Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization

Zhe He, PhD\textsuperscript{1,2}

James Geller, PhD\textsuperscript{2}, Gai Elhanan, MD\textsuperscript{3}

\textsuperscript{1}Department of Biomedical Informatics, Columbia University
\textsuperscript{2}Department of Computer Science, New Jersey Institute of Technology
\textsuperscript{3}Halfpenny Technologies, Inc

2014 AMIA Summit on Translational Bioinformatics
April 8, 2014
Disclosure

• Zhe He discloses that he has no relationships with commercial interests.
• James Geller discloses that he has no relationships with commercial interests.
• Gai Elhanan discloses that he has no relationships with commercial interests.
Learning Objective

• Use structural method to find potential concepts for enriching the conceptual content of a biomedical terminology
Overview

• Motivation
  – Exploring structural method for semantic harmonization
• Background
  – Importance of the conceptual content of SNOMED CT
• Methods
  – Structural matching of pairs of terminologies in the UMLS
• Results
  – Reusable knowledge can be derived by structural matching, including discovery of possible synonym
• Limitations and Future Work
• Conclusions
Motivation

- Need of well-developed and well-maintained terminologies
- NLP tools that process clinical text need a terminology with fruitful concepts and synonyms.
- Complex clinical research texts require combined use of multiple terminologies (Weng et al. 2010)
- Terminologies need harmonization to achieve semantic interoperability (Bittner et al. 2005)
Semantic Harmonization Between Different Terminologies

- (Weng et al. 2010)
- Harmonized existing time ontologies for annotating temporal relation in clinical narratives (Tao et al. 2011)
- Semantic harmonization efforts have recently been extended for various terminologies
  - SNOMED CT and LOINC (AMIA 2013 Informatics Year in Review)
  - SNOMED CT and ICD 11 (Rodrigues et al. 2013)
Importance of Conceptual Content of SNOMED CT

• SNOMED CT is going to be a major terminology for EHR encoding of diagnoses and problem lists by 2015
• SNOMED CT has many problems!
• Top two mentioned deficiencies of SNOMED CT (Elhanan 2011):
  – Missing concepts – 23%
  – Missing synonyms – 17%
• Users will expect SNOMED CT to have correct synonyms and sufficient concepts to be used in EHR
Leveraging Common Structure of Pairs of Terminologies in the UMLS

• **UMLS (Unified Medical Language System):**
  – More than 170 source terminologies, 8.9 million terms, 2012AB release

Breast Cancer (NCIt)  Carcinoma breast (MedDRA)  Carcinoma of breast (SNOMED)  etc.

Breast Carcinoma (UMLS)
Structurally Congruent Concepts

1) X and Y are alternative classifications
2) X can be a parent of Y
3) Y can be a parent of X
4) X and Y are synonymous
5) Structural errors in Terminology 1
6) Structural errors in Terminology 2

Cycles were eliminated during the processing
Alternative Classification

Concept A

Classified by assumption of Terminology 1 Designer

Concept X

Concept B

same concept

Concept A

Classified by assumption of Terminology 2 Designer

Concept Y

Concept B

same concept
Parent-Child Relationship

X can be a parent of Y

Y can be a parent of X

Concept X

Concept Y

Concept Y

Concept X

IS_A

IS_A
Synonymous

X and Y are synonymous but have not been identified by the UMLS

Concept X (Synonym: Y)
META Terminologies in this Study

• Metathesaurus terminologies with “PAR” relationship and “inverse_isa” relationship attribute were chosen
• SCTUSX and UWDA were excluded
• Reference Terminologies (Terminology 1):
  – MEDCIN (MEDCIN)
  – National Cancer Institute Thesaurus (NCIt)
  – Gene Ontology (GO)
  – Medical Entities Dictionary (CPM)
  – UMDNS: product category thesaurus (UMD)
  – Foundational Model of Anatomy Ontology (FMA)
• Terminology 2: SNOMED CT
# Evaluation: Pairs of Congruent Concepts Found

<table>
<thead>
<tr>
<th>Reference Terminology</th>
<th>Size of Terminology</th>
<th># of Pairs of Congruent Concepts</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>279529</td>
<td>655</td>
<td>70</td>
</tr>
<tr>
<td>NCI</td>
<td>95523</td>
<td>582</td>
<td>70</td>
</tr>
<tr>
<td>FMA</td>
<td>82062</td>
<td>116</td>
<td>70</td>
</tr>
<tr>
<td>UMD</td>
<td>15956</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>GO</td>
<td>61925</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>CPM</td>
<td>3078</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>--</td>
<td>1384</td>
<td>241</td>
</tr>
</tbody>
</table>
Review Results for Pairs of Congruent Concepts

Co-author GE reviewed the sample

<table>
<thead>
<tr>
<th>Reference Terminology</th>
<th>Sample Size</th>
<th>Alternative Classification</th>
<th>Y $\rightarrow$ X</th>
<th>X $\rightarrow$ Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>70</td>
<td>44</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>NCI</td>
<td>70</td>
<td>38</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>GO</td>
<td>6</td>
<td>2</td>
<td>--</td>
<td>4</td>
</tr>
<tr>
<td>CPM</td>
<td>7</td>
<td>5</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>UMD</td>
<td>18</td>
<td>9</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>FMA</td>
<td>70</td>
<td>45</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>241</td>
<td>143</td>
<td>36</td>
<td>21</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>59.3%</td>
<td>14.9%</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

23.6%
### Review Results for Pairs of Congruent Concepts

<table>
<thead>
<tr>
<th>Reference Terminology</th>
<th>Sample Size</th>
<th>Error in Terminology 1</th>
<th>Error in Terminology 2</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>70</td>
<td>--</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>NCI</td>
<td>70</td>
<td>--</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>GO</td>
<td>6</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>CPM</td>
<td>7</td>
<td>--</td>
<td>--</td>
<td>2</td>
</tr>
<tr>
<td>UMD</td>
<td>18</td>
<td>--</td>
<td>--</td>
<td>8</td>
</tr>
<tr>
<td>FMA</td>
<td>70</td>
<td>2</td>
<td>--</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>181</td>
<td>2</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>0.8%</td>
<td>1.7%</td>
<td><strong>14.5%</strong></td>
</tr>
</tbody>
</table>
Example: Alternative Classification

- Structure of posterior intercostal vein, C0226639
- **Eleventh** posterior intercostal vein, C0506471
- Eleventh right posterior intercostal vein, C0501203

- Structure of posterior intercostal vein, C0226639
- **Lower right** posterior intercostal veins, C1283497
- Eleventh right posterior intercostal vein, C0501203

FMA3_1

SNOMEDCT_2012_07_31
Making Explicit an Implicit Assumption of the Two Original Terminology Designers

Structure of posterior intercostal vein, C0226639

Posterior intercostal vein classified by ordinality

Eleventh posterior intercostal vein, C0506471

Eleventh right posterior intercostal vein, C0501203

FMA3_1

Structure of posterior intercostal vein, C0226639

Posterior intercostal vein classified by position

Lower right posterior intercostal veins, C1283497

Eleventh right posterior intercostal vein, C0501203

SNOMEDCT_2012_07_31
Example: Parent of the Other

- Sign and Symptoms, C0037088
- Finding by Site or System, C1333618
- Integumentary System Finding, C1291044

- Sign and Symptoms, C0037088
- Finding by site, C1290906
- Integumentary System Finding, C1291044

FMA3_1

SNOMEDCT_2012_07_31
Suggestion: Concept Import

- Sign and Symptoms, C0037088
- Finding by Site or System, C1333618
  - Finding by site, C1290906
    - Integumentary System Finding, C1291044
Example: Synonym of the Other

- Chemicals, C0220806
- Chemical Viewed Structurally, C1254350
- Organic Chemicals, C0029224
- CPM2003

- Chemicals, C0220806
- Chemical categorized structurally, C0729761
- Organic Chemicals, C0029224
- SNOMEDCT_2012_07_31
Suggestion: Merge Two Synonymous Concepts

- **Chemicals, C0220806**
- **Chemical categorized structurally, C0729761**
  - **Synonym**: Chemical Viewed Structurally
- **Organic Chemicals, C0029224**
Limitations

- Harmonization cannot be done without the consent of terminology curators
- All the terminologies are in UMLS Rich Release format
Future Work

- More complex configurations: more intermediate concepts
- Algorithm to identify the relationships between intermediate concepts in complex configurations
- Pairs of any two terminologies
Conclusions

- Six reference terminologies of the UMLS vs. SNOMED
- In a sample of 241 congruency pairs
  - 143 out of 241 (59.3%) concept pairs: alternative classification.
  - 47 out of 241 (23.6%) concept pairs: parent-child relationships.
  - 35 (14.5%) new synonyms
  - Three pairs of concepts indicated errors
- Take home message:
  - A semi-automated way based on common structure of the UMLS may complement existing human-expert centered methods to find potential concepts for import and export to a terminology.
Acknowledgment

We would like to thank:

Dr. Yehoshua Perl and Dr. Chunhua Weng

for sharing their insights and giving feedback for this work
References (1)


References (2)


Thanks!

Zhe He, PhD
zh2132@columbia.edu