Is SNOMED CT an appropriate reference terminology for mining narrative content in electronic health records?

SNOMED CT Research Webinar - Wednesday, June 29, 2022 15 UTC

Stefan Schulz
• Medizinische Universität Graz, Austria
• Averbis GmbH, Freiburg, Germany
Discussing the appropriateness of SNOMED CT – in concert with other standards – for "understanding" clinical narratives, by transforming their content into fine-grained, precise, context-aware, unambiguous and standardised representations.
<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Example</th>
<th>Elucidation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telegram style</td>
<td>“left PICA stroke, presented to ED after fall”</td>
<td>Incomplete sentences, sketchy style</td>
</tr>
<tr>
<td>Colloquialisms</td>
<td>“pothole sign”, “snorkel”</td>
<td>Milieu-specific sub-languages</td>
</tr>
<tr>
<td>Ad-hoc abbreviations</td>
<td>“infiltr”</td>
<td>Truncation (“infiltrated mucosa”)</td>
</tr>
<tr>
<td>Ambiguous short forms</td>
<td>“RTA”</td>
<td>“Road traffic accident”, “Renal-tubular acidosis”</td>
</tr>
<tr>
<td>Short forms of regional or local scope</td>
<td>“LDS Hospital”</td>
<td>“Latter-Day-Saints Hospital” (and not “Leak Detection System”)</td>
</tr>
<tr>
<td>Conventionalized Latin abbreviations</td>
<td>“V mors can dig V dext”</td>
<td>“Vulnus morsum canis digiti quinti dextri” (in some European languages)</td>
</tr>
<tr>
<td>Numeric codes</td>
<td>“45, 46 with crowns”, “VI palsy”, “2-2-2”,</td>
<td>Tooth numbers, cranial nerves, dose frequencies</td>
</tr>
<tr>
<td>Spelling errors, typos</td>
<td>“Diabetes”, “Astra-Seneca”, “Hipotireose”,</td>
<td>accidental (quick typing) or systematic (e.g. 2nd language speakers)</td>
</tr>
<tr>
<td>Spelling variants</td>
<td>“Esophagus”, “Oesophagus”</td>
<td>e.g. American vs. British English</td>
</tr>
<tr>
<td>Single noun compounds</td>
<td>“Ibuprofenintoxikation”</td>
<td>Non-lexicalized long words (in languages such as German, Swedish)</td>
</tr>
<tr>
<td>Anaphora</td>
<td>(i) “adenoCa rect pN+MX G2 (...). tumor excised in toto”</td>
<td>(i) “Tumor” coreferential to adenocarcinom described in left context</td>
</tr>
<tr>
<td>Negations</td>
<td>(ii) “no blood in stomach (...). mult mucosal erosions”</td>
<td>(ii) “mucosal erosions” refined to “erosions of gastric mucosa”</td>
</tr>
<tr>
<td>Epistemic contexts</td>
<td>“susp MI, DD lung embolism”</td>
<td>suspected diagnosis, differential diagnosis</td>
</tr>
<tr>
<td>Temporal contexts</td>
<td>“h/o Covid-19”, “Streptokokkenangina 06/16”</td>
<td>“history of”</td>
</tr>
<tr>
<td>Other contexts</td>
<td>(i) father: pancreas ca”</td>
<td>(i) family history</td>
</tr>
<tr>
<td></td>
<td>(ii) “refrained from resuscitation”</td>
<td>(ii) plans not executed</td>
</tr>
</tbody>
</table>
Will free text documentation become obsolete?

- **Natural languages**
  - Human languages: the way how we communicate
  - Clinical languages: optimised tools for communication
  - Will not disappear as long as healthcare is delivered by humans

- **Data entry:**
  - Structured documentation: redundant, error-prone data entry
  - Popularization of speech interfaces: language-centred way of human – machine interaction becoming more popular

- **Vision**
  - Disappearing boundaries between human language (spoken / written) and structured documentation
  - Technologies: language understanding
  - Resources: semantic standards, language models
A journey from text to structure

Challenges

1. Representation of multilingual clinical sublanguages
2. Semantic enrichment and disambiguation
3. Principled, guideline-driven target representations
4. Pending issues in SNOMED CT affecting its use for NLP
1. Representation of multilingual clinical sublanguages
How to enhance dictionary coverage?

• Dictionaries with clinical expressions linked to SNOMED CT: “Interface terminologies” recommended by [*] (see comparison English / Swedish)
  • Swedish: one term per concept translated
  • English: on average 2.3 terms per concept (Preferred terms, synonyms)

• German Interface Terminology for SNOMED CT [**]
  • Combination of human and machine translation, mapping and term generation
  • Currently approx. 4 M terms mapped to SNOMED CT

German Interface terminology for SNOMED CT

• Production workflow

• Assessment using annotated parallel corpus [*]

Note: low IAA in reference standard (Krippendorff’s Alpha ~0.4)

Clinical vocabulary representation: Open issues

• Manual creation / maintenance
  • Language productivity / compositionality: impossible to collect all variations in a dictionary, as well as all ambiguous readings
  • Community processes (crowdsourcing)

• The potential of machine learning
  • Synonym detection
  • Machine translation
  • Spelling correction
  • Short form resolution
  • Word sense disambiguation

• Exploiting SNOMED CTs axiomatic structure
  • Automatic generation of paraphrases for fully defined concepts
  • Enriching embedding space

Clinical training data required
2. Semantic enrichment and disambiguation
Telegram style and hidden context hides rich semantics

Pylorus and superior duodenum: Edematous thickening. Suggestive of ulcer.
Pylorus and superior duodenum:

Edematous thickening. Suggestive of ulcer.
Pylorus and superior duodenum:

- Edematous thickening.
- Suggestive of ulcer.
Pylorus and superior duodenum: Edematous thickening. Suggestive of ulcer.
Pylorus and superior duodenum:

Edematous thickening. Suggestive of ulcer.

Finding site:

- 280119005 | Pyloric structure of stomach (body structure) |
- 56734009 | Structure of superior portion of duodenum (body structure) |
- 267038008 | Edema (finding) |
- 263899003 | Thickened (qualifier value) |
- 7196007 | Suggestive of (attribute) |
- 429040005 | Ulcer (disorder) |
Pylorus and superior duodenum:
- Edematous thickening. Suggestive of ulcer.
Leveraging SNOMED CT’s ontological structure for annotation refinement: symbolic reasoning

• Concept model: relations with domain and range restrictions:
  E.g.  
  \[ i_1 \text{ Type 'Clinical finding'} \rightarrow i_1 \text{'finding site'} i_2 \]
  \[ i_2 \text{ Type 'Body structure'} \]

• Formal definitions: templates that allow typing new instances:
  E.g.  
  \[ i_1 \text{ Type 'Ulcer'} \rightarrow i_1 \text{ Type 'Ulcer of duodenum'} i_2 \]
  \[ i_2 \text{ Type 'Duodenal structure'} \]

... based on the axiom

\[ \text{Ulcer of duodenum equivalentTo Ulcer and 'finding site' some Duodenal structure} \]
Word sense disambiguation via SNOMED ontology

Nach Verbrühung großflächige Blase

https://en.langenscheidt.com/german-english/blase

https://upload.wikimedia.org/wikipedia/commons/9/96/Brandblase_Handr%C3%BCcken.jpg
Word sense disambiguation via SNOMED ontology

Nach Verbrühung großflächige Blase

i₁

385527006 | Hot water burn of skin (disorder) |

i₂

417237009 | Blister of skin AND/OR mucosa (disorder) |

i₃

89837001 | Urinary bladder structure (body structure) |

Word sense disambiguation via SNOMED ontology

- Strategies:
  - Formal: common subsumer: ‘Finding site’ some ‘Structure of skin and/or mucous membrane’
  - Graph-theoretical: distance measure

Word sense disambiguation via Ontology and entity recognizer

LT         Sept 2012       at     ADH       by Dr Parkinson       due    to      PBC
Word sense disambiguation via Ontology and entity recognizer
Word sense disambiguation via Ontology and entity recognizer

- Procedure: LT
- Date: Sept 2012
- Location: ADH
- Person: Dr Parkinson
- Condition: PBC

18027006 [Transplantation of liver (procedure)]
3653001 [Leishmania tropica (organism)]
7671002 [Antidiuretic hormone (substance)]
3798002 [Parkinsonism (disorder)]
31712002 [Primary biliary cholangitis (disorder)]
8720001000004107 [Blood culture positive for microorganism (finding)]
Semantic enrichment and disambiguation: Open issues

• Which entities qualify as candidates for relata?
  • Identification of salient entities: Centering theory / anaphora resolution
  • Possible heuristics: “Backward-Looking”: e.g., considering the closest preceding mention of a body part concept to refine a finding or procedure concept
  • Alternative: link (relation) prediction learned from annotated corpus via neural models

• Ontological alignment of entity recognition and relation recognition frameworks with SNOMED CT concept model

3. Principled, guideline-driven target representations
Logical Polysemy

12mm

Palpebral fissure = 12 mm

https://commons.wikimedia.org/wiki/File:LIG4.jpg
Logical Polysemy

• Several, ontologically related SNOMED CT concepts qualify for annotating the same entity [*]. Examples:
  • Entire anatomical entity or any part of it (“structure concept” vs. “entire concept”)
  • Disorder (as a clinically relevant occurrent) vs. the morphological structure that characterises it
  • Observable entity (which requires a value) or a grouper finding
  • Body structure Observable
  • Pharmacological substance or product that contains the substance
  • Food stuff or the biological organism the food is taken from

• Rationale for a detailed description of a principled target representation
  • As the basis of annotation guidelines
  • For consistent use of SNOMED CT within clinical information models like FHIR

• Goals:
  • one canonical representation of all paraphrases of the same clinical content
  • Several representations that are computationally equivalent **

No interoperability due to competing semantic representations
Guideline for SNOMED CT + FHIR for annotating (in progress)

Annotation guideline for annotating clinical narratives according to SNOMED CT and FHIR

Akhila Naz Kuppaesery, Alexander Beger, Larissa Hammer, Markus Kreuzhaler, Stefan Schulz
Version 20220528

Rationale
Annotated clinical corpora are the "fuel" for the successful use of text mining and AI for interpreting clinical texts and converting their content into an interoperable format such as given by SNOMED CT and FHIR.

Guideline draft (to be discussed and enriched with examples)

[BROWSER] As a reference, the SNOMED CT browser (English) is used to find the correct codes. Only active content should be used. The decision for a code should be made according to the:
- Wording of the Fully Specified Name of a concept
- The concept's text definition (if available)
- Its formal axioms
- Its parents and children

In case two concepts fit equally well, they can be added with an OR. In case the meaning of two concepts need to be combined, they should be added with an AND.
Always copy and paste the ConceptID and the term.

[PRECOORDINATION vs. POSTCOORDINATION]
Always use pre-coordinated concepts if they represent the meaning of a text passage.

[PREFERENCE] We use for our primary annotations concepts from the following hierarchies:
- Clinical Conditions (SNOMED findings / disorders / events) related to morphology, anatomy, devices, organs
- Procedures related to anatomy, devices. Procedures also used for measurement (as long as there are no observables available)
- Observables, together with qualitative values or numbers (+ units)

Others are only used in those cases where they are clinically important and not expressible with the above hierarchies, and when the interpretation of other parts of the text depends on it.
Large project for annotating German clinical narratives

- Goal: open annotated corpus of clinical documents
- Broad range of de-identified texts (informed consent)
- Annotation guidelines to follow ontological structure of SNOMED CT and FHIR
- Kick-off 2023
Boundary between SNOMED CT and FHIR

- Overlapping area:
  HL7-FHIR resources and SNOMED CT “Situation” hierarchy

- Theory [*, **]
  - FHIR: Model of use, epistemology
  - SNOMED CT: Model of meaning, ontology

- Guidance needed:
  “TermInfo” HL7 Version 3 Implementation Guide: Using SNOMED CT in HL7 Version 3 to be revisited (inactive since 2015)

- Workaround (for own annotation activities):
  represent everything in FHIR that is possible. Exclude the SNOMED situation hierarchy. Redundancy (e.g. body site) tolerated

---


4. Pending issues in SNOMED CT affecting its use for NLP
Similar, potentially overlapping, undefined concepts

Annotation experiments[*]: Agreement between annotators (despite training and guideline)
Krippendorff’s Alpha:

<table>
<thead>
<tr>
<th>SNOMED CT (EN)</th>
<th>UMLS without SNOMED CT (EN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,37</td>
<td>0,36</td>
</tr>
</tbody>
</table>

Adolescence

SNOMED CT: undefined

NCI Thesaurus: 12 – 21 years

Dictionnaire de l’Académie Nationale de Médecine: *Période de la vie comprise entre la puberté et l’âge adulte, soit entre 12 et 18 ans* [Period of life between puberty and adulthood, i.e. between 12 and 18 years]

Larousse: *Période de la vie entre l’enfance et l’âge adulte, pendant laquelle se produit la puberté et se forme la pensée abstraite.* [Period of life between childhood and adulthood, during which puberty occurs and abstract thought is formed.]

World Health Organization: 10 – 19 years
The lonely wheelchair

- 109 concepts with “wheelchair” in the name are not linked to the concept “Wheelchair device”
Similar, potentially overlapping, undefined or underdefined concepts

- Recommendation (for better use of SNOMED in NLP):
  - Formally define or describe it
  - If not formally describable:
    - Add text definition or elucidation
    - Link to authoritative text definition
    - Link to other sources (e.g. images)
  - Standards require more than just labels and hierarchical links
  - What is ill-defined will be ill-translated
    - Affects usefulness of SNOMED CT in cross-language interoperability use cases

ICD 10 mental disorders chapter

<table>
<thead>
<tr>
<th>F40</th>
<th>Phobic anxiety disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A group of disorders in which anxiety is evoked only, or predominantly, in certain well-defined situations that are not currently dangerous. As a result these situations are characteristically avoided or endured with dread. The patient’s concern may be focused on individual symptoms like palpitations or feeling faint and is often associated with secondary fears of dying, losing control, or going mad. Contemplating entry to the phobic situation usually generates anticipatory anxiety. Phobic anxiety and depression often coexist. Whether two diagnoses, phobic anxiety and depressive episode, are needed, or only one, is determined by the time course of the two conditions and by therapeutic considerations at the time of consultation.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F40.0</th>
<th>Agoraphobia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A fairly well-defined cluster of phobias embracing fears of leaving home, entering shops, crowds and public places, or travelling alone in trains, buses or planes. Panic disorder is a frequent feature of both present and past episodes. Depressive and obsessive symptoms and social phobias are also commonly present as subsidiary features. Avoidance of the phobic situation is often prominent, and some agoraphobics experience little anxiety because they are able to avoid their phobic situations.</td>
</tr>
<tr>
<td></td>
<td>Agoraphobia without history of panic disorder</td>
</tr>
<tr>
<td></td>
<td>Panic disorder with agoraphobia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F40.1</th>
<th>Social phobias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fear of scrutiny by other people leading to avoidance of social situations. More pervasive social phobias are usually associated with low self-esteem and fear of criticism. They may present as a complaint of blushing, hand tremor, nausea, or urgency of micturition, the patient sometimes being convinced that one of these secondary manifestations of their anxiety is the primary problem. Symptoms may progress to panic attacks.</td>
</tr>
<tr>
<td></td>
<td>Anthroaphobia</td>
</tr>
<tr>
<td></td>
<td>Social neurosis</td>
</tr>
</tbody>
</table>
• For the purpose of NLP, SNOMED CT is already a highly useful resource
  • Conceptual coverage and granularity
  • Ontologically grounded categories and relations
  • Formal axioms

• SNOMED CT cannot be expected to represent idiosyncratic clinical jargon
  • For NLP, interface terminologies need to be created, maintained, and linked to SNOMED CT
  • Even though, fault tolerant entity normalization will be necessary
• Really interoperable representations require in-depth annotation rules
  • Preference rules for choosing between competing codes
  • Set of canonic relations, compatible with the SNOMED CT concept model and ontological principles

• Precision medicine needs precision standards
  • Still too much underdefined and fuzzy content in SNOMED CT – difficult for annotation, difficult for interface term mapping (particularly cross-language)
  • Specification of the content of SNOMED CT concept descriptions, addition of (textual) definitions
  • Clarifying the boundary between SNOMED CT and FHIR, preventing the proliferation of incompatible annotation patterns
• Desiderata for the NLP / Text mining / AI communities
  • Produce and publish annotated corpora and language models for clinical language in main natural languages
  • Create a legal and ethical framework that allows the use of clinical texts as training resources
  • Follow standards (SNOMED CT, FHIR) and ontology principles when developing annotation schemes for clinical content
  • Leverage the potential of knowledge graph technologies
  • Optimise deep learning technology including machine translation for accelerating the acquisition of translations and synonyms

Contact: stefan.schulz@medunigraz.at