Aix-Marseille Université
Sciences économiques et sociales de la santé
& traitement de l'information médicale

January 20, 2023

Aligning biomedical terminologies

From lexical models to supervised learning

Olivier Bodenreider, MD, PhD

Senior Scientist

NIH 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 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 (2021AA)

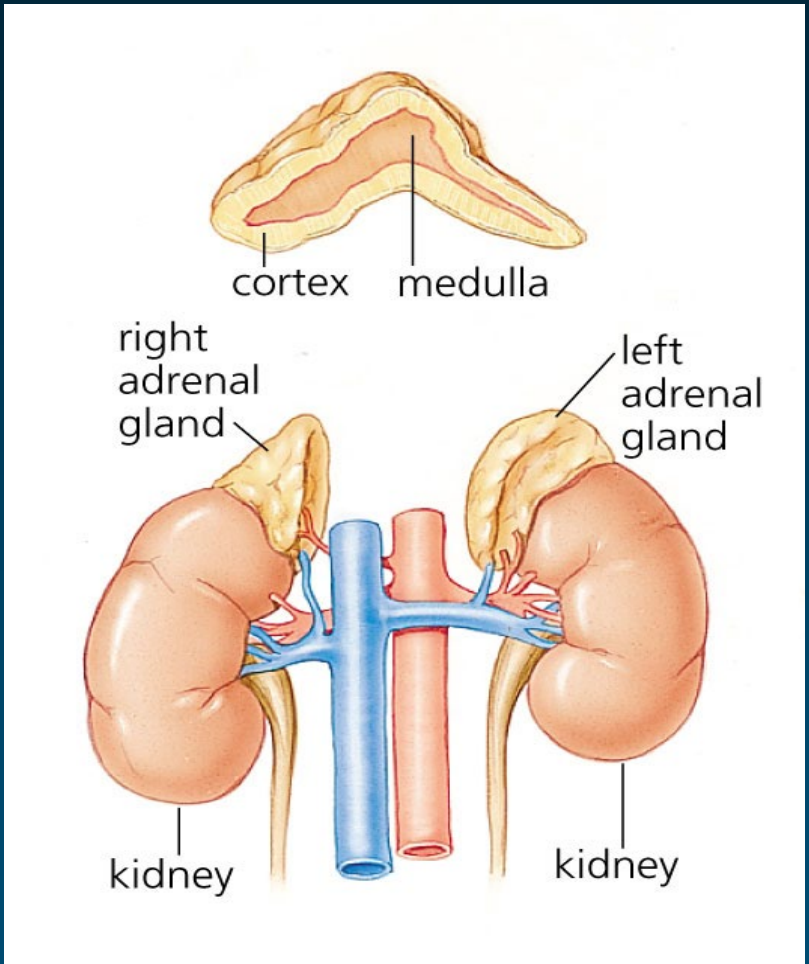
- 157 families of source vocabularies
  - Not counting 58 translations
- 25 languages
- Broad coverage of biomedicine
  - 12.5M names (normalized)
  - ~4.4M concepts
  - >10M relations
- Common presentation

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



AD in medical vocabularies

- Synonyms: different terms
  - Addisonian syndrome — eponym
  - Bronzed disease — symptoms
  - Melasma addisonii — symptoms
  - Asthenia pigmentosa — symptoms
  - Primary adrenal deficiency — clinical variants
  - Primary adrenal insufficiency — clinical variants
  - Primary adrenocortical insufficiency — clinical variants
  - Chronic adrenocortical insufficiency — clinical variants
- Contexts: different hierarchies

Organize terms

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

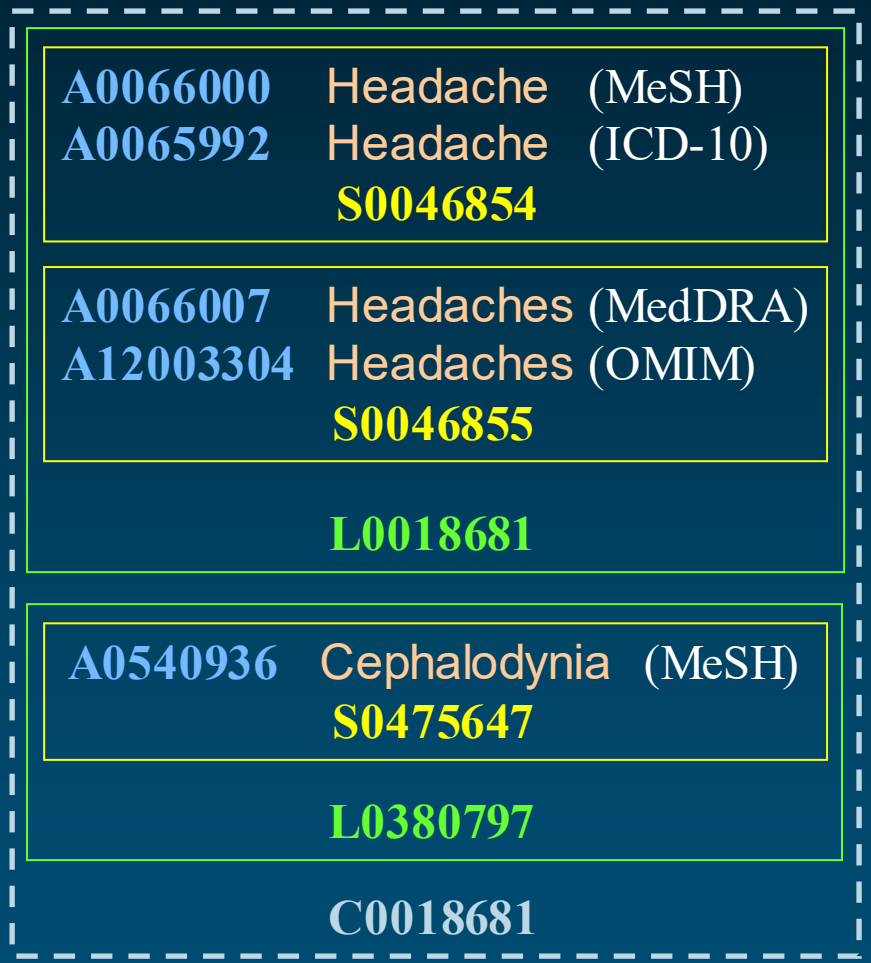
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

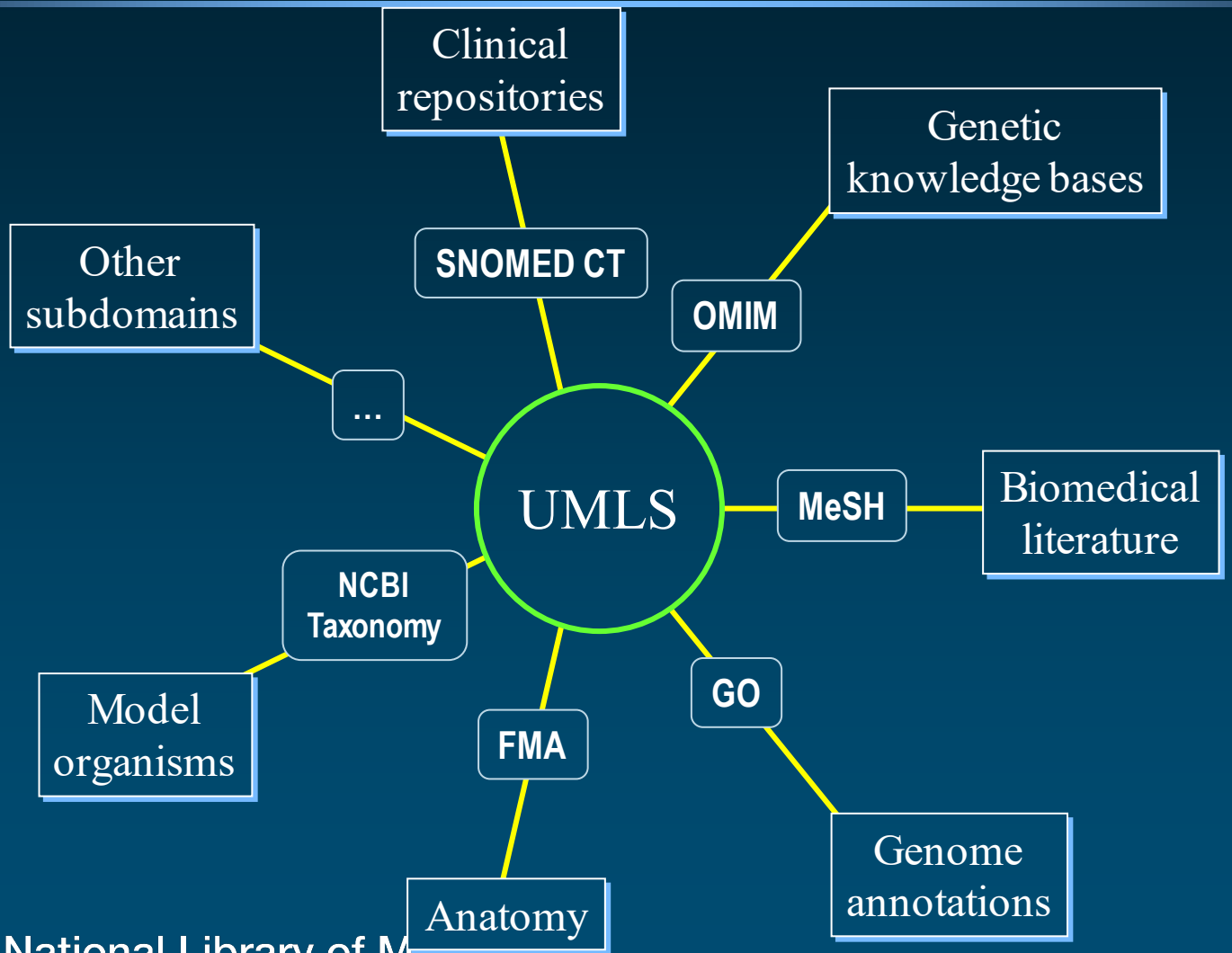
Addison's disease

Metathesaurus Concepts (2020AA)

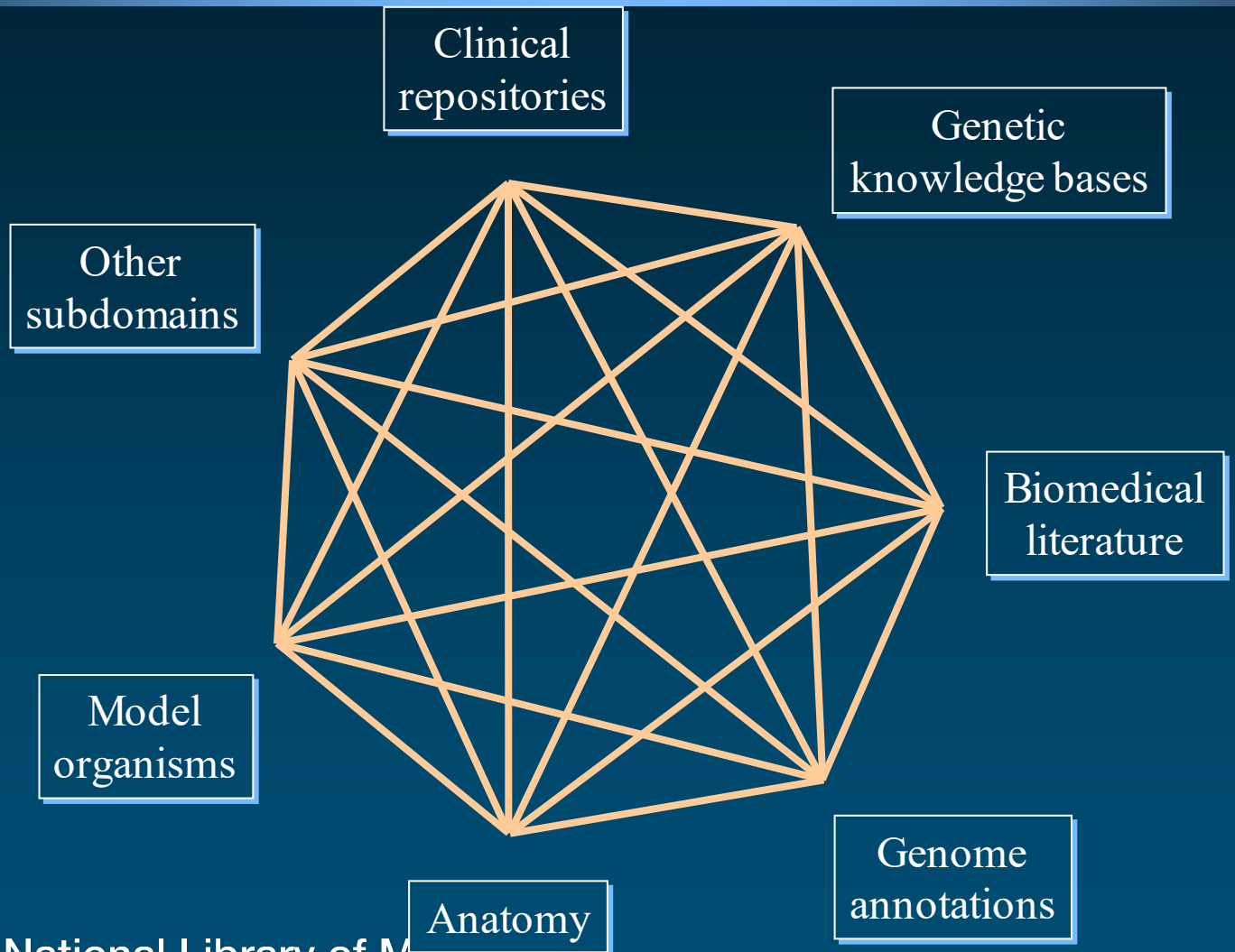
- Concept (4.3M) CUI
  - Set of synonymous concept names
- Term (12.1M) LUI
  - Set of normalized names
- String (13.2M) SUI
  - Distinct concept name
- Atom (15.5M) AUI
  - Concept name in a given source



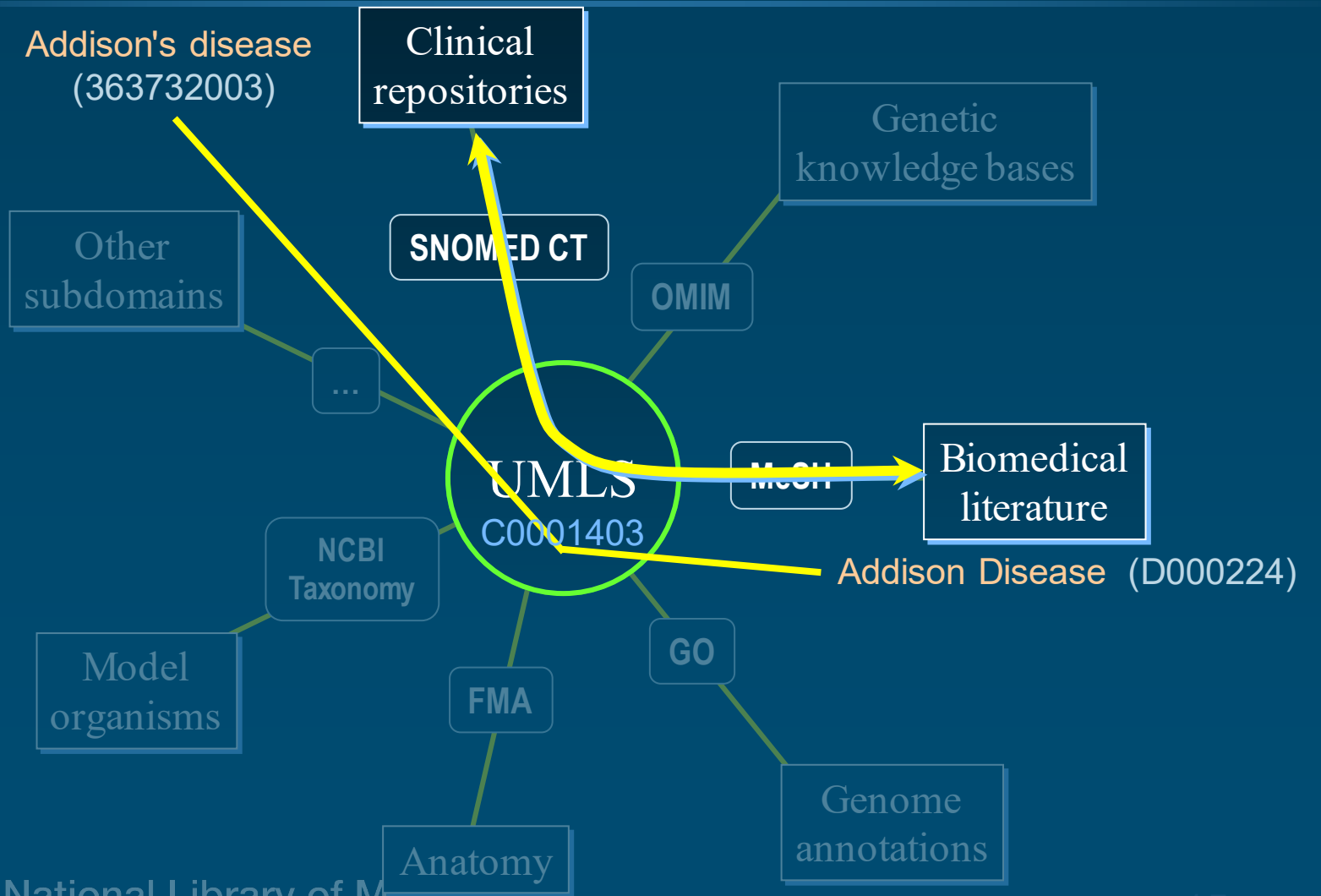
Integrating subdomains



Integrating subdomains

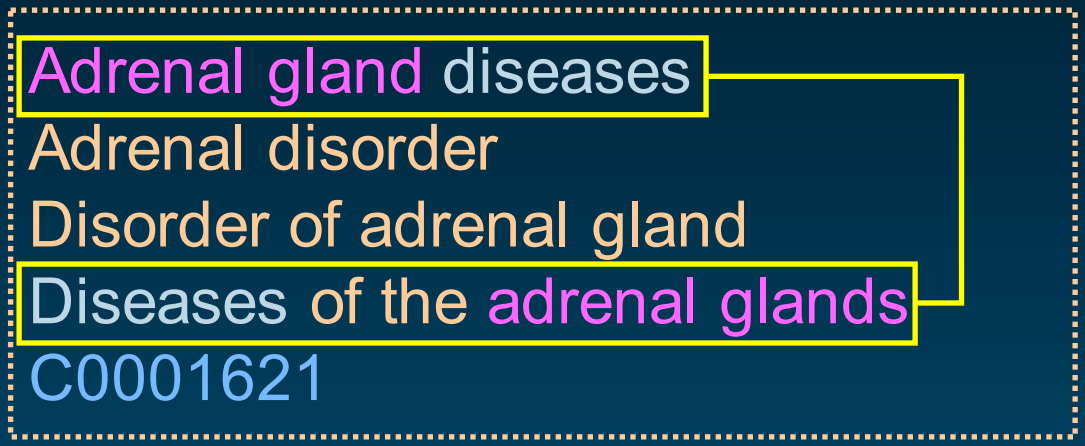


Trans-namespace integration



Lexical model of synonymy

From lexical features to synonymy



Lexical resources

SPECIALIST Lexicon and lexical tools

https://lhncbc.nlm.nih.gov/LSG/index.html

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

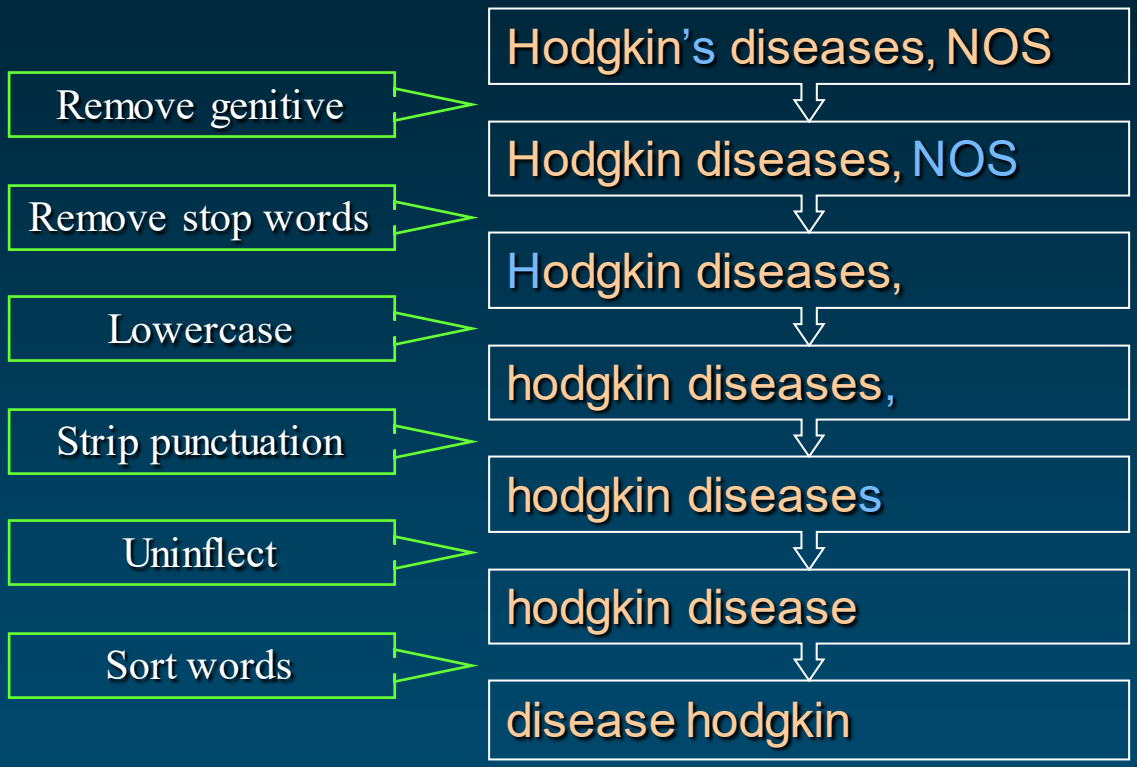
SPECIALIST Lexicon record

```
{
   base=hemoglobin      (base form)
   spelling_variant=haemoglobin
   entry=E0031208       (identifier)
   cat=noun             (part of speech)
   variants=uncount     (no plural)
   variants=reg         (plural: hemoglobins)
}
```

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



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

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

String	Source	SCUI	AUI	LUI
Headache	MSH	M0009824	A0066000	L0018681
Headaches	MSH	M0009824	A0066008	L0018681
Cranial Pains	MSH	M0009824	A1641924	L1406212
Cephalodynia	MSH	M0009824	A26628141	L0380797
Cephalodynia	SNOMEDCT_US	25064002	A2957278	L0380797
Headache (finding)	SNOMEDCT_US	25064002	A3487586	L3063036

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
- Synonyms are assigned the same CUI
- Labor-intensive and error-prone

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)
[…]

C0001403

Hodgkin Disease
Granuloma, Malignant
Hodgkin lymphoma
Malignant lymphoma, Hodgkin's
[…]

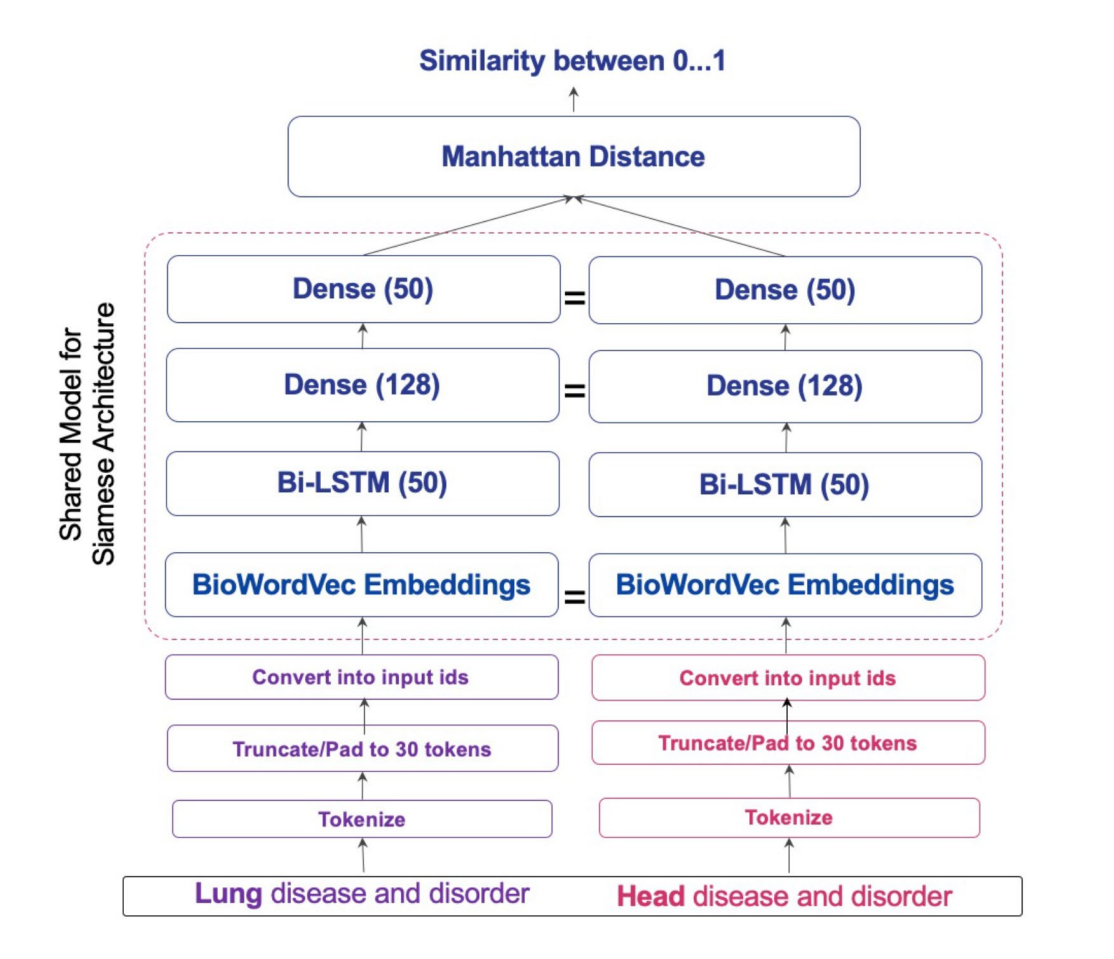
C0019829

syn("Addison Disease", "Primary hypoadrenalism") = 1
syn("Addison Disease", "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



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

Type	Positive	Negative	All
High similarity	22,324,834	55,909,551	78,234,385
Low similarity	22,324,834	55,909,551	78,234,385
No similarity	22,324,834	58,256,526	80,581,360
High+Low+No	22,324,834	170,075,628	192,400,462

*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")*

All models show good performance

Type	F1 Training	F1 Validation
High similarity	0.9521	0.9333
Low similarity	0.9887	0.9784
No similarity	**0.9958**	**0.9899**
High+Low+No	0.9480	0.9287

*Good performance against unseen data **from the same dataset***

Some models generalize poorly

Model used for testing

Model used for testing	Type	F1 High	F1 Low	F1 No	F1 H+L+N
	High similarity	0.8740	0.9117	0.9217	0.7954
	Low similarity	0.5678	**0.9654**	0.9768	0.5572
	No similarity	0.3593	0.7943	**0.9816**	0.3286
	High+Low+No	**0.8974**	0.9469	0.9549	**0.9061**

Models not trained on high lexical similarity negative examples do not generalize well

Deep learning vs. normalization and source synonymy

Type	F1 High	F1 Low	F1 No	F1 H+L+N
Deep learning (High+Low+No)	**0.8974**	**0.9469**	**0.9549**	**0.9061**
Normalization+ Source synonymy	0.7672	0.8109	0.8145	0.7651

Deep learning model largely outperforms normalization+source synonymy

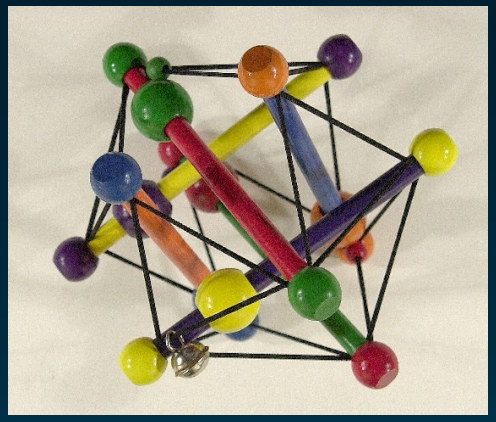
Discussion

- 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
  - Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270.
- Supervised learning approach
  - Nguyen V, Yip HY and Bodenreider O. Biomedical vocabulary alignment at scale in the UMLS Metathesaurus. *Proceedings of the Web Conference 2021 (WWW'21)*; 2672-2683.