Aligning biomedical terminologies

From lexical models to supervised learning

Olivier Bodenreider, MD, PhD

Senior Scientist

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.
Fundamental theorem of clinical data interoperability

\[ \text{I} = \text{T} + \text{M} \]

Interoperability

Terminologies
- Vocabularies
- Code sets
- Terminology Standards

Information models
- Content structure standards
- Data models
Outline

- Introduction to the UMLS Metathesaurus
- Lexical model of synonymy
- Supervised machine learning for synonymy prediction
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

- 166 families of source vocabularies
  - Not counting translations
- 27 languages
- Broad coverage of biomedicine
  - 11.8M names (normalized)
  - ~3.3M concepts
  - >10M relations
- Common presentation
UMLs Metathesaurus Example

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

<table>
<thead>
<tr>
<th>Term</th>
<th>MeSH</th>
<th>D000224</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td></td>
<td></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>
</tbody>
</table>

C0001403

Addison's disease
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- SNOMED CT
- OMIM
- MeSH
- Biomedical literature
- NCBI Taxonomy
- FMA
- GO
- Genome annotations
- Anatomy
- Other subdomains
- Model organisms

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

Clinical repositories

Genetic knowledge bases

Biomedical literature

Addison's disease (363732003)

Other subdomains

SNOMED CT

OMIM

NCBI Taxonomy

UMLS C0001403

FMA

GO

Anatomy

Genome annotations

Model organisms

Addison Disease (D000224)
Lexical model of synonymy
From lexical features to synonymy

Adrenal gland diseases
Adrenal disorder
Disorder of adrenal gland
Diseases of the adrenal glands
C0001621
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**
  - nucleus, nuclei

- **verb**
  - cauterize, cauterizes, cauterized, cauterizing

- **adjective**
  - red, redder, reddest

### Derivation

- **verb ↔ noun**
  - cauterize ↔ cauterization

- **adjective ↔ noun**
  - red ↔ 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
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

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

Remove genitive

Hodgkin’s diseases, NOS

Hodgkin diseases, NOS

Hodgkin diseases,

hodgkin diseases,

hodgkin diseases

hodgkin disease

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

normalize  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
Metathesaurus building process

❖ All terms from source vocabularies are processed
  ● Terms that have the same normalized for 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
  ● Labor-intensive and error-prone

❖ Synonyms are assigned the same CUI
### 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>
Supervised machine learning for synonymy prediction
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 

\[ \text{syn}(“Addison Disease”, “Primary hypoadrenalism”) = 1 \]
\[ \text{syn}(“Addison Disease”, “Hodgkin Disease”) = 0 \]
Early experiments Pairwise similarity

◆ Types of embeddings
  ● Word vectors for representing terms using BioWordVec (2021)
  ● Knowledge Graph Embeddings for representing the context (2022)
◆ Siamese LSTM network

◆ Results: Best model
  ● F1=0.765 (baseline: lexical similarity + source synonymy)
  ● F1=0.906 (words)
  ● F1=0.935 (context)
Recent experiments Vocabulary insertion

- Initial approach does not translate well to vocabulary insertion (inserting new terms into the Metathesaurus)

- Rethinking the approach as an entity linking problem
  - Given a new term, find the concept with which it should be associated
  - Or indicate if there is no such concept
Recent experiments  Vocabulary insertion

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Based Approximation (RBA)</td>
<td>70.1</td>
</tr>
<tr>
<td>LexLM</td>
<td>63.2</td>
</tr>
<tr>
<td>PubMedBERT</td>
<td>68.4</td>
</tr>
<tr>
<td>SapBERT</td>
<td>77.4</td>
</tr>
<tr>
<td>RBA + LexLM</td>
<td>80.4</td>
</tr>
<tr>
<td>RBA + PubMedBERT</td>
<td>83.7</td>
</tr>
<tr>
<td>RBA + SapBERT</td>
<td>90.7</td>
</tr>
<tr>
<td>Re-Ranker (PubMedBERT)</td>
<td>85.5</td>
</tr>
<tr>
<td>+ RBA Signal</td>
<td>93.2</td>
</tr>
</tbody>
</table>

- Lexical similarity + source synonymy
- Existing models
- Existing models enriched with Lexical similarity + source synonymy
- New models (re-ranking)
Discussion

🔹 Performance conserved
  - Across versions (UMLS insertion sets)
  - Across categories (UMLS semantic groups)

🔹 Importance of extending entity linking with “null injection”

🔹 The deep learning models improve when augmented with basic information (lexical similarity and source synonymy)
Overall 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
References

◆ UMLS overview

◆ Lexical approach

◆ Supervised learning approach
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