A Deep Learning Approach to Identifying Missing Hierarchical Relations in SNOMED CT

Rashmie Abeysinghe
Mentor: Dr. Olivier Bodenreider
Outline

• SNOMED CT
• Terminology Quality Assurance
• Non-Lattice Subgraphs and structural-lexical auditing methods
• Graph Neural Networks
• Sample Generation
• Architecture of the neural network
• Results
SNOMED CT

• Developed by SNOMED International*
• Largest clinical healthcare terminology in the world
• More than 350,000 active concepts
• Has 19 top-level sub-hierarchies

*http://www.snomed.org/
Terminology Quality Assurance (TQA)

• Errors exist in terminologies*
• TQA: Essential part of terminology management lifecycle
• Manual auditing: labor intensive & time consuming
• Automating TQA: active area of research

Non-Lattice Subgraphs (NLSs)

• Lattice – a desirable property for a well-formed terminology*

• Lattice – a DAG such that any two nodes have a unique maximal common descendant as well as a unique minimal common ancestor

Containment lexical pattern

*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798*
Union lexical pattern

*E*  
- Epithelial neoplasm of skin  
- Malignant neoplasm of skin  
- Malignant epithelial neoplasm of skin  
- Squamous cell carcinoma of skin

*F*  
- Epithelial neoplasm of skin  
- Malignant epithelial neoplasm of skin  
- Squamous cell carcinoma of skin

---

*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798*
Lexical Patterns

• Manually curated
• Majority of NLSs does not exhibit any of the lexical patterns currently identified
• Not sustainable
Objective

- Develop a deep learning approach that can identify the existence of and the direction of a missing is-a among two concepts in the upper/lower bounds of a non-lattice subgraph in SNOMED CT.
Handling Graph Data with Neural Networks

• Traditional neural network architectures do not handle graph data well

• Graph data are irregular:
  • Each graph has a variable size of nodes
  • Each node has a different number of neighbors
  • Neighbors of a node cannot be ordered in a graph
Graph Neural Networks (GNNs)

- Key idea: Generate node embeddings based on local neighborhoods.
- Intuition: Nodes aggregate information from their neighbors using neural networks.

Sample Generation

- 117,625 concepts in Clinical Findings
- 210,349 relations
- 14 billion non-relations!

- Most non-relations may be uninteresting
- Only non-relations types that occur in NLSs needs to be used
- To identify the non-relation type, we use edge-separation:
  - Edge-separation: sum of edges from each concept to a common ancestor

Edge-separation of A & B = 3
Sample Generation

- Edge-separation 2

- Edge-separation 3

Infection of ovary

- Mumps oophoritis
- Tuberculous oophoritis

Pelvic injury

- Contusion of sacral region
- Abrasion of pelvic region

Abrasion of perineum
## Edge-separation of non-relations in NLSs

<table>
<thead>
<tr>
<th>Edge-separation</th>
<th>Number of non-edges</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>24514</td>
<td>63.04%</td>
</tr>
<tr>
<td>3</td>
<td>5342</td>
<td>13.74%</td>
</tr>
<tr>
<td>4</td>
<td>3500</td>
<td>9.00%</td>
</tr>
<tr>
<td>5</td>
<td>2148</td>
<td>5.52%</td>
</tr>
<tr>
<td>6</td>
<td>1368</td>
<td>3.52%</td>
</tr>
<tr>
<td>7</td>
<td>926</td>
<td>2.38%</td>
</tr>
<tr>
<td>8</td>
<td>568</td>
<td>1.46%</td>
</tr>
<tr>
<td>9</td>
<td>290</td>
<td>0.75%</td>
</tr>
<tr>
<td>10</td>
<td>148</td>
<td>0.38%</td>
</tr>
<tr>
<td>11</td>
<td>64</td>
<td>0.16%</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>0.03%</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.01%</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
Sample Generation

• Positive samples:
  • 210,349 relations

• Negative samples:
  • 210,349 non-relations with same edge-separation distribution as NLS non-relations
  • 210,349 non-relations obtained by reversing existing is-a relations

• 90% for training and validation:
  • 567,942 samples

• 10% for testing:
  • 63,105 samples
Sample Subgraph Generation

• Context of a concept: consists of ancestors and descendants up to two levels.
• Concept-pair of the sample and the concepts in their contexts will define a subgraph
Toxic shock syndrome caused by methicillin resistant Staphylococcus aureus infection

Toxic shock syndrome caused by methicillin susceptible Staphylococcus aureus

Infection caused by methicillin susceptible Staphylococcus aureus

Infection caused by Staphylococcus aureus

Toxic shock syndrome

Staphylococcal toxic shock syndrome

Staphylococcal infectious disease

Disease caused by Gram-positive cocci

Disease caused by Gram-positive bacteria

Septic shock

Disease caused by Gram-positive bacteria

Disease caused by Gram-positive cocci
Negative Sample

- **A** Cutaneous diphtheria
- **B** Brucellosis of skin
- **C** Bacterial infection of skin
- **D** Bacterial infection by site
- **E** Infection of skin
- **F** Disease caused by Gram-negative bacillus
- **G** Infection caused by Corynebacterium
- **F** Brucellosis
- **F** Diphtheria
- **G** Infection caused by Corynebacterium
Node Features (at Layer-0 of GNN)

• Trained a Doc2Vec* model to obtain embeddings for preferred terms of each concept
  • Doc2Vec: Unsupervised algorithm to generate vectors for sentences/paragraphs/documents etc.
  • Vector Size: 150

Architecture of the Network

- GCN + ReLU + Dropout
- Concatenated embeddings of A & B
- FC + ReLU + Dropout
- Sigmoid
- Output
Training

• GNN library: Deep Graph Library (DGL): A python package dedicated to deep learning on graphs. Built atop Pytorch.
• Ran on NVIDIA Tesla K20X GPUs (Biowulf)
• Loss function: Binary Cross Entropy
• Optimizer: Adam
• Learning rate: 0.001
• Batch size: 128
• 100 epochs
• Performed 6-fold cross validation
## Performance

<table>
<thead>
<tr>
<th></th>
<th>Cross validation</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.8629</td>
<td>0.8439</td>
</tr>
<tr>
<td>Recall</td>
<td>0.8475</td>
<td>0.8291</td>
</tr>
<tr>
<td>F1</td>
<td>0.8552</td>
<td>0.8364</td>
</tr>
</tbody>
</table>
Relations identified correctly (true positives)

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duodenal papilla not found</td>
<td>Digestive system finding</td>
</tr>
<tr>
<td>Cyproterone adverse reaction</td>
<td>Antineoplastic adverse reaction</td>
</tr>
<tr>
<td>Mosaic trisomy 5 syndrome</td>
<td>Anomaly of chromosome pair 5</td>
</tr>
<tr>
<td>Acute hepatitis</td>
<td>Inflammatory disease of liver</td>
</tr>
</tbody>
</table>
Non-relations identified correctly (true negatives)

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic gingivitis</td>
<td>Chronic fibrous gingivitis</td>
</tr>
<tr>
<td>Finding of urine drug level</td>
<td>Acetaminophen in urine</td>
</tr>
<tr>
<td>Open fracture of proximal phalanx of left thumb</td>
<td>Closed fracture thumb proximal phalanx, head</td>
</tr>
<tr>
<td>Dextrotransposition of aorta</td>
<td>Transposition of aorta</td>
</tr>
</tbody>
</table>
### Relations not identified (false negatives)

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traumatic arthropathy of metacarpophalangeal joint</td>
<td>Traumatic arthropathy of the hand</td>
</tr>
<tr>
<td>Primary basal cell carcinoma of right lower limb</td>
<td>Basal cell carcinoma of lower extremity</td>
</tr>
<tr>
<td>Pulmonary aspiration of gastric contents</td>
<td>Pulmonary aspiration of fluid</td>
</tr>
<tr>
<td>Lesion of radial nerve</td>
<td>Radial neuropathy</td>
</tr>
</tbody>
</table>
Non-relations not identified (false positives)

<table>
<thead>
<tr>
<th>Child</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implanted defibrillator generator infection</td>
<td>Infected pacemaker</td>
</tr>
<tr>
<td>Open fracture of right angle of mandible</td>
<td>Open fracture of zygoma</td>
</tr>
<tr>
<td>Finding of cochlear function</td>
<td>Cochlear microphonic</td>
</tr>
<tr>
<td>Aluminum hydroxide overdose</td>
<td>Kaolin overdose</td>
</tr>
</tbody>
</table>
Applying the trained model on NLSs

- 7,141 small (size 4,5,6) non-lattice subgraphs in Clinical Finding subhierarchy
- 37,404 non-relations in lower and upper bounds
- 11,943 missing is-a predicted
Comparison with lexical patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th># missing is-a obtained by pattern</th>
<th># common missing is-a obtained by classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Containment</td>
<td>406</td>
<td>150</td>
</tr>
<tr>
<td>Union</td>
<td>110</td>
<td>43</td>
</tr>
<tr>
<td>Intersection</td>
<td>542</td>
<td>370</td>
</tr>
</tbody>
</table>

*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798*
Conclusions and Future Work

- We employ a deep learning approach to identify missing relations in SNOMED CT
- Samples are generated based on existing relations and non-relations in SNOMED CT
- After training the model, we apply it to lower and upper bounds of non-lattice subgraphs to identify missing relations:
- Future work: train the model on relations and non-relations similar to those that appear in non-lattice subgraphs
Acknowledgement

• Dr. Olivier Bodenreider
• Dr. Licong Cui
• Dr. Paul Fontelo
• Dr. Vinh Nguyen
• Joey Yip
• Karan Luthria
Thank you!