



Medical University of Graz

PREMEDICAL: OPTIMIERUNG DER LEXIKONABDECKUNG FÜR TEXT MINING

Stefan Schulz

PREMEDICAL - Kick-Off-Meeting

9. Mai 2022

Zwei Problemstellungen

1. SNOMED CT

- ▶ Erste Wahl für eine umfassende, feingranuläre Repräsentation klinischer Inhalte, insbesondere in Kombination mit LOINC und HL7-FHIR als Informationsmodell
- ▶ SNOMED CT in Österreich verfügbar, keine offizielle deutschsprachige Übersetzung

2. Textqualität der elektronischen Krankenakte

- ▶ zahlreiche nicht lexikalisierte / nicht lexikalisierbare Ausdrücke
 - ▶ Abkürzungen: "chron perfor. Gabla", "V. mors fel dig IV man dex"
 - ▶ Akronyme, meist ambig, oft nur lokal verständlich : "MNZN", "MM", "BHB", "EBA", "RA"
 - ▶ Schreib- und Tippfehler, Schreibvarianten: "Polenallergie", "Magenulcus", "Schüsselbein"
 - ▶ Komposita: Außenmeniscusscheibenmeniscusdeformität, Ibuprofenintoxikation

1. SCT-GIT - SNOMED CT - German Interface Terminology

- NLP erfordert Termsammlungen, die zuvorderst den *Sprachgebrauch*, nicht die *Sprachnorm* abbilden

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

- Problem: Tendenz von SNOMED CT in Richtung Sprachnorm
Beispiel "Opioid" vs. "Substance with opioid receptor agonist mechanism of action"

1. SCT-GIT - SNOMED CT - German Interface Terminology

- ▶ Seit 2014, mit Hilfe von 1-3 Studierenden der Medizin: Übersetzung eines Kernvokabulars aus englischem SNOMED CT
- ▶ Algorithmische Erzeugung von Varianten, Kombinationen, Komposita
- ▶ Priorisierung nach Korpushäufigkeit
- ▶ Gefilterte Version für NLP (max. 6 Token): derzeit für ca. 270.000 SNOMED-Konzepte 2,4 Mio Terme
- ▶ Angepasst an das Informationsextraktionssystem Averbis Health Discovery

► Kernvokabular

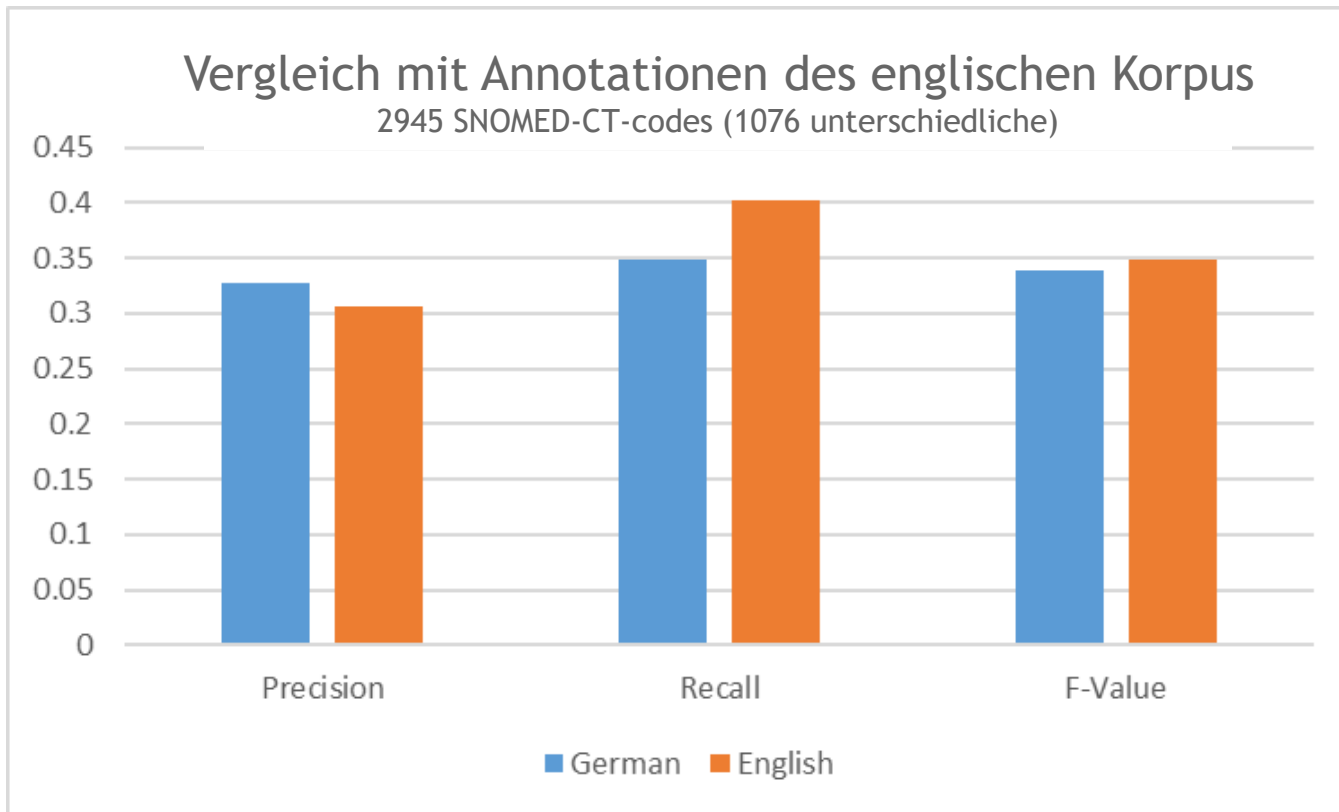
English	L Count	German 1	German 2	German 3	German 4
burn	1	1264 Brandverletzung NN F	Brandwunde NN F	Verbrennung NN F	
normal	1	1264 normales JJ	normenhaftes JJ		
ankle	1	1254 Knöchel NN M			
wrist	1	1251 Handgelenk NN N			
drug	1	1244 Wirkstoff NN M	Arznei NN F	Arzneimittel NN N	Droge NN F
second	1	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 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_		

► Automatisch generierte Interface-Terme

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

1. SCT-GIT - SNOMED CT - German Interface Terminology

► Validierung gegen klin. Parallelkorpus mit Expertenannotationen



- Unterschiede nicht signifikant zwischen
 - NLP-Tagging auf Basis des englischen SNOMED auf englischem Korpus
 - NLP-Tagging auf Basis unserer Interface-terminologie auf bedeutungsgleichem deutschem Korpus
- Inter-Annotator-Agreement der manuellen Annotationen 0,4 (Krippendorffs Alpha): bekanntes Problem großer Terminologien mit ähnlichen Konzepten und Möglichkeiten von Prä- und Postkoordination

2. Ressourcen und Algorithmen für Text Cleansing

- ▶ Ersatz nicht-lexikalisierte Formen durch lexikalisierte (und somit auf Terminologien abbildbare)
- ▶ Disambiguierung von Kurzformen:
Beispiel: "Metastasen eines MM" vs. "Hörverlust bei MM"
- ▶ Zwei Ansätze:
 - ▶ Nicht-überwachtes Lernen aus Klinikcorpus (Word embeddings)
 - ▶ PhD Michel Oleyunik: WORD2VEC F1 = 67%
 - Probleme: Größe des Trainingskorpus, Filterkriterien
 - ▶ Web-Mining (BING API Query): aus Akronym mit Kontext wird Korpus aus Suchmaschinen-Ergebnisseiten erzeugt: Ranking von Auflösungskandidaten mittels verschiedener Heuristiken
 - ▶ derzeit in Entwicklung

Beispiel: Akronymauflösung durch Web Mining

```
+++++
Acronym = "MM" -- Query = "Hörverlust bei MM"
with 200 hits, displaying score values >= 0.3
Upper case rate = 0.09
Acronyms per 1000 characters = 0.22
```

```
+++++
Word N-Gram Statistics from acronym definition patterns - 0 matches in 154 characters
```

```
+++++
COMB FREQ SCORE CANDIDATE CAPI COMP EQUI INIT STOP STRU TRAN
+++++
```

```
Word N-Gram Statistics from entire Bing dump - 35 matches in 62657 characters
```

```
+++++
COMB FREQ SCORE CANDIDATE CAPI COMP EQUI INIT STOP STRU TRAN
1.15 0014 1.00 Morbus Menière 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.32 0002 1.00 Morbus Menière - 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.32 0002 1.00 Manuelle Medizin 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.26 0002 0.80 Maßnahmen 1.00 1.00 0.80 1.00 1.00 1.00 1.00
0.16 0002 0.50 mHV Mittlerer 1.00 1.00 1.00 1.00 1.00 0.50 1.00
0.16 0002 0.50 MSD Manuals 1.00 1.00 1.00 1.00 1.00 0.50 1.00
0.16 0002 0.50 MSD Manual 1.00 1.00 1.00 1.00 1.00 0.50 1.00
0.16 0002 0.50 - MSD Manual 1.00 1.00 1.00 1.00 1.00 0.50 1.00
```

```
+++++
Acronym = "MM" -- Query = "Metastasen eines MM"
with 200 hits, displaying score values >= 0.3
Upper case rate = 0.11
Acronyms per 1000 characters = 0.35
```

```
+++++
Word N-Gram Statistics from acronym definition patterns - 0 matches in 154 characters
```

```
+++++
COMB FREQ SCORE CANDIDATE CAPI COMP EQUI INIT STOP STRU TRAN
+++++
```

```
Word N-Gram Statistics from entire Bing dump - 269 matches in 68251 characters
```

```
+++++
COMB FREQ SCORE CANDIDATE CAPI COMP EQUI INIT STOP STRU TRAN
1.30 0020 1.00 Malignes Melanom 1.00 1.00 1.00 1.00 1.00 1.00 1.00
1.10 0017 0.89 Mikrometastasen 1.00 1.00 0.89 1.00 1.00 1.00 1.00
0.93 0080 0.49 Melanom 1.00 1.00 0.49 1.00 1.00 1.00 1.00
0.87 0012 0.80 malignen Melanom 0.80 1.00 1.00 1.00 1.00 1.00 1.00
0.75 0010 0.75 Melanoms 1.00 1.00 0.75 1.00 1.00 1.00 1.00
0.71 0005 1.00 Muskuloskelettale Metastasen 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.63 0006 0.80 maligne Melanom 0.80 1.00 1.00 1.00 1.00 1.00 1.00
```