ODLS 2017 Ontologies & Data in Life Sciences

Lexical ambiguity in SNOMED CT

Stefan Schulz
Catalina Martínez-Costa
Jose Antonio Miñarro-Giménez

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

Background

- SNOMED CT: largest clinical terminology / ontology (English: 300 k concepts, 750k terms)
- Two aspects:
  - SNOMED CT as a domain ontology: Labels (FSNs), tentatively self-explaining; formal descriptions and definitions (EL++), still few free text elucidations
  - SNOMED CT as a domain terminology: at least for English, enrichment with quasi-synonyms ("interface terms")

SNOMED International (aka IHTSDO). SNOMED CT http://www.snomed.org/snomed-ct

Ontology labels vs. Interface terms

Labels

- Self-explaining
- Univocal
- Long
- Unabridged
- Unpopular
- Should be understandable independent of contex

"Primary malignant neoplasm of lung"
"Leishmania tropica"
"Electrocardiogram"
"Diagnosis"

Interface terms

- Not self-explaining
- Ambiguous
- Short
- Abridged (Acronyms)
- Popular
- Depend on user groups, by specialty, institution, dialect

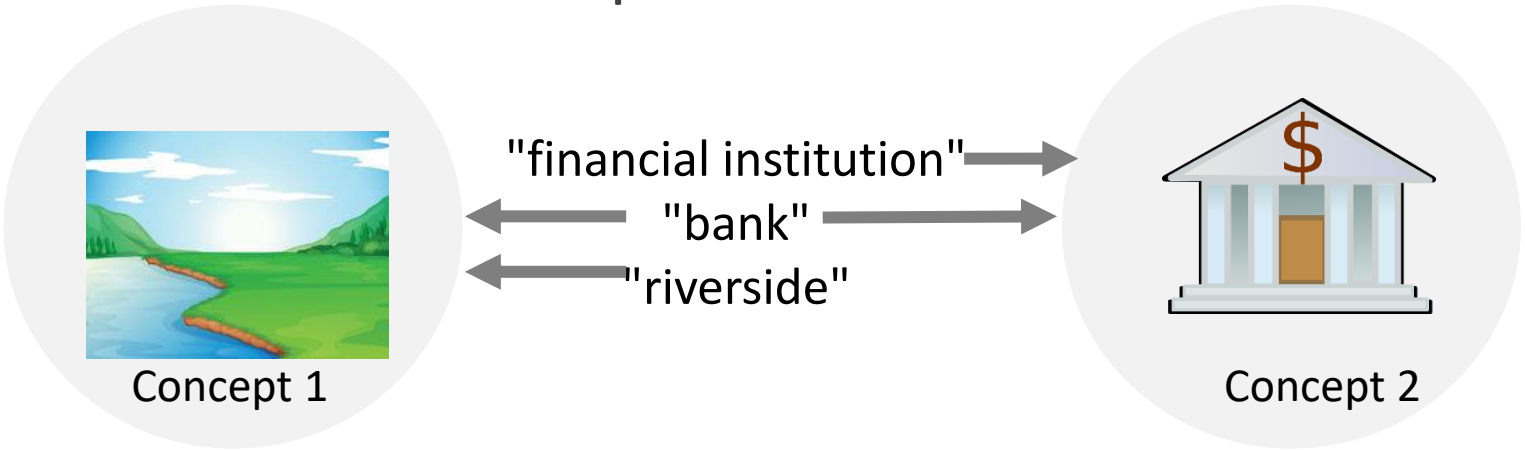
"Ca Lung "
"LT"
"ECG"
"Dx"

Popularity of terms (Pubmed titles and abstracts)

FSN (SNOMED CT)	Count	SNOMED CT synonyms	Count
Primary malignant neoplasm of lung	0	Lung cancer Bronchial carcinoma	120682 3452
Cerebrovascular accident	3819	Stroke	191559
Block dissection of cervical lymph nodes	1	Neck dissection	7512
Electrocardiographic procedure	1	Electrocardiogram ECG	33670 55120
Backache	3489	Back pain	38132
Capillary blood specimens	32	Capillary blood samples	574

Lexical ambiguity in a nutshell

- "Term" and "concept" are two fundamentally different things:
  - Concepts/classes/types/categories: units of language-independent meaning
  - (Natural language) Terms: units of language, connected to concepts



Main questions

- Why should ontologies care about ambiguity aspects at all when studying ontology?
  - User acceptance of ontology-based systems
  - Quality of structured data entry
  - Use of ontology in NLP scenarios
- How is lexical ambiguity related to the ontology issues proper?
  - Completeness and quality of ontology content
  - Complex categories

Understanding better SNOMED CT naming

- Fully specified names
  - Unique – 1 : 1 relation with codes
  - Carry a "hierarchy tag"
  - Without hierarchy tag (e.g. for term matching in texts), ambiguity may arise:
    Lymphoma (disorder) vs. Lymphoma (morphology)
- Synonyms
  - May be ambiguous
- Short forms
  - Entries not ambiguous because accompanied by expanded form, e.g.
    PIN - Prostatic intraepithelial neoplasia
    Pressure-induced nystagmus

Scrutiny of ambiguous terms in SNOMED CT

- SNOMED CT January 2017 release: Extract ambiguous entries
  - Full terms (without hierarchy tags) → D_1
  - Acronyms (without abbreviations) → D_2
- Analysis:
  - Count ambiguities and their cardinality
  - SNOMED CT hierarchies to which ambiguous terms belong
  - Ambiguous terms that are related via non-taxonomic links (e.g. **Associated morphology** or **Has active Ingredient**)
  - Ambiguous terms that are related via taxonomic links (**is-a**)
- Purpose: Detect regularities, spot errors, derive recommendations to SNOMED Intl.

Results: Frequency and Distribution

- **Frequency and distribution of ambiguous readings of SNOMED CT terms**

Dictionary	Count	Cardinality		Maximum
		Mean	Median	
D_1 (non-acronym terms)	7,439	2.02	2	6
D_2 (acronyms)	899	5.54	2	1678

Results D_1

- **Leading patterns of concept tuples connected by the same SNOMED CT (non-acronym) term**

Hierarchy tag combination patterns	Pattern count	Rate of non-taxonomic links	Rate of taxonomic links
| product | substance |	4,064	0.888	0.000
| disorder | morphologic abnormality |	1,047	0.707	0.000
| organism | organism |	221	0.000	0.452
| procedure | substance |	213	0.911	0.000
| procedure | procedure |	200	0.000	0.465
Other n-tuples ($2 \leq n \leq 6$)	1,694		

- Strict implications, e.g.
  '*Folinic acid (product)*' subclassOf '**Has active ingredient**' some '*Folinic acid (substance)*'
- "Dot types" (logical polysemy)
  '*Solar keratosis (disorder)*' subclassOf '**Associated morphology**'
  some '*Solar keratosis (morphologic abnormality)*'

Arapinis A,Vieu L (2015). A plea for complex categories in ontologies. Applied Ontology, 10(3-4), 285-296.

Results D_2

- **Leading patterns of concept tuples linked by the same acronym extracted from SNOMED CT terms**

Hierarchy tag combination Patterns	Pattern count	Rate of non-taxonomic links	Rate of taxonomic links
| disorder | disorder |	66	0.015	0.167
| disorder | procedure |	59	0.034	0.000
| procedure | procedure |	38	0.000	0.263
| procedure | substance |	33	0.333	0.000
| disorder | substance |	28	0.000	0.000
Other n-tuples ($2 \leq n \leq 1678$)	675		

- **Distribution of patterns much more evely**
- **Acronym naming pattern not specific:**
  e.g., *O/E – eye, O/E – nose, O/E – mouth, O/E – heart etc.*

Conclusion

- Degree of Lexical ambiguity in SNOMED CT moderate
- Ontological aspect:
  - ontologically dependent concepts, partly interpretable as complex categories (dot types)
- Lexical aspect:
  - Amount of ambiguous acronyms lower than expected → risk of wrong mappings
- Naming aspect:
  - Acronym – expansion patterns not specific: → wrong expansions
- **Should interface terms (synonyms) be managed by the ontology curators?**