Interface-Terminologien und Referenzterminologien

Stefan Schulz
Medizinische Universität Graz
Josef Ingenerf
Universität zu Lübeck

Beispiel: Terme in Kardiologie-Arztbriefen

- Fragestellungen:
  - Was sind Anforderungen an Medizinterminologien zur Unterstützung semantischer Annotation deutsche Medizintexte?
  - Allgemein: Wie lässt sich das Defizit deutschsprachiger terminologischer Ressourcen beheben?
  - Spezifisch: Wie macht man SNOMED CT fit für die Benutzung in deutschsprachigen Ländern

Beispiel: Terme in Kardiologie-Arztbriefen

Vorzugsterm (ICD, OPS)	Anzahl	Synonym	Anzahl
Aortenklappenstenose	3749	Aortenstenose	3126
Hirninfarkt	7	Schlaganfall	65
Elektrokardiogramm	0	EKG	12208
Koronare Herzerkrankung	331	KHK	18455
Nicht-ST-Hebungsinfarkt	498	NSTEMI	3839
Magnetresonanztomographie	2	NMR	17

Zwei Aspekte von Terminologien

- Normativ
  - Codes + Labels für definierte (Klassen von) Gegenständen
  - "sprechende“ Labels, z.B. "*Primary malignant neoplasm of lung(disorder)*".
  - Erklärende oder definierende Texte (*scope notes*)
  - Formale Beschreibungen: → formale Ontologie
- Deskriptiv
  - Tatsächlicher Sprachgebrauch: "*Lungenkrebs*", "*Bronchial-Ca*"
  - Erweiterung zu einem Thesaurus durch semantische Relationen (Synonymie, Hypernymie,…)
- Gängige Terminologiesysteme adressieren beiden Aspekte in unterschiedlichen Maß und unsystematisch

Unterscheidung Referenzterminologie - Interfaceterminologie

- H2020 Projekt ASSESS-CT

Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

  - **Referenzterminologien**:
    - Sprachunabhängige Konzepte / Codes: Eigenschaften der Objekte, die von diesen denotiert werden
    - "Sprechende"Labels in der jeweiligen Sprache, unterstützt durch textliche und / oder formale (ontologische) Definitionen
  - **Interfaceterminologien**:
    - Sammlungen von sprachlichen Ausdrücken, die in schriftlicher und mündlicher Kommunikation verwendet werden.
    - Verknüpfung zu Referenzterminologien
    - Problem: hohe Ambiguität, insbesondere von Abkürzungen (Akronymen)

Bedeutung für manuelle Annotation und klinisches Text Mining

- Interface-Terme kommen vor als:
  - Synonyme oder Entry Terms in Referenzterminologien (MeSH)
  - Separate Interface-Terminologie, gemappt auf Referenzterminologie (ICD-10 Diagnosenthesaurus)
- Terminologie-Lokalisierung via Label-Übersetzung
  - hoher Aufwand (Erfahrungen Dänemark + Schweden)
  - schlechte Benutzerakzeptanz*
  - geringer Recall beim Text Mining
- Statt dessen: inkrementelle Akquisition von Interface-Termen und Verlinkung mit Referenzterminologie
- Hypothese: Bottom-up / Crowdsourcing

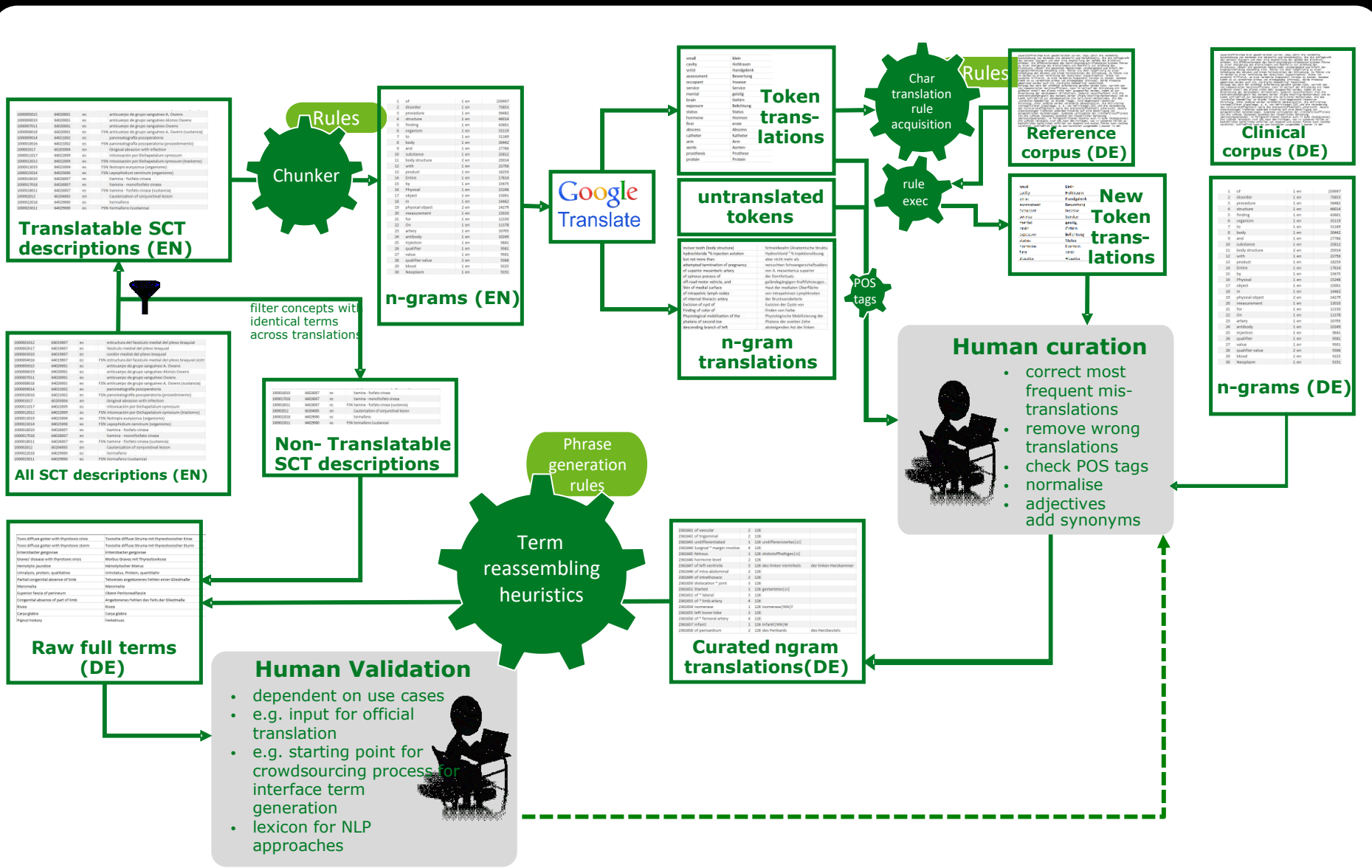
*Højen AR et al. SNOMED CT adoption in Denmark--why is it so hard? Stud Health Technol Inform. 2014;205:226-230

MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (I)

- MUG-GIT (*Medical University of Graz – German Interface Terminology*) zur maschinellen Annotation von deutschsprachigen Kliniktexten → Extrakte für semantisches Data Warehouse in Cbmed-IICCAB*
- Modularisierung (Zerlegung in N-Gramme, meist NPs und PPs), Editieren eines abgeleiteten Kernvokabular, motiviert durch hoch repetitive Teilphrasen, z.B.
  - "Magnetic resonance imaging" in 627 SNOMED -Termen
  - "second degree burn" in 166 SNOMED-Termen
- Maschinelle Vorübersetzung
- Priorisierung nach Häufigkeit:
  - Manuelle Revision und Selektion der NP-Liste
  - Anreicherung durch Terme aus anderen deutschen Terminologien und klinischen Corpora
- Regelbasierte Rekonstruktion der Komplettübersetzung

*Schulz S. Innovative Nutzung von Informationen für klinische Versorgung und Biomarkerforschung. http://goo.gl/wHMedz

MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (II)



MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (III)

- Kernvokabular:
  - gepflegt durch zwei Medizinstudentinnen unter Aufsicht
  - Priorisiert nach Use Cases
- Richtlinien
  - keine Schreibvarianten (c/k/z - Problem)
  - Akronyme nur im Kontext (kein Eintrag für "CT", aber für "Schädel-CT")
  - Vermeidung von Ambiguitäten, z.B. statt Eintrag "Aufnahme": "bei Aufnahme" (on admission), "Aufnahme von" (intake of)
- Derzeitiger Stand:
  - ca. 2 Millionen Interface-Terme
  - Automatisch generiert aus einem Kernvokabular mit deutschen 92,500 N-Grammen, verknüpft mit 85,400 englischen N-Grammen
  - Benchmark: MEDLINE extrahierter Parallelkorpus:
    Aktuelle Term-Abdeckung 33,1% für Deutsch gegenüber 55,4% (EN)

ngram - Kernvokabular

vaginal	1	1478	vaginales|JJ	Scheiden-	
fluoroscopic guidance	2	1477	Durchleuchtungskontrolle|NN|F		
disc	1	1476	Scheibe|NN|F		
lower limb	2	1473	unteres|JJ Extremität|NN|F	Bein|NN|N	
brain	1	1468	Gehirn|NN|N	Hirn|NN|N	Encephalon|NN|N
preparation	1	1464	Zubereitung|NN|F	Aufbereitung|NN|F	Präparation|NN|F
method	1	1463	Verfahren|NN|N	Methode|NN|F	
of bone	2	1462	des Knochens	_Knochen_	
Red	1	1455	rotes|JJ		
Monitoring	1	1453	Überwachung|NN|F	Monitoring|NN|N	
Computed	1	1453	berechnetes|JJ	Computer-	
phalanx	1	1449	Phalanx|NN|F		
subsp.	1	1449			
anastomosis	1	1447	Anastomose|NN|F	Anastomosierung|NN|F	
vessel	1	1446	Blutgefäß|NN|N	Gefäß|NN|N	
Computed tomography	2	1443	Computertomographie|NN|F		
uterus	1	1436	Uterus|NN|M	Gebärmutter|NN|F	
difficulty	1	1432	Schwierigkeit|NN|F		
elbow	1	1429	Ellbogen|NN|M	Cubitus|NN|M	Ellbogengelenk|NN|N
high	1	1429	hohes|JJ		
food	1	1423	Lebensmittel|NN|N	Speise|NN|F	Nahrungsmittel|NN|N
Observation	1	1423	Beobachtung|NN|F		
using fluoroscopic	2	1422			
unable	1	1421	unfähiges|JJ		
Peripheral	1	1419	peripheres|JJ		
unable to	2	1418	unfähig zu		
Vascular	1	1417	vaskuläres|JJ	Gefäß-	
using fluoroscopic guidance	3	1416	mit Durchleuchtungskontrolle		
Benign neoplasm	2	1415	gutartiges|JJ Neubildung|NN|F	gutartiges|JJ Neoplasie|NN|F	benignes|JJ Neoplasie|NN|F

Automatische generierte Interfaceterminologie

20170315_240011_002	126952004	Neoplasm of brain	Gehirnneubildung
20170315_240011_003	126952004	Neoplasm of brain	Neubildung des Hirns
20170315_240011_004	126952004	Neoplasm of brain	Hirnneubildung
20170315_240011_005	126952004	Neoplasm of brain	Neoplasie des Gehirns
20170315_240011_006	126952004	Neoplasm of brain	Gehirnneoplasie
20170315_240011_007	126952004	Neoplasm of brain	Neoplasie des Hirns
20170315_240011_008	126952004	Neoplasm of brain	Hirnneoplasie
20170315_240011_009	126952004	Neoplasm of brain	Neoplasma des Gehirns
20170315_240011_010	126952004	Neoplasm of brain	Gehirnneoplasma
20170315_240011_011	126952004	Neoplasm of brain	Neoplasma des Hirns
20170315_240011_012	126952004	Neoplasm of brain	Hirnneoplasma
20170315_241010_001	126953009	Neoplasm of cerebrum	Neubildung des Großhirns
20170315_241010_002	126953009	Neoplasm of cerebrum	Neoplasie des Großhirns
20170315_241010_003	126953009	Neoplasm of cerebrum	Neoplasma des Großhirns
20170315_242015_001	126954003	Neoplasm of frontal lobe	Neubildung des Frontallappens
20170315_242015_002	126954003	Neoplasm of frontal lobe	Neubildung des Lobus frontalis
20170315_242015_003	126954003	Neoplasm of frontal lobe	Neoplasie des Frontallappens
20170315_242015_004	126954003	Neoplasm of frontal lobe	Neoplasie des Lobus frontalis
20170315_242015_005	126954003	Neoplasm of frontal lobe	Neoplasma des Frontallappens
20170315_242015_006	126954003	Neoplasm of frontal lobe	Neoplasma des Lobus frontalis
20170315_243013_001	126955002	Neoplasm of temporal lobe	Neubildung des Temporallappens
20170315_243013_002	126955002	Neoplasm of temporal lobe	Neubildung des Lobus temporalis
20170315_243013_003	126955002	Neoplasm of temporal lobe	Neoplasie des Temporallappens
20170315_243013_004	126955002	Neoplasm of temporal lobe	Neoplasie des Lobus temporalis
20170315_243013_005	126955002	Neoplasm of temporal lobe	Neoplasma des Temporallappens

Ko-operative Entwicklung einer deutschen Interface-Terminologie

Ko-operative Entwicklung einer deutschen Interface-Terminologie

- Günstige Rahmenbedingungen
  - Datenintegration / Sekundärnutzung / semantische Suche: Thema in geförderten Großprojekten
  - Wachsendes Interesse an internationalen Terminologien (SNOMED CT, LOINC, RadLex…) und Ontologien (GO, HPO, …)
- Synergieeffekte vs. Ressourcenverschwendung
- Idee: Crowdsourcing-Plattform für Entwicklung deutschsprachigen Interface-Terminologien: GIT-CP

Zusätzliche Folien

Mögliche Spezifikationen für GIT-CP

- Web-basierte Crowdsourcing-Plattform
- Registrierung als User: Eingabe neuer Terme, Kommentieren und Bewerten bestehender Einträge
- Zentrales Datenelement:
  Mapping Interface Term – Externer Code
  "DM" - *81827009 |Diameter (qualifier value)*
- Attribute:
  - Ersteller, Erstellungsart, Datum, klinisches Fachgebiet, Nutzergruppe
    Max Muster, manuell, 20170803, Dermatologie Graz, Ärzte
  - Beispielannotation, z.B.
    "ein 3 cm im DM haltender Tumor"
  - Validierung / Kommentierung durch andere User
    John Doe, 20180912, ★★★★☆
    "Beispiel unverständlich – zusätzliche Beispiele!"

GIT-CP – Offene Fragen

- Technisch
  - Versionierung (GIT – Zielsysteme)
  - Schnittstellen zu lokalen Annotationsplatformen
  - Intelligente Tools (z.B. recommender services)
- Rechtlich / Organisatorisch
  - Koordination
  - Nachhaltige Finanzierung
  - Qualitätssicherung
  - Eigentumsrechte
  - Verwertung
  - Datenschutz

Fazit

- Text Mining deutscher medizinischer Texte benötigt Interfaceterminologien, die den alltäglichen Sprachgebrauch abbilden
- Die Verlinkung von Interfaceterminologien mit Referenzterminologien hat Priorität gegenüber der Übersetzung von Referenzterminologien
- Terminologiemanagement
  - Referenzterminologien: top-down, zentralisiert
  - Interfaceterminologien: bottom-up, dezentral
- Zeit ist reif für die kooperative, verteilte Erstellung einer medizinischen Interfaceterminologie für den deutschsprachigen Raum

Text Mining deutscher medizinischer Texte

i:DSem Workshop, 14.7.2017 @ Humboldt Universität zu Berlin 2017

Stefan Schulz

Medical University of Graz (Austria)

purl.org/steschu

Kontakt:

stefan.schulz@medunigraz.at