Aligning biomedical terminologies

From lexical models to supervised learning

Olivier Bodenreider, MD, PhD
Senior Scientist

École de Santé publique
MSO 6624 – Science des données de santé publique

Université de Montréal, Québec
February 17, 2021

National Library of Medicine
Lister Hill National Center for Biomedical Communications
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.
Outline

- Introduction to the UMLS Metathesaurus
- Lexical model of synonymy
- Supervised machine learning for synonymy
Introduction to the UMLS Metathesaurus
What does UMLS stand for?

- Unified
- Medical
- Language
- System

http://www.nlm.nih.gov/research/umls/
Motivation

- Started in 1986
- National Library of Medicine

«[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

• The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
• The second is the distribution of useful information among many disparate databases and systems.»
UMLS Metathesaurus

- 156 families of source vocabularies
  - Not counting 58 translations
- 25 languages
- Broad coverage of biomedicine
  - 12.1M names (normalized)
  - ~4.3M concepts
  - >10M relations
- Common presentation
UMLS Metathesaurus

Overview through an example
Addison’s disease

- Addison's disease is a rare endocrine disorder
- Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism
Adrenal insufficiency  Clinical variants

- **Primary / Secondary**
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland

- **Acute / Chronic**

- **Isolated / Polyendocrine deficiency syndrome**
Addison’s disease: Symptoms

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and non-exposed parts of the body)
- …
AD in medical vocabularies

**Synonyms: different terms**
- Addisonian syndrome
- Bronzed disease
- Melasma addisonii
- Asthenia pigmentosa
- Primary adrenal deficiency
- Primary adrenal insufficiency
- Primary adrenocortical insufficiency
- Chronic adrenocortical insufficiency

**Contexts: different hierarchies**
- eponym
- symptoms
- clinical
- variants
Organize terms

- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

<table>
<thead>
<tr>
<th>Term</th>
<th>Source</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>D000224</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>10036696</td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td>E27.1</td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Addison's disease
Metathesaurus Concepts (2020AA)

- **Concept** (4.3M) **CUI**
  - Set of synonymous concept names

- **Term** (12.1M) **LUI**
  - Set of normalized names

- **String** (13.2M) **SUI**
  - Distinct concept name

- **Atom** (15.5M) **AUI**
  - Concept name in a given source

<table>
<thead>
<tr>
<th>A0066000</th>
<th>Headache (MeSH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0065992</td>
<td>Headache (ICD-10)</td>
</tr>
<tr>
<td>S0046854</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A0066007</th>
<th>Headaches (MedDRA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A12003304</td>
<td>Headaches (OMIM)</td>
</tr>
<tr>
<td>S0046855</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L0018681</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>A0540936</th>
<th>Cephalodynia (MeSH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0475647</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L0380797</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>C0018681</th>
</tr>
</thead>
</table>
Integrating subdomains

Clinical repositories

Genetic knowledge bases

SNOMED CT

OMIM

MeSH

Biomedical literature

NCBI Taxonomy

UMLS

GO

Genome annotations

FMA

Model organisms

Anatomy

Other subdomains

...
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains

NILH
National Library of Medicine
Lister Hill National Center for Biomedical Communications
Trans-namespace integration

Addison's disease (36372003)

Clinical repositories

Other subdomains

Snomed CT

OMIM

Genetic knowledge bases

Biomedical literature

Addison Disease (D000224)

NCBI Taxonomy

Model organisms

FMA

GO

Genome annotations

Anatomy

UMLS C0001403
Lexical model of synonymy
From lexical features to synonymy

- Adrenal gland diseases
- Adrenal disorder
- Disorder of adrenal gland
- Diseases of the adrenal glands
- C0001621
Lexical resources

SPECIALIST Lexicon and lexical tools

SPECIALIST Lexicon

◆ Content
  ● English lexicon
  ● Many words from the biomedical domain

◆ Over 500,000 lexical items

◆ Word properties
  ● morphology
  ● orthography
  ● syntax

◆ Used by the lexical tools
Morphology

◆ Inflection
  - noun
  - verb
  - adjective

◆ Derivation
  - verb ↔ noun
ganterize ↔ cauterization
  - adjective ↔ noun
gred ↔ redness
Orthography

◆ Spelling variants

- oe/e
  - oesophagus - esophagus
- ae/e
  - anaemia - anemia
- ise/ize
  - cauterise - cauterize
- genitive mark
  - Addison's disease
  - Addison disease
  - Addisons disease
Syntax

◆ Complementation
  ● verbs
    ◆ intransitive
      I'll treat.
    ◆ transitive
      He treated the patient.
    ◆ ditransitive
      He treated the patient with a drug.
  ● nouns
    ◆ prepositional phrase
      Valve of coronary sinus

◆ Position for adjectives
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>hemoglobin</td>
<td>(base form)</td>
</tr>
<tr>
<td>spelling_variant</td>
<td>haemoglobin</td>
<td></td>
</tr>
<tr>
<td>entry</td>
<td>E0031208</td>
<td>(identifier)</td>
</tr>
<tr>
<td>cat</td>
<td>noun</td>
<td>(part of speech)</td>
</tr>
<tr>
<td>variants</td>
<td>uncount</td>
<td>(no plural)</td>
</tr>
<tr>
<td>variants</td>
<td>reg</td>
<td>(plural: hemoglobins, hemoglobins)</td>
</tr>
</tbody>
</table>
Lexical tools

- To manage lexical variation in biomedical terminologies

- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)

- Based on the SPECIALIST Lexicon

- Used by noun phrase extractors, search engines
Normalization

- Remove genitive: Hodgkin’s diseases, NOS
- Remove stop words: Hodgkin diseases, NOS
- Lowercase: Hodgkin diseases,
- Strip punctuation: hodgkin diseases,
- Uninflect: hodgkin diseases
- Sort words: disease hodgkin
Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin
Normalization Applications

- Model for lexical resemblance
- Help find lexical variants for a term
  - Terms that normalize the same usually share the same LUI
- Help find candidates to synonymy among terms
- Help map input terms to UMLS concepts
Indexes

◆ Word index
  - word to Metathesaurus strings
  - one word index per language

◆ Normalized word index
  - normalized word to Metathesaurus strings
  - English only

◆ Normalized string index
  - normalized term to Metathesaurus strings
  - English only
Lexical Variant Generation program

◆ Tool for specialists (linguists)
◆ Performs atomic lexical transformations
  ○ generating inflectional variants
  ○ lowercase
  ○ ...
◆ Performs sequences of atomic transformations
  ○ a specialized sequence of transformations provides the normalized form of a term (the *norm* program)
Principles for asserting synonymy in the UMLS Metathesaurus

- Lexical similarity is used to identify candidates for synonymy
  - Atoms that do not share a common semantics are prevented from being recognized as synonymous and grouped into the same concept

- Synonymy asserted between atoms in a source vocabulary tends to be conserved in the Metathesaurus
### Example

<table>
<thead>
<tr>
<th>String</th>
<th>Source</th>
<th>SCUI</th>
<th>AUI</th>
<th>LUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headache</td>
<td>MSH</td>
<td>M0009824</td>
<td>A0066000</td>
<td>L0018681</td>
</tr>
<tr>
<td>Headaches</td>
<td>MSH</td>
<td>M0009824</td>
<td>A0066008</td>
<td>L0018681</td>
</tr>
<tr>
<td>Cranial Pains</td>
<td>MSH</td>
<td>M0009824</td>
<td>A1641924</td>
<td>L1406212</td>
</tr>
<tr>
<td>Cephalodynia</td>
<td>MSH</td>
<td>M0009824</td>
<td>A26628141</td>
<td>L0380797</td>
</tr>
<tr>
<td>Cephalodynia</td>
<td>SNOMEDCT_US</td>
<td>25064002</td>
<td>A2957278</td>
<td>L0380797</td>
</tr>
<tr>
<td>Headache (finding)</td>
<td>SNOMEDCT_US</td>
<td>25064002</td>
<td>A3487586</td>
<td>L3063036</td>
</tr>
</tbody>
</table>
Metathesaurus building process

- All terms from source vocabularies are processed
  - Terms that have the same normalized form are candidates for synonymy
    - Unless they bear different semantics
  - Synonymy indicated by source vocabularies tends to be preserved

- All candidates (from normalization or sources) are reviewed manually

- Synonyms are assigned the same CUI
Supervised machine learning for synonymy
Intuition

- Large collection of synonymy assertions in Metathesaurus can be used for supervised learning
  - Positive examples: terms from the same concept
  - Negative examples: terms from different concepts

- Possible features
  - Lexical (words in a term)
  - Semantic (semantics of the source)
  - Relations to other terms
Synonymy function

Addison Disease
Primary hypoadrenalism
Primary adrenocortical insufficiency
Addison's disease (disorder)

Hodgkin Disease
Granuloma, Malignant
Hodgkin lymphoma
Malignant lymphoma, Hodgkin’s

\[\text{syn}(\text{“Addison Disease”}, \text{“Primary hypoadrenalism”}) = 1\]
\[\text{syn}(\text{“Addison Disease”}, \text{“Hodgkin Disease”}) = 0\]
Neural network architecture

- **Word embeddings**
  - Word vectors for representing terms
  - Using BioWordVec
- **Siamese LSTM network**
- **Similarity function = Manhattan distance**
Neural network architecture

Similarity between 0...1

Manhattan Distance

Dense (50) = Dense (50)
Dense (128) = Dense (128)
Bi-LSTM (50) = Bi-LSTM (50)
BioWordVec Embeddings = BioWordVec Embeddings

Convert into input ids
Truncate/Pad to 30 tokens
Tokenize

Lung disease and disorder
Head disease and disorder
Learning experiments

◆ Hypotheses

  ● More difficult to predict synonymy among lexically different terms than lexically similar terms

  ● More difficult to predict non-synonymy among lexically similar terms than among lexically different terms

◆ Experiments

  ● Different degrees of lexical similarity among negative examples used for learning
## Datasets

<table>
<thead>
<tr>
<th>Type</th>
<th>Positive</th>
<th>Negative</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>High similarity</td>
<td>22,324,834</td>
<td>55,909,551</td>
<td>78,234,385</td>
</tr>
<tr>
<td>Low similarity</td>
<td>22,324,834</td>
<td>55,909,551</td>
<td>78,234,385</td>
</tr>
<tr>
<td>No similarity</td>
<td>22,324,834</td>
<td>58,256,526</td>
<td>80,581,360</td>
</tr>
<tr>
<td>High+Low+No</td>
<td>22,324,834</td>
<td>170,075,628</td>
<td>192,400,462</td>
</tr>
</tbody>
</table>

**Positive (selected pairwise within concepts)**

(“Addison Disease”, “Primary hypoadrenalism”)

**Negative (selected pairwise between concepts)**

- high sim: (“Addison Disease”, “Hodgkin Disease”)
- low sim: (“Fracture of left rib”, “Traumatic hematoma of left kidney”)
- no sim: (“Addison Disease”, “Hodgkin lymphoma”)

---

**National Library of Medicine**

Lister Hill National Center for Biomedical Communications
All models show good performance

<table>
<thead>
<tr>
<th>Type</th>
<th>F1 Training</th>
<th>F1 Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High similarity</td>
<td>0.9521</td>
<td>0.9333</td>
</tr>
<tr>
<td>Low similarity</td>
<td>0.9887</td>
<td>0.9784</td>
</tr>
<tr>
<td>No similarity</td>
<td><strong>0.9958</strong></td>
<td><strong>0.9899</strong></td>
</tr>
<tr>
<td>High+Low+No</td>
<td>0.9480</td>
<td>0.9287</td>
</tr>
</tbody>
</table>

*Good performance against unseen data from the same dataset*
Some models generalize poorly

<table>
<thead>
<tr>
<th>Model used for testing</th>
<th>F1 High</th>
<th>F1 Low</th>
<th>F1 No</th>
<th>F1 H+L+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>High similarity</td>
<td>0.8740</td>
<td>0.9117</td>
<td>0.9217</td>
<td>0.7954</td>
</tr>
<tr>
<td>Low similarity</td>
<td>0.5678</td>
<td>0.9654</td>
<td>0.9768</td>
<td>0.5572</td>
</tr>
<tr>
<td>No similarity</td>
<td>0.3593</td>
<td>0.7943</td>
<td>0.9816</td>
<td>0.3286</td>
</tr>
<tr>
<td>High+Low+No</td>
<td>0.8974</td>
<td>0.9469</td>
<td>0.9549</td>
<td>0.9061</td>
</tr>
</tbody>
</table>

Models not trained on high lexical similarity negative examples do not generalize well
### Deep learning vs. normalization and source synonymy

<table>
<thead>
<tr>
<th>Type</th>
<th>F1 High</th>
<th>F1 Low</th>
<th>F1 No</th>
<th>F1 H+L+N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning (High+Low+No)</td>
<td>0.8974</td>
<td>0.9469</td>
<td>0.9549</td>
<td>0.9061</td>
</tr>
<tr>
<td>Normalization+Source synonymy</td>
<td>0.7672</td>
<td>0.8109</td>
<td>0.8145</td>
<td>0.7651</td>
</tr>
</tbody>
</table>

*Deep learning model largely outperforms normalization+source synonymy*
Discussion

◆ Encouraging results
  ● Outperforms Normalization+Source synonymy
◆ Inclusion of lexically similar terms among negative examples is key to performance
◆ Areas for improvement
  ● More sophisticated embeddings (e.g., BERT)
  ● Integration of context (source synonymy, relations)
◆ Applications
  ● Integration of new terminology into Metathesaurus
Summary
Summary

- The UMLS Metathesaurus is a biomedical terminology integration system
- Metathesaurus construction has relied on a lexical model for synonymy and human review
- Supervised machine learning approaches to predicting synonymy have shown promising results
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider

National Library of Medicine
Lister Hill National Center for Biomedical Communications
References

◆ UMLS overview


◆ Supervised learning approach