Using Natural Language Processing (NLP) for Annotating German Clinical Narratives with SNOMED CT

Stefan Schulz, Markus Kreuzthaler, David Hashemian Nik, Larissa Hammer, Michaela Schneider

Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz, Austria

Introduction

Rationale for clinical text mining

SNOMED CT EXPO 2020
Virtual Conference October 8-9

- In most clinical information systems, existing structured data are
  - Incomplete, often only encompassing codes for major procedures and diagnoses
  - Biased, due to the purposes for which they were acquired (e.g.)
  - Error-prone, due to clinicians' lack of interest in structured documentation paralleling free-text narratives
- Factors complicating mining information from clinical narratives
  - Compact, telegram-style language, high frequency of short forms
  - Dynamic clinical jargon, not well represented by clinical terminologies
  - Highly contextualized
- How to bridge this gap?
  - Using SNOMED CT (supported by information models)
  - Using comprehensive language resources linked to SNOMED CT

Web frequencies of Fully Specified Names (FSNs)

SNOMED CT EXPO 2020
Virtual Conference October 8-9

- Frequency of FSNs and their translations
  - English: "Secondary malignant neoplasm of liver" 1,500 hits
  - Swedish: "sekundär malign levertumör" 0 hits
  - German: "Sekundäre maligne Neoplasie der Leber“ 0 hits
- Frequency of popular synonyms
  - English: "Hepatic metastasis“ 46,600 hits
  - Swedish: "levermetastasen“ 1,470 hits
  - German: "Lebermetastasen“ 13,600 hits
- Similar findings in clinical corpora
  - e.g. no single occurrence of “Elektrokardiogramm” in 30k cardiology notes

Term matching with localised SNOMED CT versions

- EU coordination & support action ASSESS CT, 2016 recommendations:
  - Term matching with localised SNOMED CT versions insufficient
  - Fully Specified Names / Preferred Terms: poor coverage of clinical jargon
  - Compared to International English SNOMED CT, which contains synonyms
  - Recommends user interface terminologies linked to SNOMED CT
- Characteristics of (user) interface terminologies
  - Capture the language actually used by clinicians / laypersons
  - Bottom-up instead of top-down approach
  - Use cases: term retrieval, value set creation, NLP

Kalra, D., et al. (2016). Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. ASSESS CT Recommendations. http://assess-ct.eu/final-brochure.html

Resource

German-language interface terminology

- Low-resource activity initiated 2014 (senior terminologist, 3 medical students)
- Automated generation of German terms out of a core vocabulary with human-curated machine translations (single-word and short-term) extracted from English SNOMED CT descriptions. Enrichment by synonyms as a key activity, using n-gram hit lists extracted from clinical corpora
- Natural-language generator produces variants and combinations, including single-word compounds. No translation of FSN, no term preferences
- Scoring according to occurrence and frequency in reference corpora and term collections, lexical patterns and anti-patterns
- Filtered version for NLP (max 6 tokens, minimization of ambiguities)

Hashemian Nik, D., et al. (2019). Building an Experimental German User Interface Terminology Linked to SNOMED CT. Stud Health Technol Inform, 264:153-157

Core terminology

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_		

Scored interface terminology

SNOMED ID	Score	English FSN	German 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

Experiment

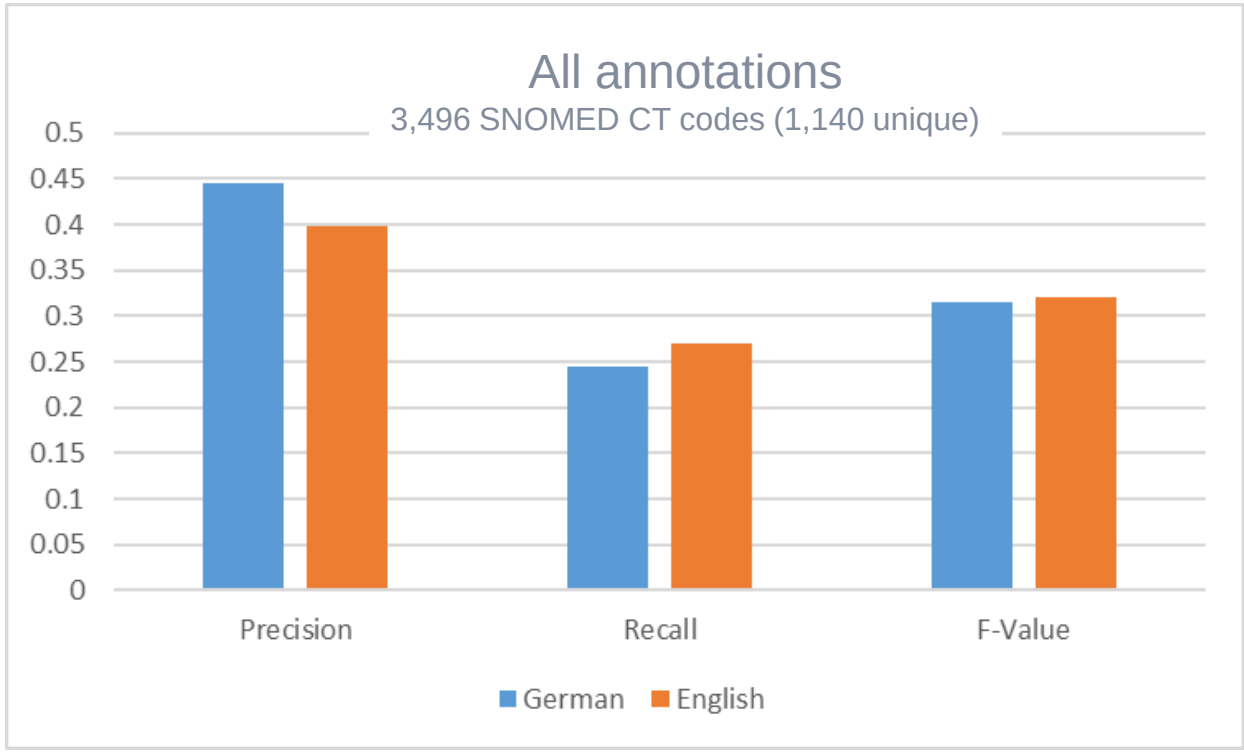
Comparison English - German

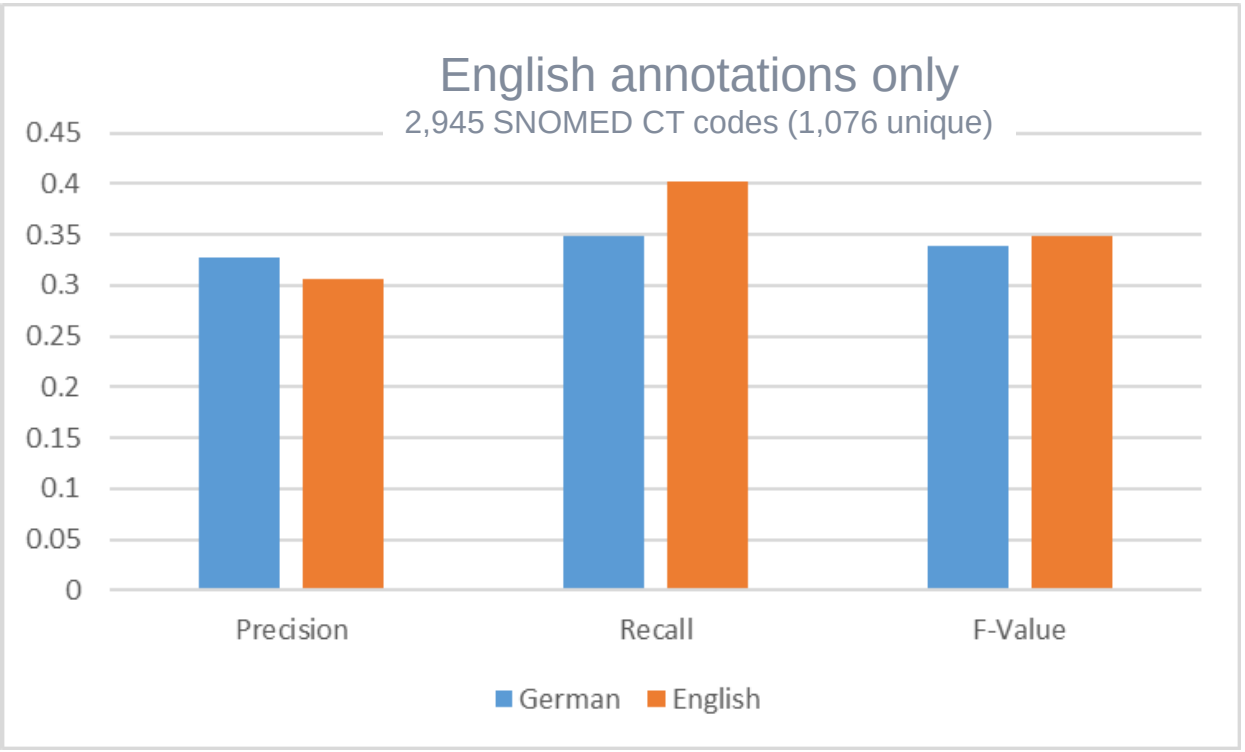
- Terminologies
  - Complete March 2020 International Version Description table: 1.2 M active entries
  - NLP extract of German Interface Terminology: 1.8 M entries
- ASSESS-CT parallel corpus
  - Snippets of clinical documents, different clinical specialties and source languages
  - On average 3650 words per language
  - English, Dutch, Swedish and French versions annotated by terminology experts with SNOMED CT (2015)
  - Reference standards: pooled (all annotations), English annotations only
- NLP system
  - Averbis Health Discovery for German and English (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)

Results

Term detection English - German





- Differences not significant
- Reported inter-annotator agreement 0.4 (Krippendorff's Alpha)
- Pre-coordinated concepts privileged by annotation guidelines

Discussion & Conclusions

F-values not satisfactory

- Known issues with terminology grounding of clinical texts
  - Not specific to SNOMED CT (cf. ASSESS CT report)
  - Fine-grained conceptual distinctions in large terminologies
  - Ambiguous terms, particularly acronyms and elliptic expressions (“fundus”)
- Pre coordination vs. post-coordination
  - Text: “The lateral epicondyle of the left elbow was broken”
  - Human coders: 208271008 |Closed fracture distal humerus, lateral epicondyle
  - Machine: 72704001 |Fracture + 73451009 |Structure of lateral epicondyle of humerus + 7771000 |Left (qualifier value)|
- How to improve?
  - Symbolic reasoning: exploiting defining axioms of SNOMED CT concepts
  - Neural ML: exploiting phrase-level similarities; short form expansion + disambiguation

Encouraging for interface terminology approach

- German interface terminology behaves as well on German texts as English SNOMED CT descriptions on English text
  - Remarkable due to absence of German SNOMED CT translation and low-resource terminology-building approach
  - Puts the benefits of the “traditional” terminology translation process into perspective
  - Example Swedish SNOMED CT translation: > 8 M €, but much lower term matching rate compared to English on same corpus (cf. ASSESS-CT)
- Conclusion
  - At least for NLP: Interface terminology construction more cost-effective
  - Independent of language: still a long way to go to really satisfactory text mining results of real-world clinical texts

Miñarro-Giménez, J.A., et al. (2019) Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform.:123:37-48

Thank you for your attention

Stefan Schulz | Markus Kreuzthaler | David Hashemian-Nik | Larissa Hammer | Michaela Schneider

Contact: stefan.schulz@medunigraz.at