Using lexical and structural features for quality assurance of biomedical ontologies

*Application to SNOMED CT*

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.
SNOMED Clinical Terms
SNOMED CT Characteristics

- Developed by SNOMED International
  - Consortium of 30 member countries
- Largest clinical terminology in the world
  - ~320,000 active concepts
  - ~1 million terms ("descriptions")
- Major organizing principles
  - Logical definitions (incomplete: many primitives)
  - Built using description logics (EL++)

http://www.snomed.org/
SNOMED CT  Top level

- SNOMED CT Concept
  - Body structure (body structure)
  - Clinical finding (finding)
  - Environment or geographical location (environment / location)
  - Event (event)
  - Observable entity (observable entity)
  - Organism (organism)
  - Pharmaceutical / biologic product (product)
  - Physical force (physical force)
  - Physical object (physical object)
  - Procedure (procedure)
  - Qualifier value (qualifier value)
  - Record artifact (record artifact)
  - Situation with explicit context (situation)
  - SNOMED CT Model Component (metadata)
  - Social context (social concept)
  - Special concept (special concept)
  - Specimen (specimen)
  - Staging and scales (staging scale)
  - Substance (substance)
SNOMED CT Example

Parents
- Operation on appendix (procedure)
- Partial excision of large intestine (procedure)

 Appendectomy (procedure)
SCTID: 80146002
80146002 | Appendectomy (procedure) |
Appendectomy
Excision of appendix
Appendicectomy
Appendectomy (procedure)

Procedure site - Direct → Appendix structure
Method → Excision - action

Children (8)
- Appendectomy with drainage (procedure)
- Emergency appendectomy (procedure)
- Excision of appendiceal stump (procedure)
- Excision of ruptured appendix by open approach (procedure)
- Incidental appendectomy (procedure)
- Interval appendectomy (procedure)
- Laparoscopic appendectomy (procedure)
- Non-emergency appendectomy (procedure)
---

**SNOMED CT Challenges**

- **Legacy**
  - Many primitive concepts
  - Not amenable to automatic DL classification

- **Maintenance**
  - Developed by many human editors
  - Error prone

- **Quality assurance**
  - Difficult due to its size
  - Ontology design patterns ("concept model")
    - Difficult to apply retrospectively

---

*Lister Hill National Center for Biomedical Communications*
Quality assurance approaches
Quality assurance approaches

- Three types of QA approaches applied to SNOMED CT by researchers

  - Lexical
  - Structural
  - Semantic
Lattice-based structural QA

◆ Lattice
  ● Specific type of directed acyclic graph (DAG)
  ● Any two nodes have a unique maximal common descendant, as well as a unique minimal common ancestor

◆ Non-lattice fragments are often indicative of a problem in ontology construction
  ● Missing hierarchical relation
  ● Missing intermediary concept
Example of non-lattice fragment

Example #1

Jiang et al., JAMIA, 2009

Diagram:
- Hypophysectomy
  - Partial hypophysectomy
  - Transcranial hypophysectomy
    - Transsphenoidal hypophysectomy
    - Partial excision of pituitary gland by transfrontal approach
  - Partial excision of pituitary gland by transsphenoidal approach

Upper bounds
Lower bounds
Missing intermediary concept

Example #1

- Hypophysectomy
  - Partial hypophysectomy
  - Transcranial hypophysectomy
  - Partial transcranial hypophysectomy
    - Partial excision of pituitary gland by *transfrontal* approach
  - Transsphenoidal hypophysectomy
    - Partial excision of pituitary gland by *transsphenoidal* approach
Non-lattice fragment in SNOMED CT

Example #2

Acute respiratory disease

Chronic bronchitis

Acute exacerbation of chronic bronchitis

Chronic obstructive bronchitis

Acute exacerbation of chronic obstructive bronchitis
Missing hierarchical relation

- Acute respiratory disease
- Chronic bronchitis
- Acute exacerbation of chronic bronchitis
- Acute exacerbation of chronic obstructive bronchitis
- Chronic obstructive bronchitis
Non-lattice fragment in SNOMED CT

- Acute respiratory disease
- Chronic bronchitis
- Acute exacerbation of chronic bronchitis
- Acute exacerbation of chronic obstructive bronchitis
- Chronic obstructive bronchitis
Lexical differences among terms are often indicative of semantic relations among them.
# Suggested missing hierarchical relations

<table>
<thead>
<tr>
<th>Child name</th>
<th>Parent name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alveolar bone graft to mandible</td>
<td>Alveolar bone graft</td>
</tr>
<tr>
<td>Basal cell carcinoma of skin of lip</td>
<td>Carcinoma of lip</td>
</tr>
<tr>
<td>Carcinoma in situ of palate</td>
<td>Palate carcinoma</td>
</tr>
<tr>
<td>Chronic bacterial otitis externa</td>
<td>Chronic otitis externa</td>
</tr>
<tr>
<td>Congenital vascular anomaly of eyelid</td>
<td>Vascular anomaly of eyelid</td>
</tr>
<tr>
<td>Electrocoagulation of retina for repair of tear</td>
<td>Repair of retina</td>
</tr>
</tbody>
</table>
Objectives

- To combine lexical and structural QA approaches to automatically and precisely identifying missing hierarchical relations and missing concepts in SNOMED CT
- To suggest remediation for such inconsistencies

Materials: September 2015 version of SNOMED CT (U.S. edition)
Methods & Results
Overview

- Identifying non-lattice pairs and subgraphs
- Identifying lexical patterns indicative of missing concepts and relations
- Analyzing non-lattice subgraphs with lexical patterns
- Evaluation
Identifying non-lattice pairs and subgraphs

◆ Hadoop-based technique
◆ 30 hours to analyze all pairs of SNOMED CT concepts
◆ Aggregation of non-lattice pairs with the same shared ancestors into non-lattice subgraphs
  ● Smaller subgraphs contained in larger subgraphs
◆ 631,006 non-lattice pairs
◆ 171,011 non-lattice subgraphs
  ● Focus on small subgraphs
Lexical patterns (1) Containment

- The set of words for one concept in the upper (resp. lower) bounds is contained in the set of words for another concept in the upper (resp. lower) bounds.
- Suggests a *missing hierarchical relation* between concepts in the upper (resp. lower) bounds.
- 736 small non-lattice subgraphs with this pattern.
Lexical patterns (1) Containment

- Non-lattice subgraph
- Suggested remediation

Duodenal ulcer with perforation AND obstruction ⊃ Chronic duodenal ulcer with perforation AND obstruction
Lexical patterns (2) Intersection

- The intersection of sets of words for concepts in the lower bounds is equal to the set of words for some concept in the upper bounds.
- Suggests a *missing hierarchical relation* between concepts in the upper bounds.
- 1085 small non-lattice subgraphs with this pattern.
Lexical patterns (2) Intersection

Non-lattice subgraph

Suggested remediation

Irritable bowel syndrome

Irritable bowel syndrome
variant of childhood

Irritable bowel syndrome
with diarrhea

Disorder of colon

Disorder of colon

Irritable bowel syndrome
variant of childhood

Irritable bowel syndrome
with diarrhea
Lexical patterns (3) Union

- The union of the sets of words for concepts in the upper bounds is equal to the set of words for some concept in the lower bounds.
- Suggests a missing hierarchical relation between concepts in the lower bounds.
- 164 small non-lattice subgraphs with this pattern.
Lexical patterns (3) Union

Non-lattice subgraph

Suggested remediation

Epithelial neoplasm of skin $\bigcup$ Malignant neoplasm of skin

Malignant epithelial neoplasm of skin
Lexical patterns (4) Union-Intersection

- The union of the sets of words for concepts in the upper bounds is equal to the intersection of sets of words for concepts in the lower bounds.
- Suggests a *missing intermediary concept* between the upper bounds and the lower bounds.
- 61 small non-lattice subgraphs with this pattern.
Lexical patterns (4) Union-Intersection

Non-lattice subgraph

Suggested remediation

Neoplasm of right upper lobe of lung

Malignant neoplasm of upper lobe of lung

Secondary malignant neoplasm of right upper lobe of lung

Primary malignant neoplasm of right upper lobe of lung

Malignant neoplasm of right upper lobe of lung

Secondary malignant neoplasm of right upper lobe of lung

Primary malignant neoplasm of right upper lobe of lung

Secondary malignant neoplasm of right upper lobe of lung

Primary malignant neoplasm of right upper lobe of lung

Neoplasm of right upper lobe of lung ∪ Malignant neoplasm of upper lobe of lung = Secondary malignant neoplasm of right upper lobe of lung ∩ Primary malignant neoplasm of right upper lobe of lung
Evaluation

- 59 subgraphs independently reviewed by 2 experts after triaging
  - Differences resolved by discussion
- All contained errors – 61 errors
  - Missing hierarchical relation: 59
  - Missing intermediary concept: 2
- Lexical patterns
  - Containment: 34; Intersection: 14; Union: 8; U/I: 3
- Suggested remediation
  - Accepted for 53 subgraphs
  - Rejected for 6 subgraphs (deeper modeling issues)
Discussion
Significance

- Most terminology QA techniques merely identify potential errors

- Our approach
  - Identified unreported errors
    - Confirmed by experts
  - Suggested appropriate remediation in many cases

- Should greatly facilitate error correction by the developers of SNOMED CT

- Scalable and applicable to other terminologies
Limitations and future work

◆ Suggested remediation (e.g., to add missing hierarchical relations) is based on the inferred concept hierarchy of SNOMED CT
  ● Does not address the root cause (e.g., incomplete/inaccurate logical definition)
  ● Root cause needs to be addressed by the SNOMED CT editors

◆ Only 4 lexical patterns considered
  ● Could be refined with additional patterns

◆ Only used the preferred terms
  ● Could also use synonyms
Current work

◆ Maximize the chances for word set inclusion among concepts within the non-lattice subgraph (beyond the words of the terms themselves)

◆ Enrich the lexical definitions with
  ● Words from ancestor concepts (within the non-lattice subgraph)
  ● Additional hypernyms extracted from hierarchically related concepts (within the non-lattice subgraph)
Additional hypernyms

L6 = \{\text{fracture, subluxation, of, perilunate, joint}\}

L4 = \{\text{fracture, dislocation, of, perilunate, joint}\}
Lexical features enriched with words from ancestors

L1 = \{fracture, dislocation, of, lunate\}  
L2 = \{fracture, subluxation, of, wrist\}  
L3 = \{fracture, subluxation, of, lunate, wrist\}  
L4 = \{fracture, dislocation, of, perilunate, joint, lunate\}  
L5 = \{open, fracture, subluxation, lunate, dislocation, of, wrist\}  
L6 = \{fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist\}
Lexical features enriched with ancestors and hypernyms

- $L_1 = \{\text{fracture, dislocation, of, lunate}\}$
- $L_2 = \{\text{fracture, subluxation, of, wrist, dislocation}\}$
- $L_3 = \{\text{fracture, subluxation, of, lunate, wrist, dislocation}\}$
- $L_4 = \{\text{fracture, dislocation, of, perilunate, joint, lunate}\}$
- $L_5 = \{\text{open, fracture, subluxation, lunate, dislocation, of, wrist}\}$
- $L_6 = \{\text{fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist}\}$
Missing hierarchical relation

1. Fracture dislocation of lunate
   \[ L_1 = \{\text{fracture, dislocation, of, lunate}\} \]

2. Fracture subluxation of wrist
   \[ L_2 = \{\text{fracture, subluxation, of, wrist, dislocation}\} \]

3. Fracture subluxation of lunate
   \[ L_3 = \{\text{fracture, subluxation, of, lunate, wrist, dislocation}\} \]

4. Fracture dislocation of perilunate joint
   \[ L_4 = \{\text{fracture, dislocation, of, perilunate, joint, lunate}\} \]

5. Open fracture subluxation lunate
   \[ L_5 = \{\text{open, fracture, subluxation, lunate, dislocation, of, wrist}\} \]

6. Fracture subluxation of perilunate joint
   \[ L_6 = \{\text{fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist}\} \]
Non-lattice $\rightarrow$ lattice subgraph

1. Fracture dislocation of lunate
2. Fracture subluxation of wrist
3. Fracture subluxation of lunate
4. Fracture dislocation of perilunate joint
5. Open fracture subluxation lunate
6. Fracture subluxation of perilunate joint
Summary

◆ The combination of structural (non-lattice) and lexical (words from terms) features is effective in identifying missing hierarchical relations in SNOMED CT with high precision

◆ This approach would be applicable to a large number of ontologies