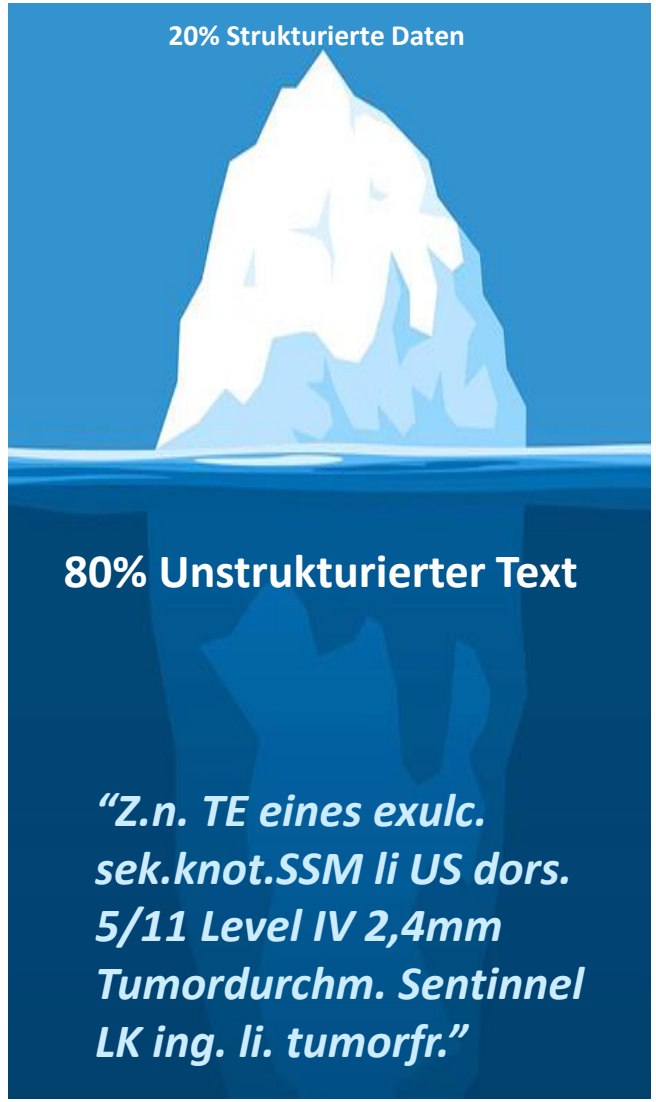
Entwicklung einer deutschsprachigen Interface-Terminologie für SNOMED CT

Stefan Schulz

Medizinische Universität Graz

Averbis GmbH, Freiburg

steschu@gmail.com



- Ein Großteil der relevanten EHR-Inhalte ist gering strukturierter Text.
- Medizinische Texte sind knapp formuliert, kontextbezogen und unterschiedlicher Qualität
- Viele Anwendungsfälle erfordern die Extraktion kodierter Inhalte aus klinischen Texten

Medizinische Universität Graz

averbis text analytics

Eignet sich SNOMED CT zur Informationsextraktion aus medizinischen Texten?

Häufigkeit von SNOMED Preferred Terms und ihrer Übersetzungen	Treffer Google*
– Englisch: "Secondary malignant neoplasm of liver"	100
– Schwedisch: "sekundär malign levertumör"	1
– Deutsch: "Sekundäre maligne Neoplasie der Leber"	1

Häufigkeit klinisch gebräuchlicher Synonyme	
– Englisch: "liver metastases"	1.230.000
– Schwedisch: "levermetastaser"	217.000
– Deutsch: "Lebermetastasen"	204.000

Ähnliche Beobachtungen in klinischen Korpora / PubMed:

Z.B. kein einziger Treffer für “Elektrokardiogramm” in 30.000 Kardiologie-Arztbriefen

*https://www.google.com/search?q=%22Secondary+malignant+neoplasm+of+liver%22

Faktoren, die die Extraktion von SNOMED-CT-Codes aus Kliniktexten erschweren

- Telegrammstil, Häufigkeit von **Kurzformen**, v.a. Akronymen *("ED 9/19, Fil. pulm., IDDM")*
- Dynamischer klinischer **Jargon** *("Biontech-Impfung", "Dexamethasongabe", "N. coli")*
- Ellipsen, Anaphern, **kontext**bezogene Wortbedeutungen *("nach Lyse", "die Mukosa")*
- Eine Übersetzung von SNOMED CT, die sich auf die Vorzugsterme beschränkt, wird der klinischen Sprache vielfach nicht gerecht
- Informationsextraktion mittels NLP (Natural Language Processing) erfordert eine Terminologie, die den Klinikjargon abbildet und mit SNOMED CT verknüpft ist
  1. durch Anreicherung einer **SNOMED-CT-Übersetzung** mit entsprechenden Synonymen (-> EN)
  2. durch nutzerseitige Erstellung und Pflege sogenannter **Interface-Terminologien**: dokumenten-nahe Termkollektionen, die mit SNOMED-CT-Codes (und ggf. postkoordinierten Ausdrücken) verknüpft werden → Empfehlung ASSESS-CT, 2016 *

* D Kalra, S Schulz, D Karlsson, R Vander Stichele, R Cornet, K Rosenbeck Gøeg, G Cangioli, C Chronaki, R Thiel, S Thun, V Stroetmann(2016). Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. ASSESS CT Recommendations. http://assess-ct.eu/final-brochure.html

Deutschsprachige Interface-Terminologie für SNOMED CT

- Seit 2014, mit Hilfe von 1-3 Medizinstudenten: Erstellung und Pflege eines Kernvokabulars aus englischen SNOMED CT-Beschreibungen
- Algorithmische Erzeugung von Varianten und Kombinationen, einschließlich Komposita
- Bewertung nach Vorkommen und Häufigkeit in Referenzkorpora und -terminologien, lexikalischen Patterns und Anti-Patterns
- Gefilterte Version für NLP (max. 6 Token): derzeit für ca. 270.000 SNOMED-Konzepte 2,4 Mio Terme
- Angepasst an das in den MI-I-Konsortien verwendete NLP-System Averbis Health Discovery, kann aber auch in anderen NLP-Pipelines verwendet werden

Hashemian Nik D, Kasáč Z, Goda Z, Semlitsch A, Schulz S. Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform. 2019 Aug 21;264:153-157

Kernvokabular

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 Gliedmaße|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_		

Automatisch generierte Interface-Terme

SNOMED ID	Score	Fully Specified Name (Englisch)	Deutscher Interface-Term
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirninfarkt verursacht durch Stenose der A. carotis
99451000119105	0.833	Cerebral infarction due to stenosis of carotid artery (disorder)	Hirninfarkt 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)	Gehirninfarkt verursacht durch Verengung der Halsschlagader

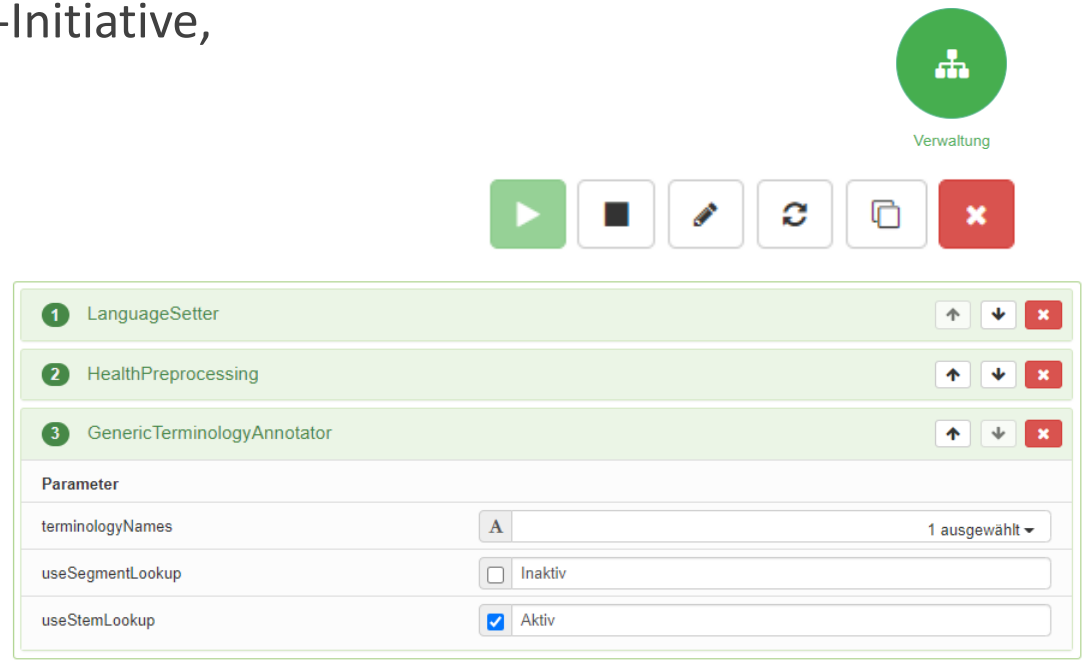
Einbindung der Interface-Terminologie in Averbis Health Discovery

- Zugang zu SNOMED CT – Interfaceterminologie

  - Gehostet von der Medizinischen Universität Graz
  - Wichtig: Keine Übersetzung sondern Map eigener Terminologie nach SNOMED CT
  - Nutzungsbedingungen: frei für Medizininformatik-Initiative, ansonsten in Absprache mit Med. Univ. Graz

- Anwendung

  - Hochladen des OBO-Files in Terminologie-Verwaltung
  - *Discharge-Pipeline* "klonen"
  - *Generic Terminology Annotator* mit *stemLookup*:
  - Terminologieexport (aus Terminologieverwaltung)
  - Annotator starten (auch über API)



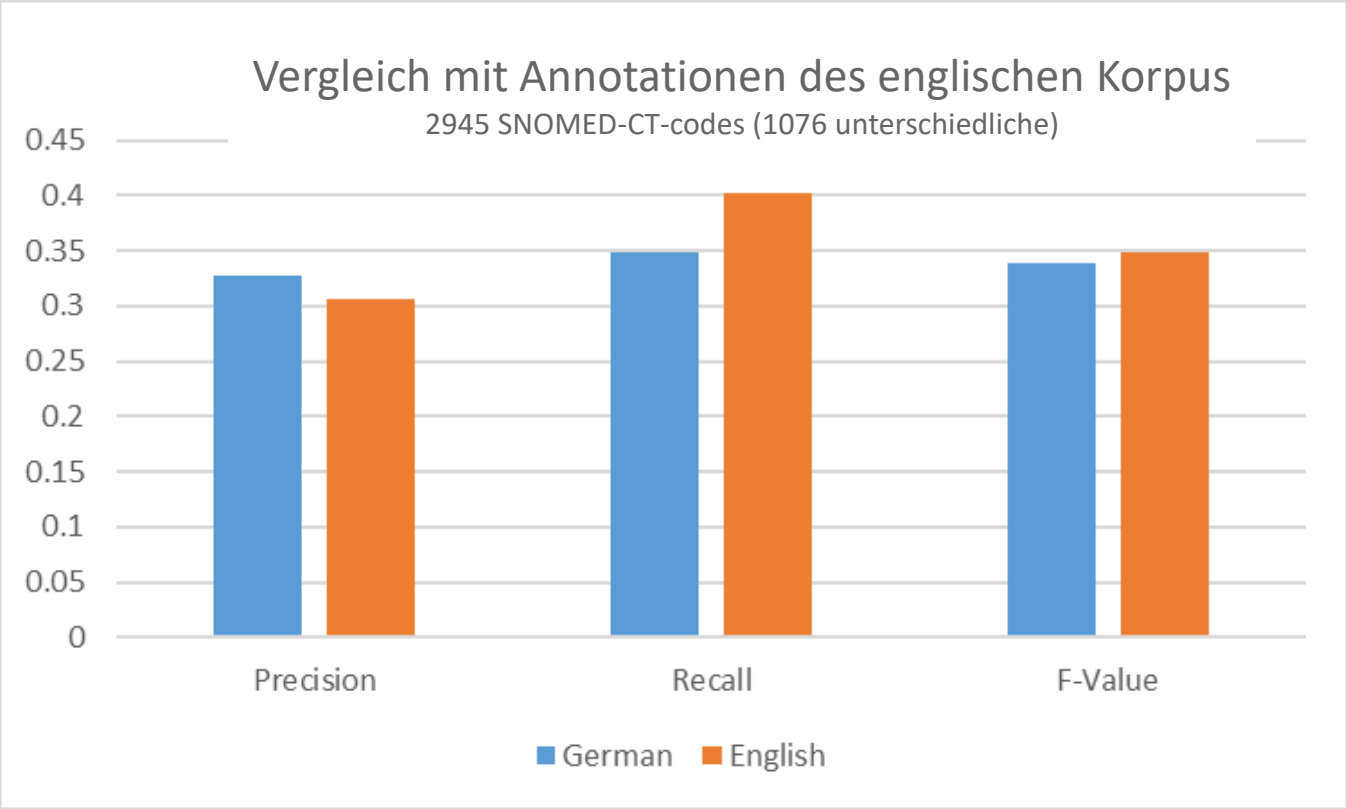
http://user.medunigraz.at/stefan.schulz/mugit/ (siehe ReadMe-File)

Validierung mit Parallelkorpus

- Terminologien:
  Englisch: SNOMED CT-Version März 2020: 1,2 Mio. aktive "Descriptions"
  Deutsch: NLP-Auszug der deutschen Interface-Terminologie: 1,8 Mio. Einträge
- Benchmark: ASSESS-CT Parallelkorpus
  - Ausschnitte aus klinischen Dokumenten, klinischen Fachgebieten und Ausgangssprachen, durchschnittlich 3650 Wörter pro Sprache
  - Englische, niederländische, schwedische und französische Version, annotiert von Terminologieexperten mit SNOMED CT (2015)
- Referenzstandard: SNOMED-CT-Annotationen der englischen Version des Parallelkorpus
- NLP-System: Averbis Health Discovery für Deutsch und Englisch (www.averbis.com)

Miñarro-Giménez, J.A., et al. (2018). Qualitative analysis of manual annotations of clinical text with SNOMED CT.
PLoS One. Dec 27:3(12)

Ergebnisse



- Unterschiede nicht signifikant zwischen
  - maschineller Annotation mittels der englischen SNOMED-Descriptions auf englischem Korpus
  - maschineller Annotation mittels der deutscher Interface-terminologie auf bedeutungsgleichem deutschem Korpus
- Inter-Annotator-Agreement der manuellen Annotationen war nur 0,4 (Krippendorffs Alpha), trotz Annotationsrichtlinien*

*Miñarro-Giménez JA, Martínez-Costa C, Karlsson D, Schulz S, Gøeg KR. Qualitative analysis of manual annotations of clinical text with SNOMED CT. PLoS One. 2018 Dec 27;13(12):e0209547.

Diskussion

- Schlechte Übereinstimmung bei der Erstellung von Referenzannotationen: Bekanntes Problem des Terminologiemappings (nicht spezifisch für SNOMED CT, vgl. ASSESS CT-Bericht*)
  - Feinkörnige konzeptuelle Unterscheidungen in großen Terminologien ("anxiety"-"fear"-"phobia")
  - Mehrdeutige Terme, insbesondere Akronyme und elliptische Ausdrücke ("Fundus", "Corpus")
- Prä-Koordination vs. Post-Koordination
  - Text: „Der laterale Epikondylus des linken Ellenbogens war gebrochen“
  - Mensch: 208271008 |Closed fracture distal humerus, lateral epicondyle + 7771000 |Left
  - Maschine: 72704001 |Fracture + 73451009 |Structure of lateral epicondyle of humerus + 7771000 |Left|
- Wie lässt sich das verbessern?
  - Ausnutzung definierender Axiome von SNOMED-CT-Konzepten (Beschreibungslogik)
  - Neuronales Maschinelles Lernen: Lernen von Ähnlichkeiten via Graph Embeddings, Präprozessieren des Input-Texts durch kontextsensitive Expansion von Kurzformen, Disambiguierung, Terminologielernen

* D Kalra, S Schulz, D Karlsson, R Vander Stichele, R Cornet, K Rosenbeck Gøeg, G Cangioli, C Chronaki, R Thiel, S Thun, V Stroetmann(2016). Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. ASSESS CT Recommendations. http://assess-ct.eu/final-brochure.html

Fazit

- Die deutsche Interface-Terminologie zeigt bei deutschen Texten gleiche Performance wie die englischen SNOMED CT-Descriptions bei parallelem englischem Text.
  Das ist bemerkenswert aufgrund des Fehlens einer amtlichen deutschen SNOMED-CT-Übersetzung und des ressourcenarmen Ansatzes zur Erstellung der Interface-Terminologie
- Konstrast: Schwedische SNOMED-CT-Übersetzung: > 8 Mio. €, aber viel niedrigere Term-Matching-Rate im Vergleich zu Englisch auf demselben Korpus (vgl. ASSESS-CT), da nur ein Term pro Konzept
- Für NLP scheint die Interface-Terminologie sinnvoll, auch parallel und evtl. zur Unterstützung einer amtlichen Übersetzung
- Bis zu wirklich zufriedenstellenden Text Mining-Ergebnissen aus realen klinischen Texte ist es noch ein langer Weg. Dennoch: dank der verfügbaren Terminologie und einfach zu nutzenden Tools, wie die in den MI-I-Konsortien verfügbare Averbis-Pipeline kann mit geringem Aufwand getestet werden.

BEISPIEL PRÄ-POSTKOORDINATION

Text 1	Asserted SNOMED concepts	Implied SNOMED CT concepts
Im rechten	Right (qualifier value)	is-a → Right (qualifier value)
Großzehennagel	Structure of nail unit of great toe (body structure)	is-a → Great toe structure (body structure); is-a → Structure of nail unit of toe (body structure) is-a → Nail unit structure (body structure)
fanden sich		
Candida-Spezies	Genus Candida (organism)	is-a → Genus Candida (organism)
als Ursache der		
Infektion	Infectious process (qualifier value)	is-a → Infectious process (qualifier value)
Text 2	**Asserted SNOMED concepts**	**Implied SNOMED CT concepts**
Candida-Onychomykose	Candidiasis of nails (disorder)	∃ causative agent → Genus Candida (organism); ∃ pathological process → Infectious process (qualifier value); ∃ finding site → Nail unit structure (body structure)
Großzehe rechts	Structure of right great toe (body structure)	is-a → Great toe structure (body structure); ∃ laterality → Right (qualifier value)

NLP – ARCHITEKTUR AVERBIS