Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping

Presented By: Dr. Chris J. Lu

(NIH/NLM/LHNCBC)
2017.08.24
(Health Data Science 13)

- Chris Lu (E): chlu@mail.nih.gov
Concept Mapping:
- From terms to (UMLS) concepts

Many to many mapping
- A term could have many concepts
  - WSD (Word Sense Disambiguation)
- A concept could have many terms
  - Normalization
  - Subterm substitution
  - …
### Lexical Tools – Norm [1-2]

| q0: map Unicode symbols to ASCII     | “Behçet’s Diseases, NOS”                     |
| g: remove genitives                  | "Behçet's Diseases, NOS"                    |
| rs: remove parenthetic plural forms  | "Behçet Diseases, NOS"                      |
| o: replace punctuation with spaces   | "Behçet Diseases, NOS"                      |
| t: strip stop words                  | Behçet Diseases NOS                         |
| l: lowercase                         | Behçet Diseases                             |
| B: uninflect each words in a term    | behçet diseases                             |
| Ct: retrieve citations               | behçet disease                              |
| q7: Unicode core Norm                | behcet disease                              |
| q8: strip or map non-ASCII char       | behcet disease                              |
| w: sort words by order               | behcet disease                              |
NLP – Norm (Texture Variations)

- behcet disease
- behçet disease
- behcet diseases
- behçet diseases
- behçet's disease
- behçet’s disease
- behcet's disease
- behçet's disease, nos
disease, behçet
diseases, behçet
...

Indexed Database
Normalized String

Terms in Corpus

normalize

Index

- C0004943
- Behcet Syndrome

Indexed Database
Normalized String

Terms in Corpus

normalize

Index

- behcet disease
- behçet disease
- behcet diseases
- behçet diseases
- behçet's disease
- behçet’s disease
- behçet's disease
- behçet's disease, nos
disease, behçet
diseases, behçet
...
Concept Mapping – Norm

- Behcet’s Disease
- Disease, Behcet

Indexed Database

Normalized String

Results that matches the normalized query
- C0004943
- Behcet Syndrome

Indexed Database

Normalized String

MRXNS_ENG.RRF
UMLS Metathesaurus

- **UMLS Normalized Files**
  - Normalized words: MRXNW_ENG.RRF
  - Normalized strings: MRXNS_ENG.RRF
Subterm Substitution (Synonym) [5-6]

calcaneal fractures* → heel bone fractures

Indexed Database Normalized String

None

C0006655: • calcaneal • heel bone

C0281926 Fracture of calcaneus

* PMID: 1118604, 1165396, ..
Element Synonyms

[Element Synonym]
Subterms Substitution

calcaneal fractures

heel bone fractures

Norm

bone fracture heel

Indexed Database

Normalized String

[UMLS Synonyms]
Expanded Terms for Concept Mapping: Normalization

C0006655:
- calcaneal
- heel bone
- calcaneus
- ...

[sPair: calcaneal|heel bone]

C0281926:
- **Key|calcaneus fracture**
  - fractured calcaneus
  - fracture; calcaneus
  - fracture of calcaneus
  - calcaneus fracture
  - calcaneus fractures
  - calcaneus; fracture
- **Key|bone fracture heel**
  - heel bone fracture
  - heel bone; fracture
  - fracture; heel bone
  - ...

Fracture of calcaneus

Indexed Database

Normalized String

C0281926
Fracture of calcaneus

MedInfo2017
Multiple Substitutions [7-9]

Indexed Database Normalized String

pneumonia **due to** other **virus***

Norm

due other pneumonia virus

None

pneumonia **by** other **viral**

Norm

due other pneumonia virus

other pneumonia viral

C0678226: • due to
• by

C0521026: • virus
• viral

* VA14760, HA480.80, ..
Recursive Substitutions

- chromosomal aberrance*
  - Norm
  - aberrance chromosomal
  - None

- chromosomal aberrant
  - E0006478:
    - aberrant
    - aberration
    - aberrance
    - aberrancy

- chromosomal aberration
  - Norm
  - aberration chromosomal

Indexed Database Normalized String

- C0443127

* PMID: 11172638, 25543836, ..

- C0008625
  - Chromosome Aberrations
Real-time Model [5-6]

Tokenization & NER
- Documents
- Paragraphs
- Sentences
- Phrases
- Terms
- Tokens (words)
- NER
- ...

Free Text

STMT

Norm Term

UMLS - Indexed Database
Normalized Term

CUI

Yes

No

Same LexRecord

Related LexRecords

WSD

Ranking

Subterm Substitution
(synonyms, derivations, etc.)
Pre-Processing Model

Terms in Corpus
Texture Variations
- Spelling variants
- Inflectional Variants
- Synonyms
- Derivations
...

Norm

Indexed Database Normalized String
calcaneal fracture
C0281926

Enhanced UMLS Indexed Database Normalized String
calcaneal fracture
Norm
calcaneal fracture
C0281926
Fracture of calcaneus
Synonym Sets

- UMLS Synonyms (13M)
- The SPECIALIST Lexicon Synonyms, 2016- (~5K)
- Others
  - UMLS-Core Projects (~12K)
  - Synonym set by Randy Miller, (~15K)
  - dictionary.com, thesaurus.com,
  - WordNet ([https://wordnet.princeton.edu](https://wordnet.princeton.edu))
  - etc..
Objectives

To generate a standalone set of element synonyms (sPairs) for effective UMLS concept mapping

- **Scope:**
  - include all synonymous terms in Lexicon (LexSynonyms)
  - grow with the SPECIALIST Lexicon
  - a thorough set of element synonyms (to increase recall)

- **Feature requirements:**
  - better performance: increase recall and preserve precision
  - resolve known issues (near-synonyms, POS ambiguity, include multiword synonyms, etc.)
  - cognitive synonyms (to preserve precision)
Enhanced Requirements [5-11]

- Element synonyms for subterm substitution
  - R1: Cognitive synonyms (not near-synonyms)
  - R2: POS (meaning shift)
  - R3: Source: CUI (UMLS) and other source information
  - R4: Expansions of abbreviations and acronyms
  - R5: Word level (single POS): single words and multiwords
- ...
R1: Cognitive Synonym (Quality)

- Two properties:
  - **Commutativity**: \((x = y) \rightarrow (y = x)\)
  - joy|noun|enjoy|verb \rightarrow enjoy|verb|joy|noun
  - bi-directional (sPair)
  - **Transitivity**: \(((x = y) \text{ and } (y = z)) \rightarrow (x = z)\)
  - enjoy|verb \rightarrow joy|noun \rightarrow happy|adj
  - multiple (recursive) substitutions
  - sClass (synonym class)
- Prevent precision issues by near-synonyms.
## Near-Synonyms

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>Synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0000869</td>
<td>Acacia</td>
<td>locust tree</td>
<td>Though both the acacia &amp; locust tree are members of Leguminosae (pea, bean), they do seem to refer to different trees.</td>
</tr>
<tr>
<td>C0003353</td>
<td>Antigua</td>
<td>Anguilla</td>
<td>The islands of Antigua &amp; Anguilla are both in the West Indies, but are not the same place.</td>
</tr>
<tr>
<td>C0032639</td>
<td>Pons</td>
<td>metencephalon</td>
<td>The metencephalon, per unabridged.merriam-webster.com includes the cerebellum and pons, and is different from the pons</td>
</tr>
</tbody>
</table>
Acacia & Locust Tree

- C0000869

Acacia

Locust Tree
Metencephalon & Pons

- C0032639

**Hinbrain: Metencephalon**

b) metencephalon

- **pons**
  - Contains pneumotaxic centre which fine tunes breathing rate
  - Relays information between cerebellum and cerebrum

- **cerebellum**
  - Feedback center for execution of motor movements
  - Controls posture and balance

- **reticular formation**
  - Nuclei diffusely located through the brainstem*
  - Regulates wakefulness and muscle tone

*the term “brainstem” refers to the medulla oblongata, pons, and the midbrain
### R2: POS – Meaning Shift

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>Synonym</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0004063</td>
<td>Assault</td>
<td>Mug</td>
<td>The noun mug means a large cup, while the verb mug refers to assault.</td>
</tr>
<tr>
<td>C0001774</td>
<td>Agaricales</td>
<td>Mushroom</td>
<td>The verb (to) mushroom means increase, spread, or develop rapidly. It does not refer to Agaricales while the noun is a synonym.</td>
</tr>
<tr>
<td>C0003842</td>
<td>Arteries</td>
<td>Arterial</td>
<td>The noun arterial refers to roads, not circulatory anatomy, unlike the adjective arterial.</td>
</tr>
</tbody>
</table>

**Example:**

- **mug** | **verb**  ≠  **mug** | **noun**
The patient **expired** 1 day later.

Pressure of CO2 in **expired** air ...

Disposal of **expired** drug ...

---

**CUI:** C0011065
- PT: Cessation of life
  - died
  - dead
  - death
  - deceased
  - ...

**CUI:** C0231800
- PT: Expiration, Function
  - exhaled
  - expiratory
  - expiration
  - ...

**CUI:** C1704631
- PT: Expiration
  - expire
  - expiration
  - ...

---

*MedInfo2017*
R4: Acronym/Abbreviation

- ER (27): emergency room | efficacy ratio | ejection rate | evoked response | extended release | external resistance | eye research | energy restriction | …

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Term</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0003023</td>
<td>Angola</td>
<td>ago</td>
</tr>
<tr>
<td>C0001857</td>
<td>AIDS related complex</td>
<td>arc</td>
</tr>
<tr>
<td>C3714936</td>
<td>Non-Compliant ADaM Datasets Domain</td>
<td>ax</td>
</tr>
</tbody>
</table>
Computer-aided System

**Candidate sClasses**
- UMLS sClasses
  - MRCONSO.RRF
  - English terms with same CUI

**Manual Tagging**
- Tagged by 2 linguists
  - Ensure cognitive synonyms

**Synonym Generation**
- sPairs Generating
  - Source: EUI and CUI
    - Add spelling variants
    - Add nominalization

**Refined sClasses (Filter & Matchers)**
- Must be a base form in the Lexicon
- POS: noun, verb, adjective
- Remove chemicals and drugs (STI)
- Remove acronyms or abbreviations
- Add EUI and CUI
- Remove spelling variants
- Remove nominalization

MedInfo2017
Example: sClass & Tags (POS)

#SYNONYM_CLASS|C0003842|Arteries
noun|E0010481|arteria|Y
noun|E0010531|artery|Y
noun|E0694191|arterial|N
adj|E0010482|arterial|Y
#SYNONYM_CLASS|C0004063|Assault
verb|E0041250|mug|Y
noun|E0010822|assault|Y
noun|E0041249|mug|N
...

MedInfo2017
sPairs Generation

UML-Sourced
- Retrieve synonym candidates (sClasses)
- Tag sClasses
- Generate sPairs (CUI)

Lexicon-Sourced
- Generate sPairs from nominalizations (EUI)

NLP Project-Sourced
- Generate sPairs from Lexical Tools, 2016 (NLP-LVG)

**Synonym-1** | **POS-1** | **Synonym-2** | **POS-2** | **Source**
---|---|---|---|---
mug | verb | assault | noun | C0004063
assault | noun | mug | verb | C0004063
... | ... | ... | ... | ...

MedInfo2017
Results – 2017 Release

2017 LexSynonyms

<table>
<thead>
<tr>
<th></th>
<th>Candidates</th>
<th>Tagged</th>
<th>Completion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sClass</td>
<td>22,779</td>
<td>7,686</td>
<td>33.74%</td>
</tr>
<tr>
<td>Synonyms</td>
<td>80,913</td>
<td>29,990</td>
<td>37.06%</td>
</tr>
</tbody>
</table>

Synonyms (sPairs):

<table>
<thead>
<tr>
<th>Year</th>
<th>CUI</th>
<th>EUI</th>
<th>NLP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>5,198 (100%)</td>
<td>5,198</td>
</tr>
<tr>
<td>2017</td>
<td>118,468 (62%)</td>
<td>67,584 (35%)</td>
<td>4,792 (3%)</td>
<td>190,844</td>
</tr>
</tbody>
</table>

36.71 growth

Format:

<table>
<thead>
<tr>
<th>Synonym-1</th>
<th>POS-1</th>
<th>Synonym-2</th>
<th>POS-2</th>
<th>Source</th>
</tr>
</thead>
</table>

MedInfo2017
Evaluation

- **Model:**
  - STMT (Sub-Term Mapping Tools) [6]:
    - Real-time subterm substitution features for concept mapping
    - Easy configurable options for element synonym set

- **Data:**
  - UMLS-Core Project:
    - Top 95% used terms from 8 hospitals.
    - Assigned CUI(s) to 13,076 terms
    - 2,755 terms of them do not have mapped concept through normalization in UMLS.2016AB
    - Gold Standard: 2,755 terms mapped to 2,756 CUIs
**Evaluation Model**

- **Input Terms (13,076)**
- **2,755 terms (~21% no CUI found)**

**Indexed Database Normalized String, 2016 AB**

10,321 terms (CUI found)

**STMT**

- Norm
- Subterm Substitutions

**Element Synonym Sets**

- STMT
- STMT + LexSynonym.2016
- STMT + LexSynonym.2017
- LexSynonym.2016
- LexSynonym.2017

**Results**
Gold Standard: 2,755 terms mapped to 2,756 CUIs

Element sets:

- **STMT**: a validated project specific synonym set for UMLS-Core project
- **About 75% of STMT element synonyms are duplicated in LexSynonym.2017, while only ~3% are duplicated in LexSynonym.2016.**

<table>
<thead>
<tr>
<th>Element Synonym Set</th>
<th>N. Size</th>
<th>T.P.</th>
<th>F.P.</th>
<th>F.N.</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>STMT [6]</td>
<td>7,873</td>
<td>690</td>
<td>353</td>
<td>2,066</td>
<td>66.16%</td>
<td>25.04%</td>
<td>0.3633</td>
<td>7:57</td>
</tr>
<tr>
<td>STMT + LexSynonym.2016</td>
<td>12,681</td>
<td>691</td>
<td>358</td>
<td>2,065</td>
<td>65.87%</td>
<td>25.07%</td>
<td>0.3632</td>
<td>5:31</td>
</tr>
<tr>
<td>STMT + LexSynonym.2017</td>
<td>151,913</td>
<td>828</td>
<td>424</td>
<td>1,928</td>
<td>66.13%</td>
<td>30.04%</td>
<td>0.4132</td>
<td>9:18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Element Synonym Set</th>
<th>N. Size</th>
<th>T.P.</th>
<th>F.P.</th>
<th>F.N.</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LexSynonym.2016</td>
<td>5,070</td>
<td>9</td>
<td>12</td>
<td>2,747</td>
<td>42.86%</td>
<td>0.33%</td>
<td>0.0065</td>
<td>0:16</td>
</tr>
<tr>
<td>LexSynonym.2017</td>
<td>149,912</td>
<td>287</td>
<td>117</td>
<td>2,469</td>
<td>71.04%</td>
<td>10.41%</td>
<td>0.1816</td>
<td>3:19</td>
</tr>
</tbody>
</table>
## Summary & Conclusion

<table>
<thead>
<tr>
<th>Objective &amp; Requirements</th>
<th>Check</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone element synonym set</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>All synonymous terms in the Lexicon</td>
<td>1/3 Yes</td>
<td>~ 1/3 completed</td>
</tr>
<tr>
<td>Grows with the SPECIALIST Lexicon</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Element synonyms, not expanded terms (Over-generated issues)</td>
<td>Yes</td>
<td>Must be in the Lexicon (430K, &lt; 2% of UMLS synonyms)</td>
</tr>
<tr>
<td>R1: Cognitive Synonym</td>
<td>Yes</td>
<td>Done in tagging (cognitive synonyms)</td>
</tr>
<tr>
<td>R2: Include POS</td>
<td>Yes</td>
<td>Provide POS in sClass by Lexicon</td>
</tr>
<tr>
<td>R3: Include source (CUI, EUI, etc.)</td>
<td>Yes</td>
<td>Provide source in sClass (CUI, EUI, etc.)</td>
</tr>
<tr>
<td>R4: Exclude Acronym/abbreviation</td>
<td>Yes</td>
<td>Removed in sClass by Lexicon</td>
</tr>
<tr>
<td>R5: Include Single words and multiwords</td>
<td>Yes</td>
<td>Terms in the Lexicon include both</td>
</tr>
<tr>
<td>Improve NLP performance</td>
<td>Yes</td>
<td>Improve recall and preserve precision</td>
</tr>
</tbody>
</table>
Future Work

- Complete all candidate sClasses in the future releases
- Update annually on Lexicon and Lexical Tools release with the latest Lexicon and UMLS Metathesaurus
- Include more project specific synonym set from other NLP resources (UMLS-Core, Randy Milller, etc.)
- Performance tests on NLP applications
Acknowledgements

- Supported by the Intramural Research Program of the NIH/NLM

- Co-authors (NLM):
  - Destinee Tormey
  - Dr. Lynn McCreedy
  - Allen C. Browne

- Dr. Kin Wah Fung, Dr. Marcelo Fiszman, Guy Divita, Willie Rogers, James Mork, and Francois Lang
Questions

- Chris Lu (E): chlu@mail.nih.gov