UMLS-INTERFACE AND UMLS-SIMILARITY:

OPEN SOURCE SOFTWARE FOR MEASURING PATHS AND SEMANTIC SIMILARITY

Bridget McInnes
Ted Pedersen
Serguei Pakhomov
**OBJECTIVE**

Develop tools to automatically compute the semantic similarity between two concepts in the biomedical domain using measures originally developed for general English using the Unified Medical Language System (UMLS)
**Motivation**

- Clustering symptoms and disorders found in the text of clinical reports for post-marking medication safety and surveillance
- Identification of patients for clinical studies
- Improving the sensitivity of document retrieval of scientific journals and clinical reports
- Development of terminologies and ontologies
- Clustering of biomedical documents
- Word sense disambiguation
Unified Medical Language System

- Knowledge representation framework
- Contains 3 Main components:
  - Metathesaurus
  - Semantic Network
  - SPECIALIST Lexicon
Metathesaurus

- Semi-automatically integrates biomedical concepts from over a 100 controlled medical terminologies

- Source vocabularies are organized based on their Atomic Unique Identifiers

- Metathesaurus is organized based on their Concept Unique Identifier (CUI)
Concept Unique Identifiers (CUIs)

- **CUI**
  - **C0009264**
  - Cold Temperature

- **AUI**
  - A15588749
  - Cold Temperature

- **AUI**
  - A3292554
  - Low Temperature

- **MSH**

- **SNOMED-CT**
CUI Information

- The concepts (AUIs) from the source vocabularies may contain information about the concept such as its
  - Definition
  - Relation information between the concepts

- The information from the AUIs can be obtained through their respective CUIs
RELATIONS BETWEEN CUIs IN MSH

AUI A15588749 Cold Temperature

is-a

AUI A0123939 Temperature

PAR/CHD (MSH)

CUI C0009264 Cold Temperature

CUI C0039476 Temperature

MSH
RELATIONS BETWEEN CUIs IN SNOMED-CT

SNOMED-CT

is-a

AUI A3292554 Low Temperature

AUI A2887140 Temperature

PAR/CHD (SNOMED-CT)

CUI C0009264 Cold Temperature

CUI C0039476 Temperature
MULTIPLE RELATIONS

PAR/CHD (SNOMED-CT)

- CUI C0009264 Cold Temperature
- CUI C0039476 Temperature

PAR/CHD (MSH)

- CUI C0009264 Cold Temperature
- CUI C0039476 Temperature
**Relation Information**

**MRHIER**

AUI A0123939 Temperature

Relation

AUI A15588749 Cold Temperature

**MRREL**

CUI C0039476 Temperature

Relation

CUI C0009264 Cold Temperature
MRREL AND MRHIER

- **MRHIER**
  - Contains the full path to root relations between AUIs from each of the sources
    - is-a
    - part-of

- **MRREL**
  - Contains the pairwise relations between CUIs
  - Relations:
    - PAR/CHD
    - RB/RN

- It is possible to generate MRHEIR from MRREL except for the following sources:
  - AIR
  - MSH
  - SNM2
  - USPMG
  - OMS
CUI VERSUS AUI HIERARCHY

The benefit of using CUIs
- Ability to obtain the relation information between concepts across sources
- Ability to obtain the relation information between concepts using more than one type of relation:
  - PAR/CHD – parent/child (relation in MRHIER)
  - RB/RN – narrower/broader
  - SIB – sibling
  - RL – concepts are similar or ‘alike’

The benefit of using AUls
- Ability to obtain relation information (PAR/CHD) between concepts in the same source very quickly
- incorporates tree positional information for sources such as MSH

UMLS-Query by Shah and Musen, 2008
UMLS-Interface

- Perl interface to the UMLS present locally in a MySQL database.

- Its main purpose is to returns path information about CUIs using the relation information in MRREL
  - All possible paths to the root
  - Shortest path between two concepts
UMLS-SIMILARITY

- A suite of perl modules that implement a number of path-based semantic similarity measures to determine the similarity between two CUIs in the UMLS
  - Measures are path-based because they rely on the location of the concepts in a hierarchy
  - The path information is obtained using UMLS-Interface

- Semantic Similarity Measures:
  - Path measure
  - Conceptual Distance (Rada, et. al, 1989)
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Nguyen and Al-Mubaid, 2006
SEMANTIC SIMILARITY EXAMPLE

Path measure

\[ \text{Sim}(c_1, c_2) = \frac{1}{N} \]

where \( N \) = \# links in the shortest path between the two concepts \( c_1 \) and \( c_2 \)
SIMILARITY GIVEN SPECIFIED SOURCES

C0005898 Body Regions

C0229962 Anatomic Part

C0015385 Limbs

PAR (SNOMED-CT)

PAR (MSH)
SIMILARITY GIVEN SPECIFIED SOURCES

Similarity = 1/1

C0005898 Body Regions

C0015385 Limbs

C0229962 Anatomic Part

PAR (SNOMED-CT)

PAR (MSH)
SIMILARITY GIVEN SPECIFIED SOURCES

\[ \text{Similarity} = \frac{1}{2} \]

- **C0005898 Body Regions**
- **C0015385 Limbs**
- **C0229962 Anatomic Part**

PAR (SNOMED-CT)
SIMILARITY GIVEN SPECIFIED SOURCES

Similarity = 1/1

C0005898
Body
Regions

PAR
(MSH)

C0015385
Limbs

C0229962
Anatomic
Part

PAR
(SNOMED-CT)

PAR
(SNOMED-CT)
SIMILARITY GIVEN SPECIFIED RELATIONS

C0005898
Body
Regions

C0015385
Limbs

C0229962
Anatomic
Part

PAR (MSH)

RB (MSH)

RB (MSH)
SIMILARITY GIVEN SPECIFIED RELATIONS

Similarity = 1/1
SIMILARITY GIVEN SPECIFIED RELATIONS

C0005898 Body Regions

C0015385 Limbs

PAR (MSH)

C0229962 Anatomic Part

RB (MSH)

Similarity = 1/2
SIMILARITY GIVEN SPECIFIED RELATIONS

Similarity = 1/1

C0005898
Body Regions

PAR (MSH)

C0015385
Limbs

C0229962
Anatomic Part

RB (MSH)
FUNCTIONAL VALIDATION

Comparison with Previous Work:
- Pedersen, et al. 2007
- Nguyen and Al-Mubaid, 2006
- Caviedes and Cimino, 2004
Semantic Similarity Measures
- Path
- Leacock and Chodorow, 1998

Source
- SNOMEDCT

Data
- 29 medical terms pairs
- Similarity determined by:
  - 9 Medical Coders
  - 3 Physicians
- 4 Point Scale
  - 4 – practically synonymous
  - 3 – related
  - 2 – marginally related
  - 1 - unrelated

Spearman’s Rank Correlation Coefficient
COMPARISON WITH PEDERSEN, ET AL.

- Semantic Similarity Measures
  - Path
  - Leacock and Chodorow, 1998

- Source: SNOMED-CT from UMLS 2008AB

- Relations: PAR/CHD

- Comparison with human annotations
  - Spearman Rank Correlation Coefficient
## Comparison with Pedersen, et al.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Physician</th>
<th>Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>Pedersen, et. al.</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.35</td>
</tr>
<tr>
<td>Leacock and Chodorow</td>
<td>Pedersen, et. al.</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.35</td>
</tr>
</tbody>
</table>
## Comparison with Pedersen, et al.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Physician</th>
<th>Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>path Pedersen, et. al.</td>
<td>0.36</td>
<td>0.51</td>
</tr>
<tr>
<td>UMLS-Similarity</td>
<td>0.35</td>
<td>0.50</td>
</tr>
<tr>
<td>Leacock and Chodorow</td>
<td>0.35</td>
<td>0.50</td>
</tr>
<tr>
<td>UMLS-Similarity</td>
<td>0.35</td>
<td>0.50</td>
</tr>
</tbody>
</table>
NGUYEN AND AL-MUBAID

- Semantic Similarity Measures
  - Nguyen and Al-Mubaid, 2006
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Path

- Source: MSH

- Same Dataset created by Pedersen, et al.
  - Data
    - 29 medical terms pairs
    - Similarity determined by:
      - 9 Medical Coders
      - 3 Physicians

- Spearman’s Rank Correlation Coefficient
COMPARISON WITH NGUYEN AND AL-MUBAID

- Semantic Similarity Measures
  - Nguyen and Al-Mubaid, 2006
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Path

- Source: MSH from UMLS 2008AB

- Relations: PAR/CHD

- Comparison with human annotations
  - Spearman Rank Correlation Coefficient
### Comparison with Nguyen and Al-Mubaid

<table>
<thead>
<tr>
<th>Measure</th>
<th>Physician</th>
<th>Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>Nguyen and Al-Mubaid</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.49</td>
</tr>
<tr>
<td>Leacock and Chodorow</td>
<td>Nguyen and Al-Mubaid</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.49</td>
</tr>
<tr>
<td>Wu and Palmer</td>
<td>Nguyen and Al-Mubaid</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.45</td>
</tr>
<tr>
<td>Nguyen and Al-Mubaid</td>
<td>Nguyen and Al-Mubaid</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>UMLS-Similarity</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Cavieges and Cimino

Semantic Similarity Measure
- Conceptual Distance – Rada, et al.

Source: MSH

Relations: PAR/CHD

Data
- 10 medical terms pairs using following CUIs
  - Digestive system disease: C0012242
  - Peptic esophagitis: C0014869
  - Psychotherapy: C0033968
  - Thirst: C0039971
  - Thoracic duct: C0039979
COMPARISON WITH CAVIEDES AND CIMINO

Semantic Similarity Measures
  • Conceptual Distance
    ○ Originally proposed by Rada, et. al., 1989

Source: MSH from UMLS 2008AB
  • Relations: PAR/CHD

Comparison between the Conceptual Distance Scores
## Comparison with Caviedes and Cimino

<table>
<thead>
<tr>
<th>CUI Pairs</th>
<th>Caviedes and Cimino</th>
<th>UMLS-Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0012242-C0014869</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C0012242-C0033968</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C0033968-C0039971</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C0012242-C0039971</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>C0012242-C0039979</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>C0033968-C0039979</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>C0014869-C0033968</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>C0014869-C0039971</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>C0014869-C0039979</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>C0039971-C0039979</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>
## Comparison with Caviedes and Cimino

<table>
<thead>
<tr>
<th>CUI Pairs</th>
<th>Caviedes and Cimino</th>
<th>UMLS-Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0012242-C0014869</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C0012242-C0033968</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C0033968-C0039971</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C0012242-C0039971</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td><strong>C0012242-C0039979</strong></td>
<td>7</td>
<td><strong>6</strong></td>
</tr>
<tr>
<td>C0033968-C0039979</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>C0014869-C0033968</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>C0014869-C0039971</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>C0014869-C0039979</strong></td>
<td>10</td>
<td><strong>11</strong></td>
</tr>
<tr>
<td>C0039971-C0039979</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>C0039971-C0039979</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>
RESULTS

The results show that UMLS-Similarity can be used to reproduce the results reported by:

- Pedersen, et al.
- Caviedes and Cimino
RESULTS

The correlation results obtained by UMLS-Similarity and reported by Nguyen and Al-Mubaid vary

- Different versions of MSH were used to conduct the experiment
- Possibly different mappings of the terms to CUIs in MSH were used
- Information used by Nguyen and Al-Mubaid comes directly from MSH which is located in MRHEIR and as PAR/CHD relations in MRREL
  - It is not possible to generate MRHIER from MRREL because the full path-to-root is a transitive closure of the pairwise PAR/CHD relations which does not hold true for MSH because a MSH concept may have different children depending on its tree position
CONCLUSIONS

- **UMLS-Similarity**
  - Used to determine the similarity between two concepts given a specified set of sources and relations
  - Contains the following similarity measures
    - Path measure
    - Conceptual Distance proposed Rada, et. al. 1989
    - Leacock and Chodorow, 1998
    - Wu and Palmer, 1994
    - Nguyen and Al-Mubaid, 2006

- **UMLS-Interface**
  - Used to obtain path information about a CUI given a specified set of sources and relations
FUTURE WORK

○ UMLS-Interface
  • Improve the efficiency in which the path information is stored

○ UMLS-Similarity
  • Information Content Similarity Measures
    ○ Resnik, 1995
    ○ Jiang and Conrath, 1997
    ○ Lin, 1997
  • Relatedness Measures
    ○ Patwardhan, 2003
**Take Home Message #1**

UMLS-Interface can be used to extract path information about a concept given a specified set of sources and relations.
UMLS-Similarity can be used to compute the semantic similarity between two concepts given a specified set of sources and relations.
Availability

- UMLS-Interface
  - http://search.cpan.org/dist/UMLS-Interface

- UMLS-Similarity
  - http://search.cpan.org/dist/UMLS-Similarity
THANK YOU

We would like to thank
- Kin Wah Fung,
- Olivier Bodenreider,
- Jan Willis
- Lan Aronson

The research was supported in parts by:
- Fellowships:
  - NLM Research Participation Program
  - GAANN fellowship from US Dept. of Ed.
- Grants
  - IR01LM009623-01A2 from NIH, NLM