

Clinical Document representation

Stefan Schulz*, Markus Kreuzthaler**

Institut für Medizinische Informatik, Statistik und
Dokumentation

Medizinische Universität Graz

CBMed Biomarker Research, Graz

*stefan.schulz@medunigraz.at

**markus.kreuzthaler@medunigraz.at

Goals

- To optimise and enrich output of NLP analysis of **German**-language **clinical** documents
- To make it compatible with standards:
 - Ontology standards: **SNOMED** CT, LOINC
 - Standardised medical information templates: **FHIR** (provides context to semantic IDs)
- Three examples:
 1. German **Interface terminology** development
 2. Disambiguation of **short forms**
 3. Identification of **semantic relations**

1. German Interface terminology development

- Problem:
 - Diversity and idiosyncrasy of clinical language
 - Generally: ontology labels do not reflect real use of language ("Sekundäre maligne Neoplasie der Leber" vs. "Lebermetastasen")
 - Currently no German translation of SNOMED CT
- Resource:
 - Since 2014, low-resourced activities (paid medical students): token n-gram translations (EN –> DE) and annotations (POS, gender, number) from English SNOMED CT dictionary
 - Algorithmic creation of variants including single-word compounds
 - Scoring and Filtering (corpus occurrence, character sequence patterns)
 - Currently 2.4 Million terms (limit: 6 tokens)
 - Performance: same as English (term matching against annotated parallel corpus)

Core vocabulary

English	L Count	German 1	German 2	German 3	German 4
burn	1 1264	Brandverletzung NN F	Brandwunde NN F	Verbrennung NN F	
normal	1 1264	normales JJ	normenhaftes JJ		
ankle	1 1254	Knöchel NN M			
wrist	1 1251	Handgelenk NN N			
drug	1 1244	Wirkstoff NN M	Arznei NN F	Arzneimittel NN N	Droge NN F
second	1 1244	zweites JJ	Sekunde NN F	Sekunden-	%VOID% 2. %VOID%
uncertain	1 1227	unsicheres JJ			
abdominal	1 1222	abdominales JJ	Bauch-	abdominelles JJ	
membrane	1 1210	Membran NN F			
liver	1 1207	Hepar NL N	Leber NN F		
microgram	1 1202	%VOID% µg %VOID%	Mikrogramm NN N	Mikrogramm NL N	
middle	1 1193	mittleres JJ	Mitte NN F	Mittel--	
ulcer	1 1180	Ulzeration NN F	Ulkus NN N	Geschwür NN N	
upper limb	2 1180	oberes JJ Extremität NN F	Arm NN M	oberes JJ Gliedermaße NN F	OE NL F
fluoroscopic	1 1171	Durchleuchtungs-	durchleuchtungsgestütztes JJ	fluoroskopisches JJ	
effect	1 1170	Effekt NN M	Auswirkung NN F	Wirkung NN F	Folge NN F
service	1 1158	Service NN M	Dienst NN M	Service NN N	
vehicle	1 1154	Fahrzeug NN N			
external	1 1149	äußeres JJ	externes JJ	auswärtiges JJ	
internal	1 1149	inneres JJ	internes JJ	internistisches JJ	
of foot	2 1149	des Fußes	_Fuß_		

Scored interface vocabulary

SNOMED ID	Score	Fully Specified Name (Englisch)	Deutscher Interface-Term
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirnfarkt verursacht durch Stenose der A. carotis
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirnfarkt verursacht durch Stenose der A. karotis
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Schlaganfall wegen Stenose der Halsschlagader
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Insult wegen Stenose der Halsschlagader
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Schlaganfall wegen Karotisstenose
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Insult wegen Karotisstenose
99451000119105	0.800	Cerebral infarction due to stenosis of carotid artery (disorder)	Gehirnfarkt verursacht durch Verengung der Halsschlagader

German Interface terminology development

- Current state
 - Experimental use in Averbis Health Discovery
 - Experimental use by industry partners
- Future directions
 - More automation of
 - synonym / variant detection
 - quality control
 - Periodic updates

By machine learning using reference corpora (clinical, public)
 - Fuzzy term matching, matching out-of-language terms

2. Disambiguation of short forms

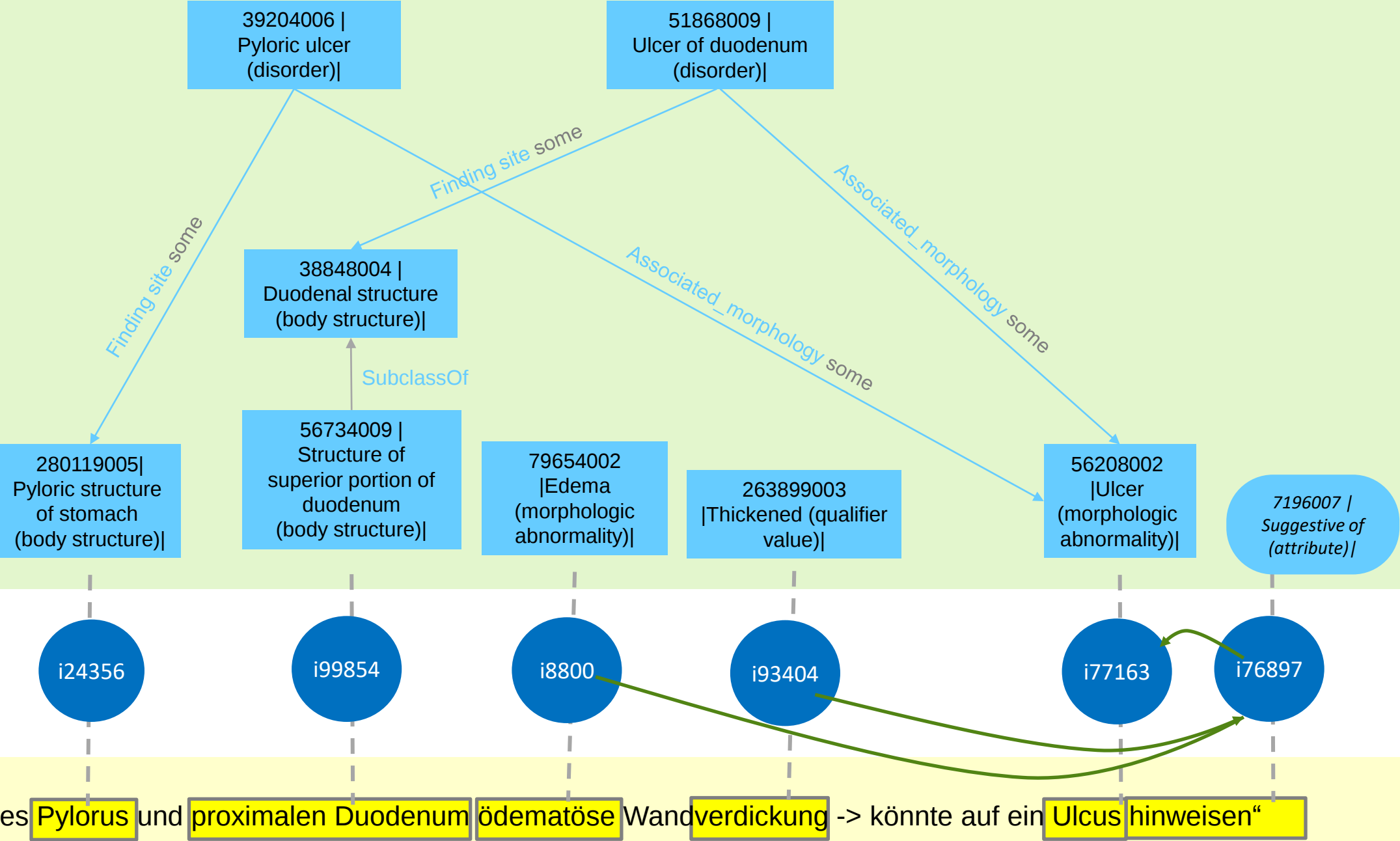
- Clinical texts are "infested" by short forms
 - Acronyms and abbreviations
 - Rarely introduced
 - Highly ambiguous
 - Not lexicalised
 - Confusion with non-abbreviations (roman numbers, capitalised text)
 - Institution-specific, specialty-specific
 - Often never expanded in clinical corpora
- Some attempts
 - Detection and disambiguation of short forms with "."
 - Disambiguation from clinical corpora (embeddings): <https://github.com/bst-mug/acres>
 - Harvesting acronym definitions + context from Web resources

“Z.n. TE eines exulc.
sek.knot.SSM li US dors.
5/11 Level IV 2,4mm
Tumordurchm.
Sentinell LK ing. li.
tumorfr.”

Identification of semantic relations

- Common problems in clinical text:
 - Temporal relations: "Arztbrief vom 5.11.2020"
"Max Mustermann *13.3.1979"
"Streptokokkenangina im Kindesalter"
"Niereninsuffizienz ED 4/2011"
"dialysepflichtig seit 3 Jahren"
 - Nominal anaphora: "akute Erosionen der Magenschleimhaut Die Schleimhautläsionen"
(correspond to taxonomic relations in the underlying ontology)
 - Bridging anaphora: "im Bereich proximalen Duodenum ... könnte auf ein Ulcus hinweisen"
(correspond to non-taxonomic relations in the underlying ontology, like location, part-of)
 - "Semantic similarity" relations ()
- Converting linear NLP output to graph
 - Using ontological structure of underlying ontology (SNOMED CT EL++ axioms)
(inferring "duodenal ulcer" out of "duodenum" and "ulcer")
 - Learning graph embeddings

SNOMED CT
classes and
relations



Semantically close primitive concepts

