical	3000PA$_J$ [4], Bronco150 [5]
Medical	JSynCC [6], GGPOnc [3], PubMed
General	Krauts [10], WikiWars_De [9]

Features

- Linguistic Features (Word count, Sentences, Stop words)
- Occurrences of UMLS [1] terms (Anatomy, Disorders, Living Being)
- Medications from *Rote Liste*

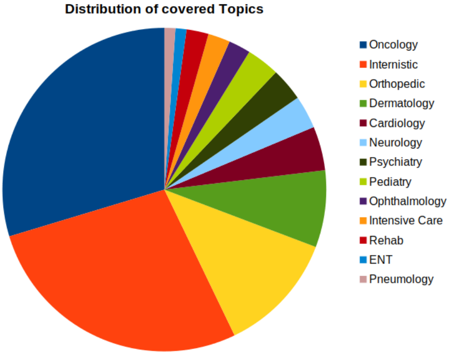
Clustering

- Max. 1000 documents per dataset
- Normalization (per document word count) and scaling
- K-means [8] and t-SNE [7]

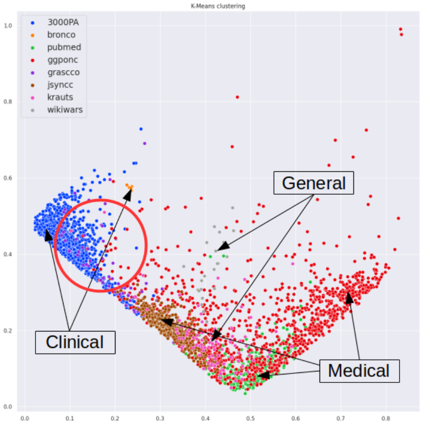
Corpus description

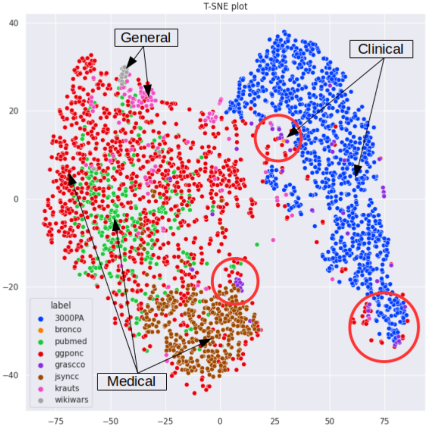
Key Data

Documents	63
Sentences	5,430
Tokens	43.667
Licence	CC0 1.0 Universal
Sex distr.	1:1
In-patient	60 %



Visualization





Conclusion

- First freely shareable German clinical corpus (Creative Commons licence)
- Preliminary evidence for closeness to clinical documents
- 63 documents, ca. 5k sentences, ca. 43k token
- Small in size → addressed in future releases
- Documents may be medically implausible

References

[1] Olivier Bodenreider. “The Unified Medical Language System (UMLS): integrating biomedical terminology”. 2004.

[2] Florian Borchert et al. “GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers”. June 2022.

[3] Florian Borchert et al. “GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines”. Nov. 2020.

[4] Udo Hahn et al. “3000PA - Towards a National Reference Corpus of German Clinical Language”. 2018.

[5] Madeleine Kittner et al. “Annotation and initial evaluation of a large annotated German oncological corpus”. 2021.

[6] Christina Lohr et al. “Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus”. May 2018.

[7] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. 2008.

[8] David J. C. MacKay. Information Theory, Inference & Learning Algorithms. 2002.

[9] Jannik Strötgen and Michael Gertz. “WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions”. 2011.

[10] Jannik Strötgen et al. “KRAUTS: A German Temporally Annotated News Corpus”. May 2018.

GraSCCo

A Fully Shareable, Multiply-Alienated German Clinical Text Corpus

Luise Modersohn*[A,B], Stefan Schulz*[C],
Christina Lohr[B] und Udo Hahn[B]

[A] AIIM, Technical University of Munich
[B] JULIE Lab, Friedrich Schiller University Jena
[C] Institute for Medical Informatics, Statistics and Documentation, Med Uni Graz

* These authors contributed equally

MEDIZIN INFORMATIK INITIATIVE

SMITH
Smart Medical Information
Technology for Healthcare

DIFUTURE
Data Integration for Future Medicine