Extraction of UMLS Concepts using Apache cTAKES for German Language

Fachhochschule Dortmund
University of Applied Sciences and Arts

M. Becker, B. Böckmann

Introduction

- Automatic information extraction of medical concepts and classification with semantic standards from medical reports is useful for standardization and for clinical research
- This is an approach for an UMLS concept extraction with a customized natural language processing pipeline for German clinical notes using Apache cTAKES
- The objective is, to test a natural language processing tool for German language if it is suitable to identify UMLS concepts and map these with SNOMED-CT

Introduction

- Main goal:
  - access knowledge from unstructured German clinical text for:
    - electronic patient records
    - research
    - personalized guideline-based treatment recommendations
  - *using an open source natural language processing Tool*
  - *using international semantic standards*
  - *using formalized german guidelines*

Motivation & Introduction

- structured vs. unstructured information
- most available tools for Textmining and also terminologies are in English language
- evaluation problem: no existing German gold standard corpus
- Formalized German Guidelines have no link to the unstructured data

Methods

- Unified Medical Language System (UMLS)
  - compendium of many controlled vocabularies in the biomedical sciences
  - Metathesaurus → Concept Unique Identifier (CUI)
  - available in German language
  - By extracting the UMLS concepts it is possible, to map this concepts to other non-German speaking classifications and ontologies

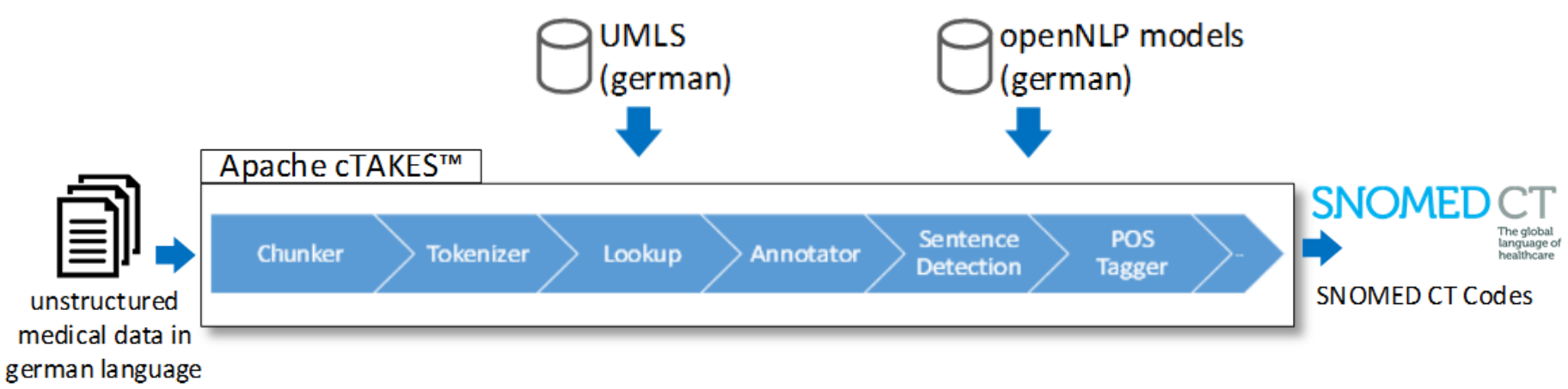
Methods

- SNOMED-CT
  - Systematized Nomenclature of Medicine - Clinical Terms
  - SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world
  - Not available in German language

Methods

- Apache cTAKES
  - One of the most proven open source natural language processing tools
  - Apache cTAKES already offers a variety of algorithms for text analysis and information extraction
  - It can normalize to domain ontologies such as SNOMED-CT using UMLS concepts

- Architecture Overview



Methods

- Gold Standard
  - ShARe/CLEF eHealth 2013 shared task 1 training set
    - 199 clinical notes
    - only in English notes
  - translated into German (Google Translator & manually)

Type	#Note	#CUI
ALL	199	2798
DISCHARGE	61	1969
ECHO	42	479
RADIOLOGY	42	257
ECG	54	93

Results

- Pipeline with English clinical notes:

Type	Recall	Precision	F1
DISCHARGE	0.71	0.27	0.39
ECHO	0.56	0.26	0.35
RADIOLOGY	0.69	0.23	0.35
ECG	0.81	0.25	0.38

- Pipeline with German translated clinical notes:

Type	Recall	Precision	F1
DISCHARGE	0.37	0.30	0.33
ECHO	0.47	0.51	0.49
RADIOLOGY	0.37	0.28	0.32
ECG	0.39	0.25	0.30

Discussion

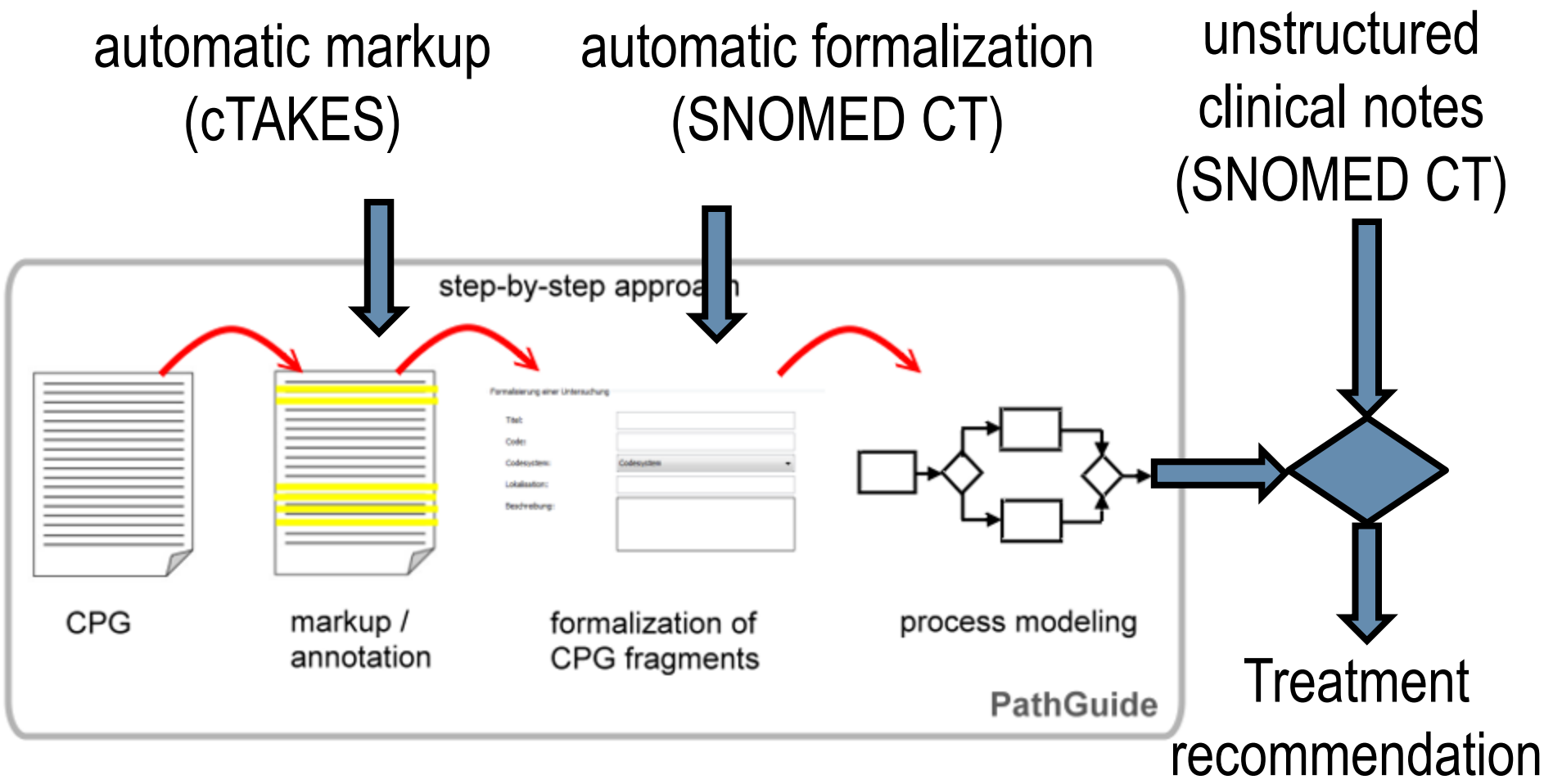
- The UMLS database is not as extensive for the German Language (196842 entries) like the English (5571374 entries) UMLS database
- German pipeline has a better average precision because it also identifies less false positive concepts
  - → F1 value on German echo notes is higher than the value of the English pipeline
- *no German stemming has been integrated into the pipeline at the time of the results publication and the German OpenNLP models have not yet been trained to medical notes*

future work

- The next steps:
  - NLP:
    - implement Context Analyzer
    - implement German stemming
    - expand UMLS Database
    - training of the openNLP models
  - Treatment recommendations:
    - combining SNOMED-CT Codes & formalized guidelines for personalized guideline-based treatment recommendations

future work

- Treatment recommendations:



Vielen Dank für Ihre Aufmerksamkeit

Fachhochschule Dortmund
University of Applied Sciences and Arts

Matthias Becker, M.Sc.
FB Informatik, Medizinische Informatik
Emil-Figge-Str. 42 - 44227 Dortmund

Matthias.Becker@fh-dortmund.de
www.inf.fh-dortmund.de