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

• 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)

- 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

Using NLP for Annotating German Clinical Narratives with SNOMED CT

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).

<table>
<thead>
<tr>
<th>English</th>
<th>L Count</th>
<th>German 1</th>
<th>German 2</th>
<th>German 3</th>
<th>German 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>burn</td>
<td>1</td>
<td>1264 Brandverletzung</td>
<td>NN</td>
<td>F</td>
<td>Brandwunde</td>
</tr>
<tr>
<td>normal</td>
<td>1</td>
<td>1264 normales</td>
<td>JJ</td>
<td>normenhaftes</td>
<td>JJ</td>
</tr>
<tr>
<td>ankle</td>
<td>1</td>
<td>1254 Knöchel</td>
<td>NN</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>wrist</td>
<td>1</td>
<td>1251 Handgelenk</td>
<td>NN</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>drug</td>
<td>1</td>
<td>1244 Wirkstoff</td>
<td>NN</td>
<td>M</td>
<td>Arznei</td>
</tr>
<tr>
<td>second</td>
<td>1</td>
<td>1244 zweites</td>
<td>JJ</td>
<td>Sekunde</td>
<td>NN</td>
</tr>
<tr>
<td>uncertain</td>
<td>1</td>
<td>1227 unsicheres</td>
<td>JJ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abdominal</td>
<td>1</td>
<td>1222 abdominales</td>
<td>JJ</td>
<td>Bauch-</td>
<td>abdominelles</td>
</tr>
<tr>
<td>membrane</td>
<td>1</td>
<td>1210 Membran</td>
<td>NN</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>1</td>
<td>1207 Hepar</td>
<td>NL</td>
<td>N</td>
<td>Leber</td>
</tr>
<tr>
<td>microgram</td>
<td>1</td>
<td>1202 µg</td>
<td>µg %VOID%</td>
<td>Mikrogramm</td>
<td>NN</td>
</tr>
<tr>
<td>middle</td>
<td>1</td>
<td>1193 mittleres</td>
<td>JJ</td>
<td>Mitte</td>
<td>NN</td>
</tr>
<tr>
<td>ulcer</td>
<td>1</td>
<td>1180 Ulzeration</td>
<td>NN</td>
<td>F</td>
<td>Ulkus</td>
</tr>
<tr>
<td>upper limb</td>
<td>2</td>
<td>1180 oberes</td>
<td>JJ Extremität</td>
<td>NN</td>
<td>F</td>
</tr>
<tr>
<td>fluoroscopic</td>
<td>1</td>
<td>1171 Durchleuchtungs-</td>
<td></td>
<td>durchleuchtungsgestütztes</td>
<td>JJ</td>
</tr>
<tr>
<td>effect</td>
<td>1</td>
<td>1170 Effekt</td>
<td>NN</td>
<td>M</td>
<td>Auswirkung</td>
</tr>
<tr>
<td>service</td>
<td>1</td>
<td>1158 Service</td>
<td>NN</td>
<td>M</td>
<td>Dienst</td>
</tr>
<tr>
<td>vehicle</td>
<td>1</td>
<td>1154 Fahrzeug</td>
<td>NN</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>external</td>
<td>1</td>
<td>1149 äußeres</td>
<td>JJ</td>
<td>externes</td>
<td>JJ</td>
</tr>
<tr>
<td>internal</td>
<td>1</td>
<td>1149 inneres</td>
<td>JJ</td>
<td>internes</td>
<td>JJ</td>
</tr>
<tr>
<td>of foot</td>
<td>2</td>
<td>1149 des Fußes</td>
<td></td>
<td><em>Fuß</em></td>
<td></td>
</tr>
<tr>
<td>SNOMED ID</td>
<td>Score</td>
<td>English FSN</td>
<td>German Interface Term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>----------------------------------------------------</td>
<td>------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Hirninfarkt verursacht durch Stenose der A. carotis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Hirninfarkt verursacht durch Stenose der A. karotis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Schlaganfall wegen Stenose der Halsschlagader</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Insult wegen Stenose der Halsschlagader</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Schlaganfall wegen Karotisstenose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.833</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Insult wegen Karotisstenose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99451000119105</td>
<td>0.800</td>
<td>Cerebral infarction due to stenosis of carotid artery (disorder)</td>
<td>Gehirninfarkt verursacht durch Verengung der Halsschlagader</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)

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

Thank you for your attention

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

Contact: stefan.schulz@medunigraz.at