Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping

Presented By: Dr. Chris J. Lu

(NIH/NLM/LHNCBC)

2017.08.24

(Health Data Science 13)

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
- Chris Lu (E): chlu@mail.nih.gov

Table of Contents

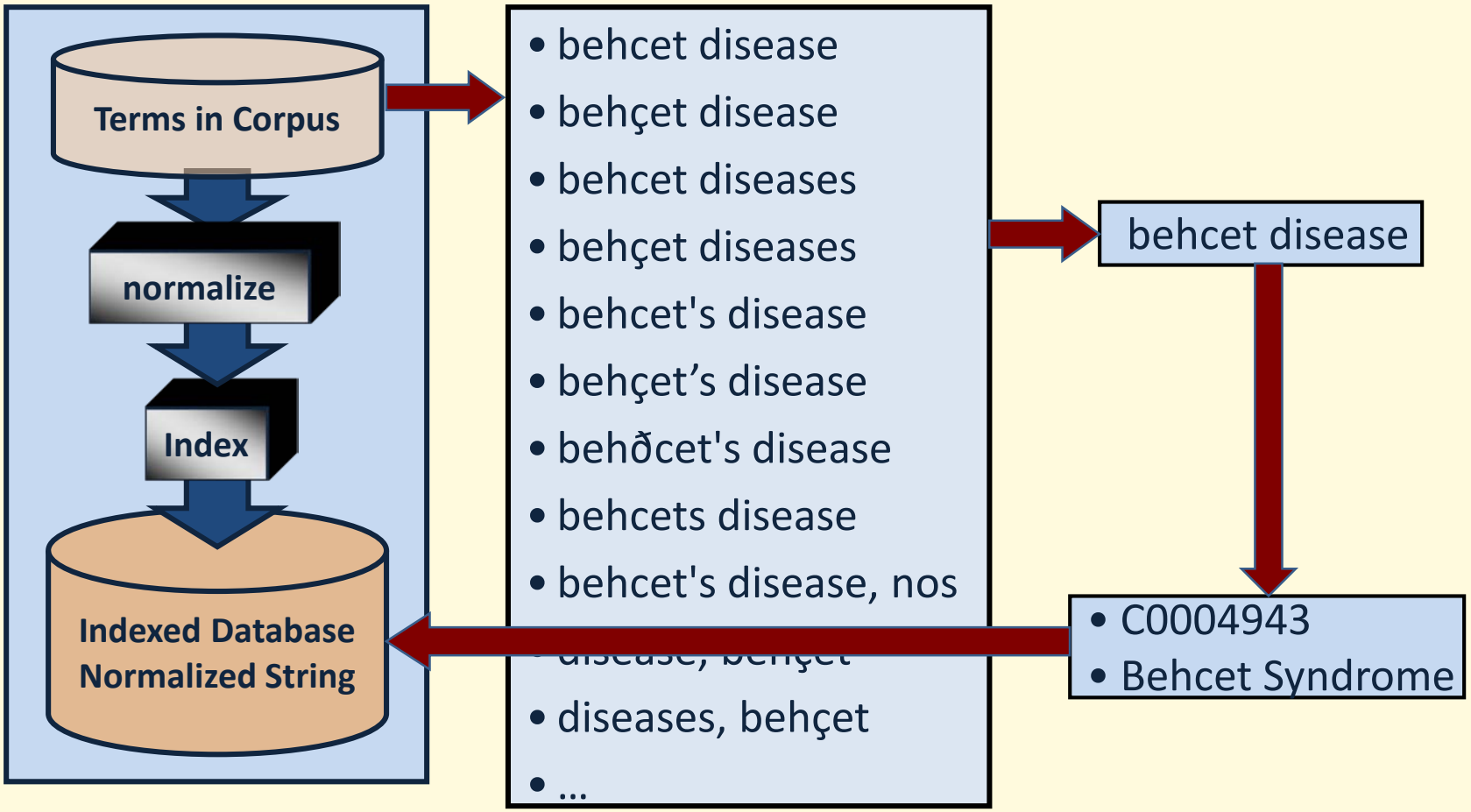
Introduction

- 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
    - …

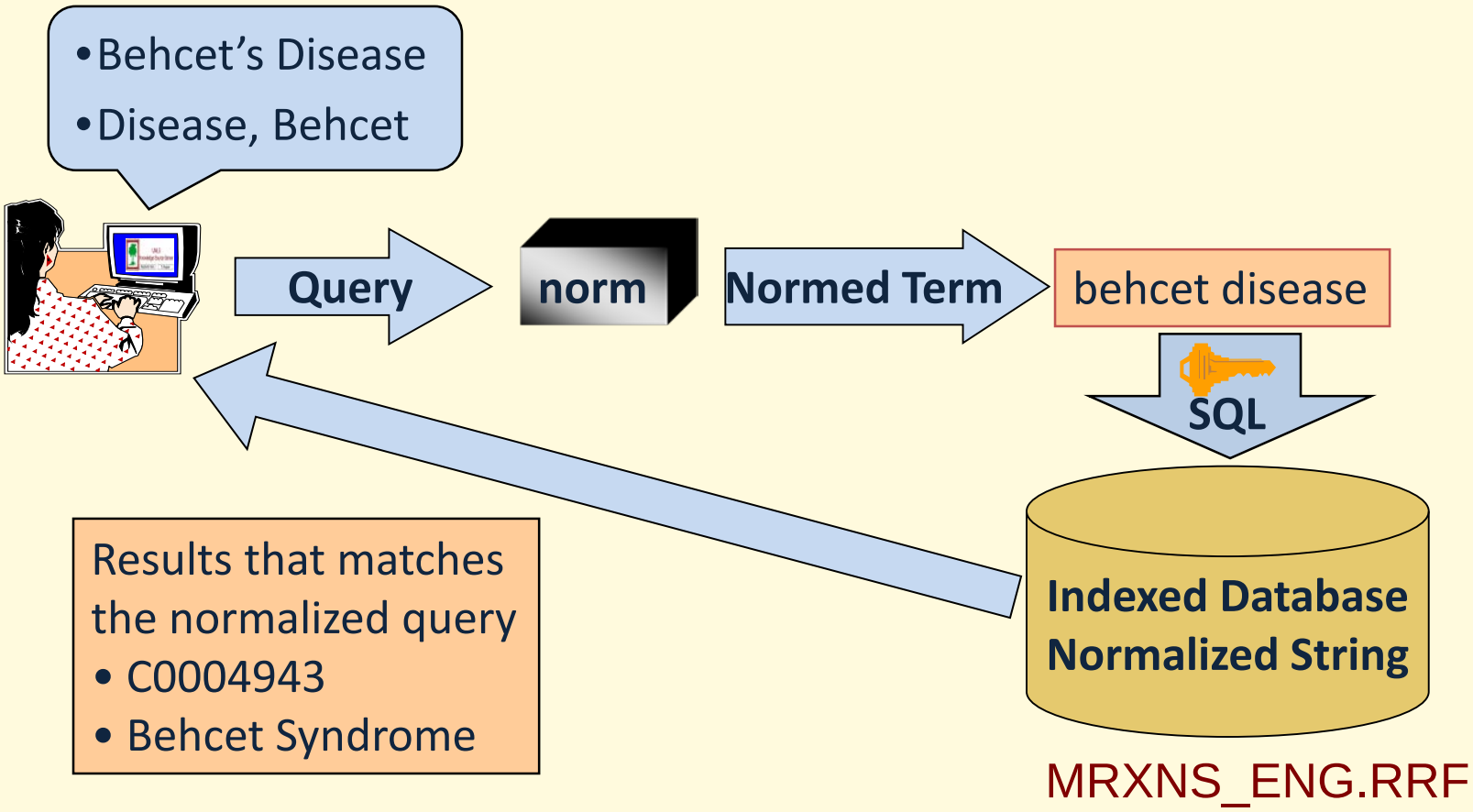
Lexical Tools – Norm [1-2]

	“Behçet’s Diseases, NOS“
q0: map Unicode symbols to ASCII	"Behçet's Diseases, NOS"
g: remove genitives	"Behçet Diseases, NOS"
rs: remove parenthetic plural forms	"Behçet Diseases, NOS"
o: replace punctuation with spaces	Behçet Diseases NOS
t: strip stop words	Behçet Diseases
l: lowercase	behçet diseases
B: uninflect each words in a term	behçet disease
Ct: retrieve citations	behcet 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)

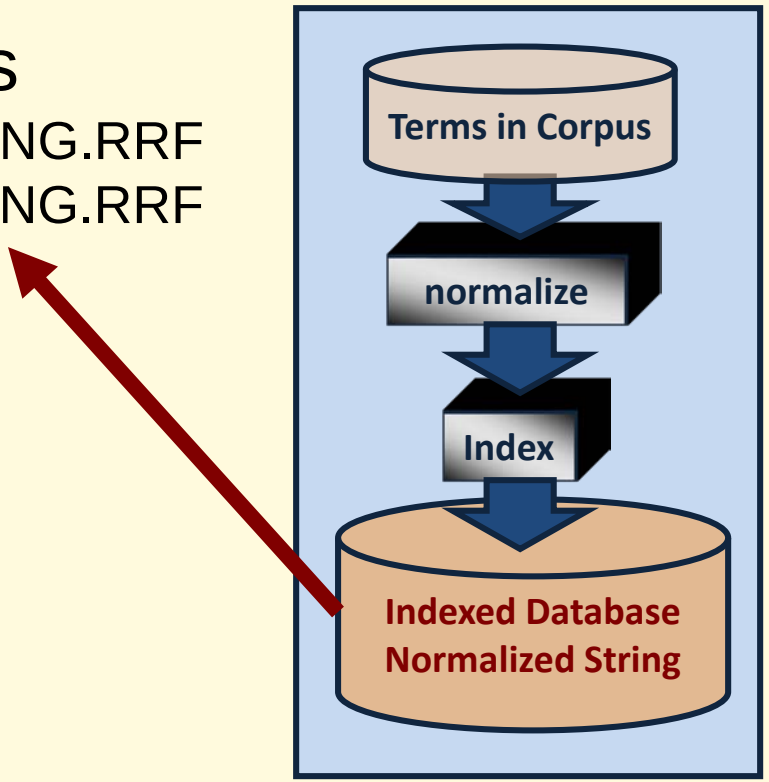


Concept Mapping – Norm

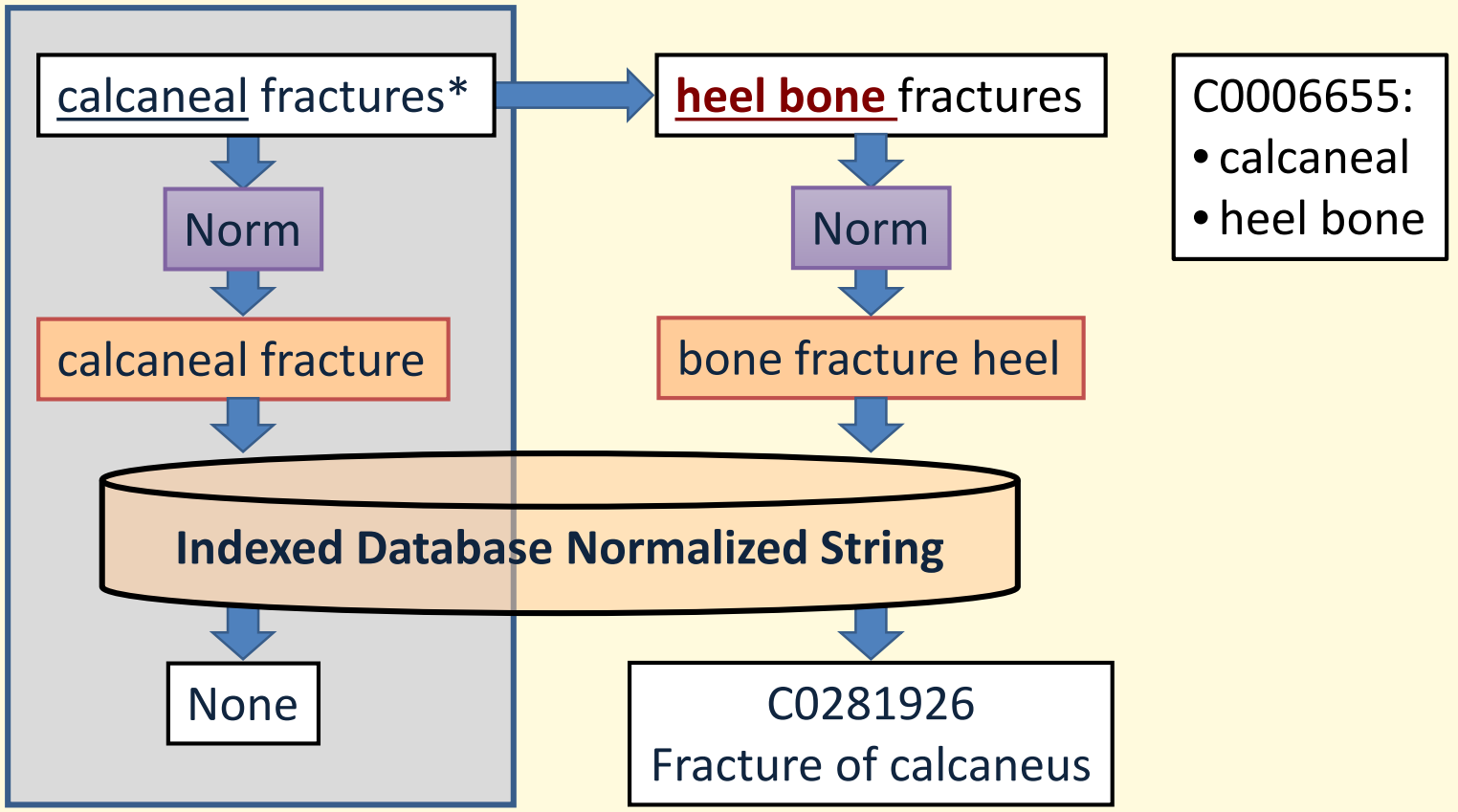


UMLS Metathesaurus

- UMLS Normalized Files
  - Normalized words: MRXNW_ENG.RRF
  - Normalized strings: MRXNS_ENG.RRF

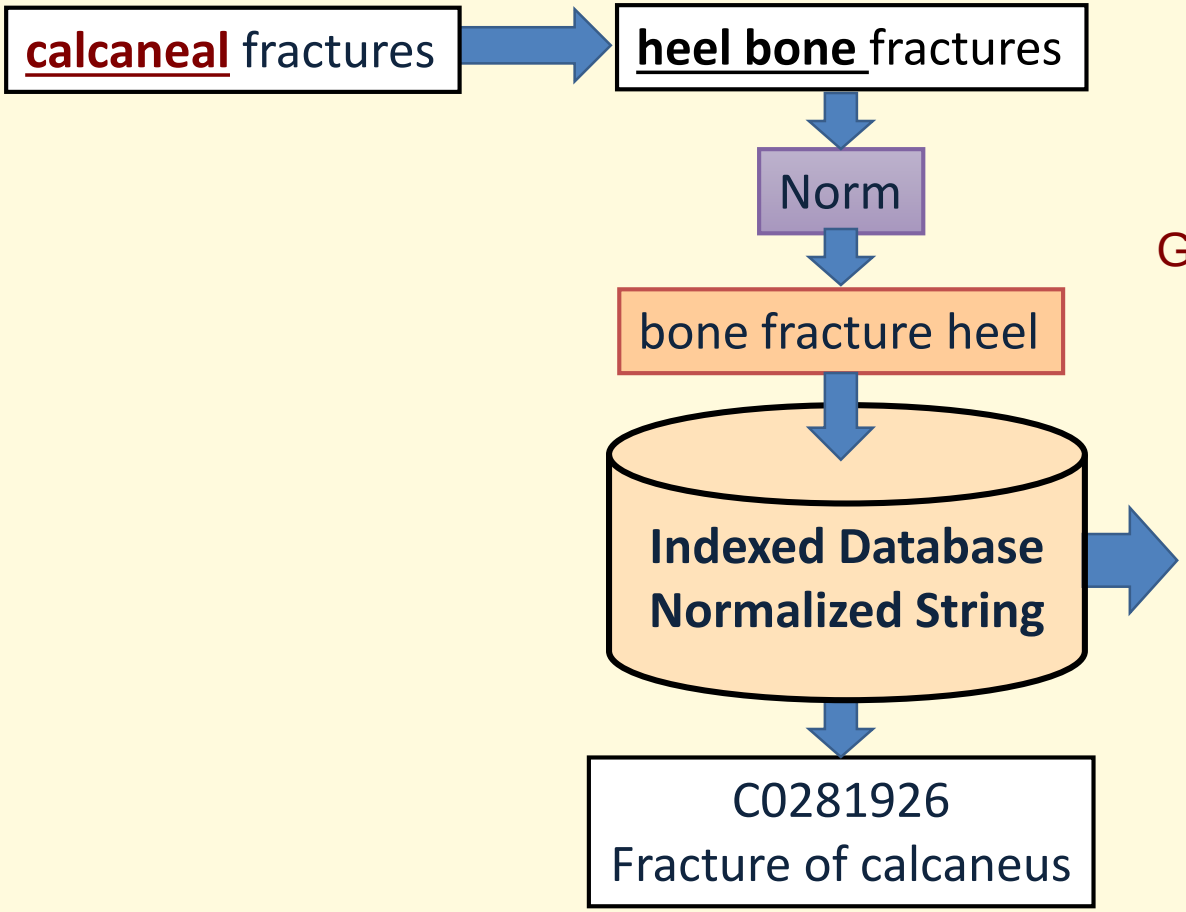


Subterm Substitution (Synonym) [5-6]



* PMID: 1118604, 1165396, ..

UMLS Synonyms

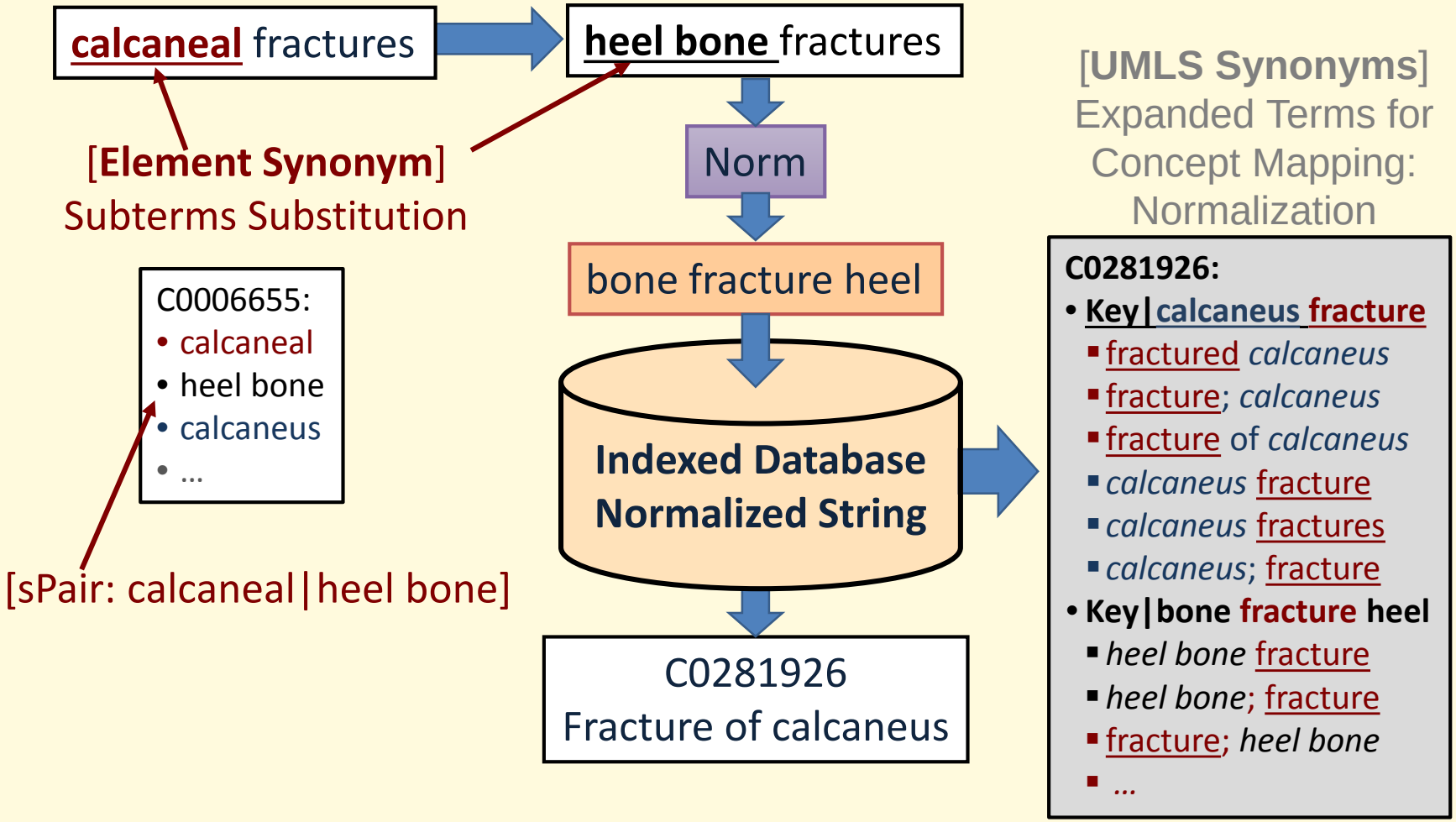


[**UMLS Synonyms**]
Expanded Terms for Concept Mapping: Grouped by Normalization

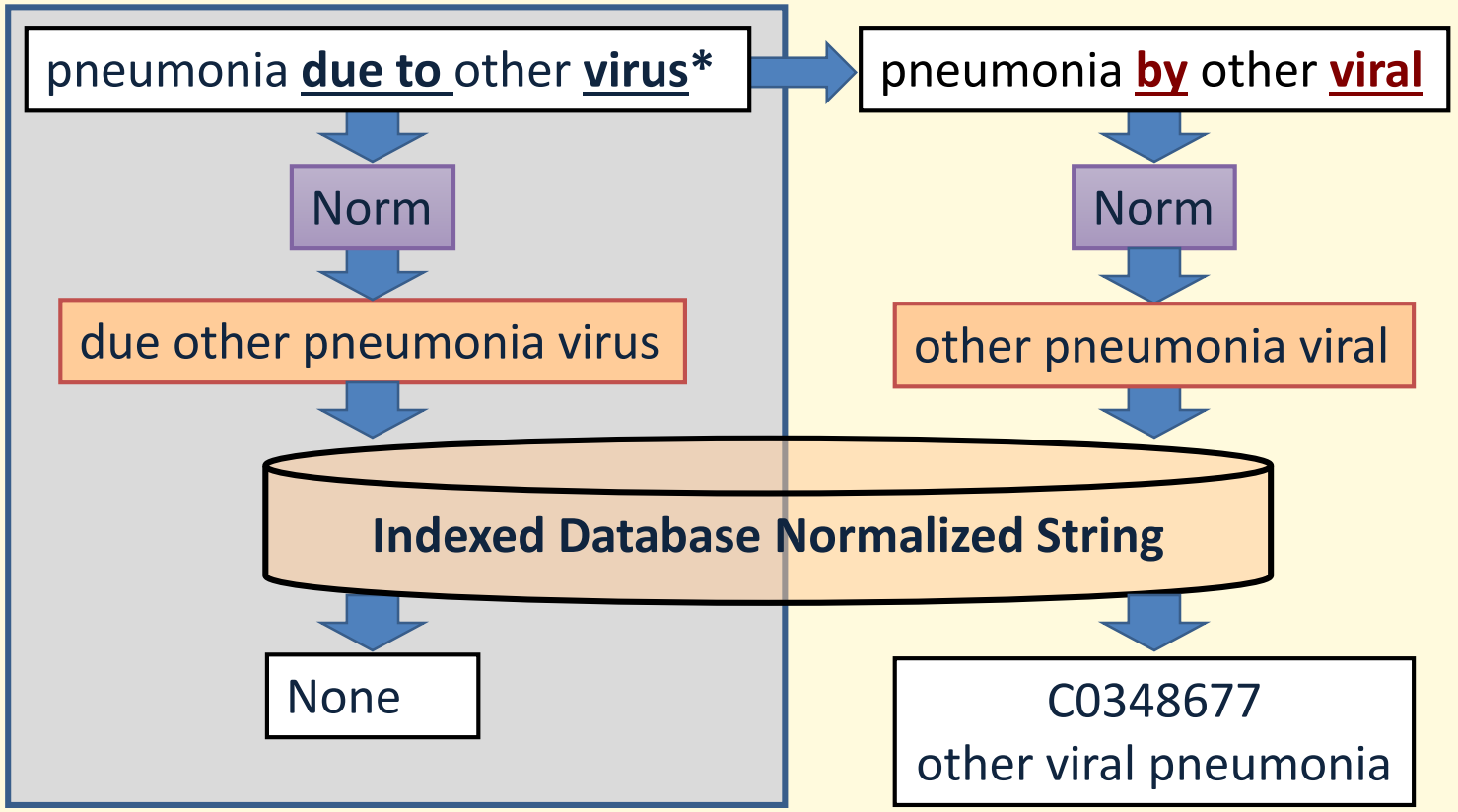
C0281926:

- **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*
  - …

Element Synonyms

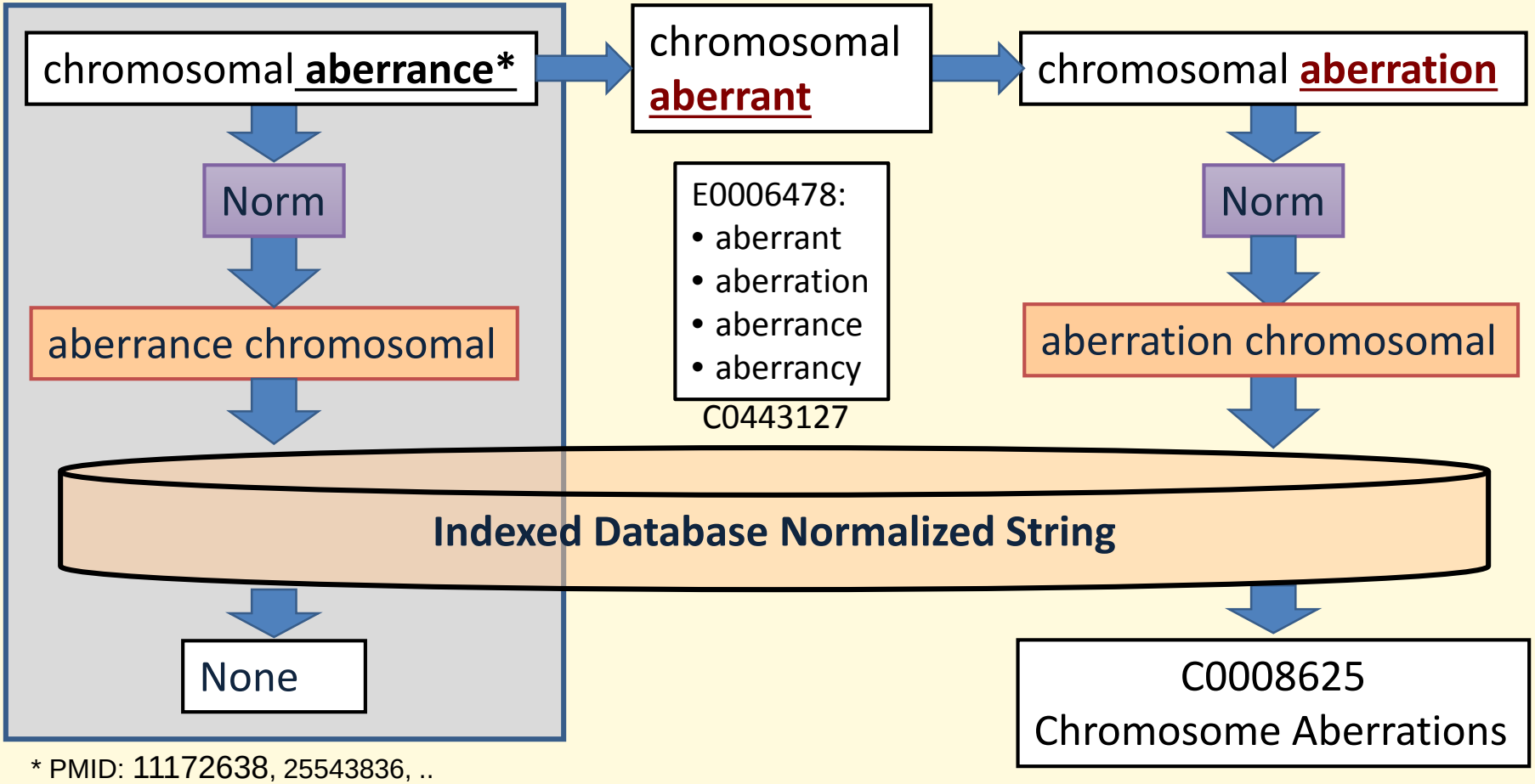


Multiple Substitutions [7-9]



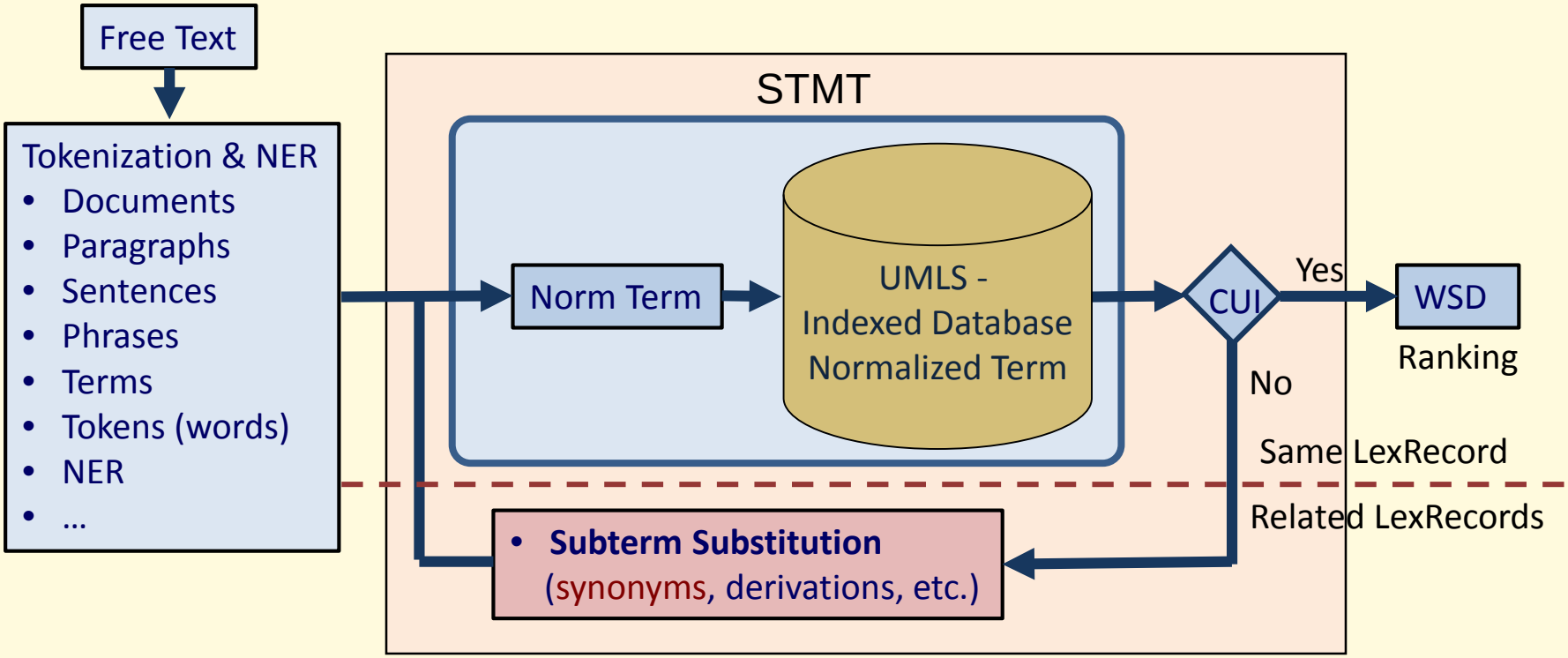
* VA14760, HA480.80, ..

Recursive Substitutions

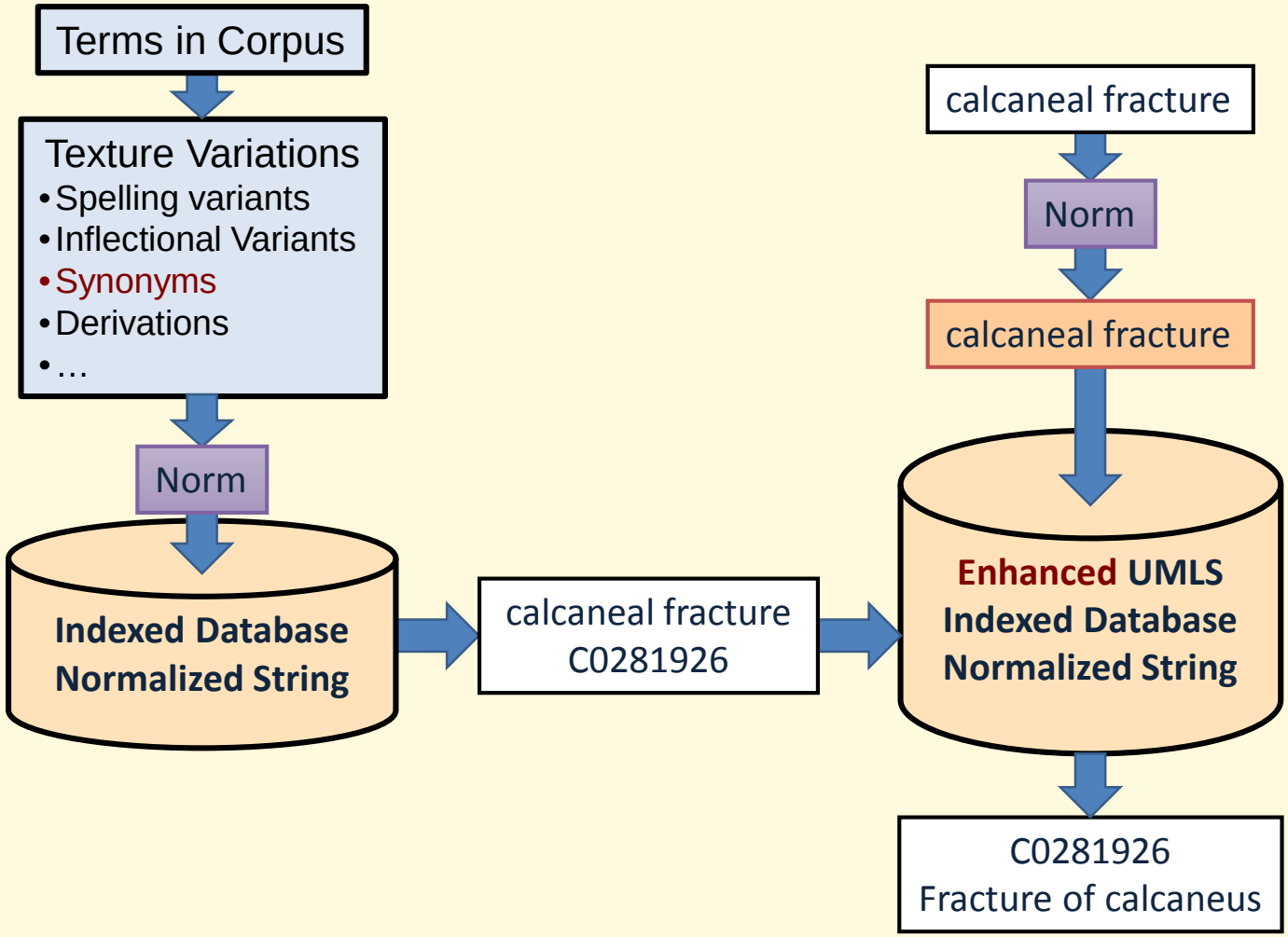


* PMID: 11172638, 25543836, ..

Real-time Model [5-6]



Pre-Processing Model [7-13]



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)
  - etc..

Objectives

- To generate a standalone set of element synonyms (sPairs) for effective UMLS concept mapping
  - Scope:
    - o include all synonymous terms in Lexicon (LexSynonyms)
    - o grow with the SPECIALIST Lexicon
    - o a thorough set of element synonyms (to increase recall)
  - Feature requirements:
    - o better performance: increase recall and preserve precision
    - o resolve known issues (near-synonyms, POS ambiguity, include multiword synonyms, etc.)
    - o 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)

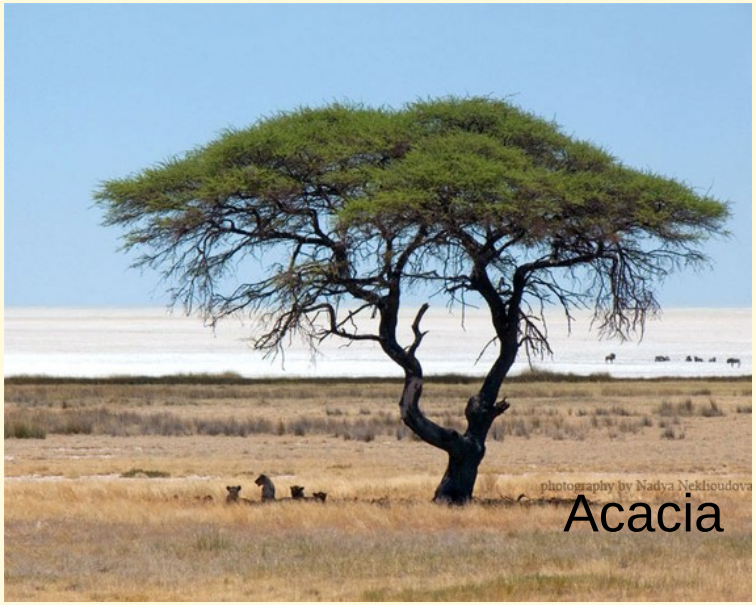
- Two properties:
  - **Commutativity: (x = y) -> (y = x)**
    - o joy|noun|enjoy|verb -> enjoy|verb|joy|noun
    - o bi-directional (sPair)
  - **Transitivity: ((x = y) and (y = z)) -> (x = z)**
    - o enjoy|verb -> joy|noun -> happy|adj
    - o multiple (recursive) substitutions
    - o sClass (synonym class)
- Prevent precision issues by near-synonyms.

Near-Synonyms

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

Acacia & Locust Tree

- C0000869



Anguilla & Antigua

- C0003353

Metencephalon & Pons

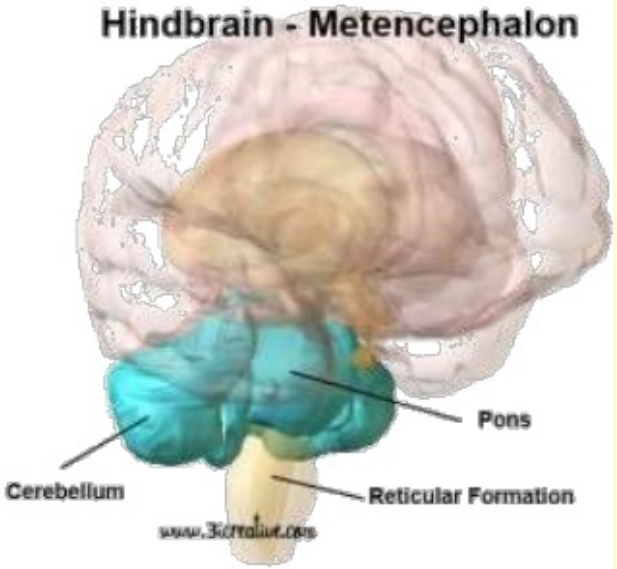
- C0032639

Hinbrain: *Metencephalon*

20

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

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

mug|verb

≠

mug|noun

R3: Source: CUI, EUI, …

The patient **expired** 1 day later.

CUI: C0011065
PT: Cessation of life
died
dead
death
deceased
…

Pressure of CO2 in **expired** air …

CUI: C0231800
PT: Expiration, Function
exhaled
expiratory
expiration
…

Disposal of **expired** drug …

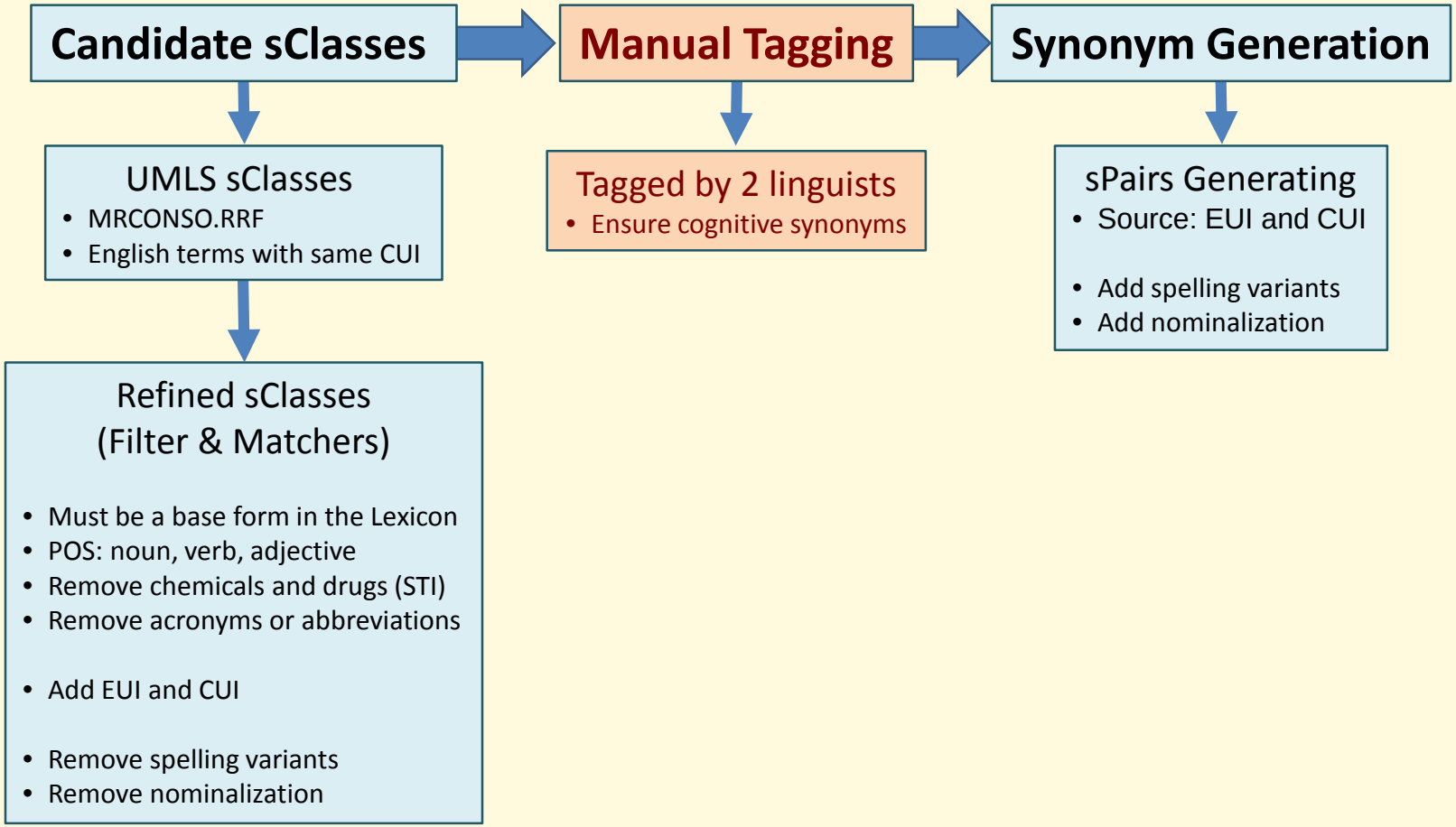
CUI: C1704631
PT: Expiration
expire
expiration
…

R4: Acronym/Abbreviation

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

CUI	Preferred Term	Synonym
C0003023	Angola	ago
C0001857	AIDS related complex	arc
C3714936	Non-Compliant ADaM Datasets Domain	ax

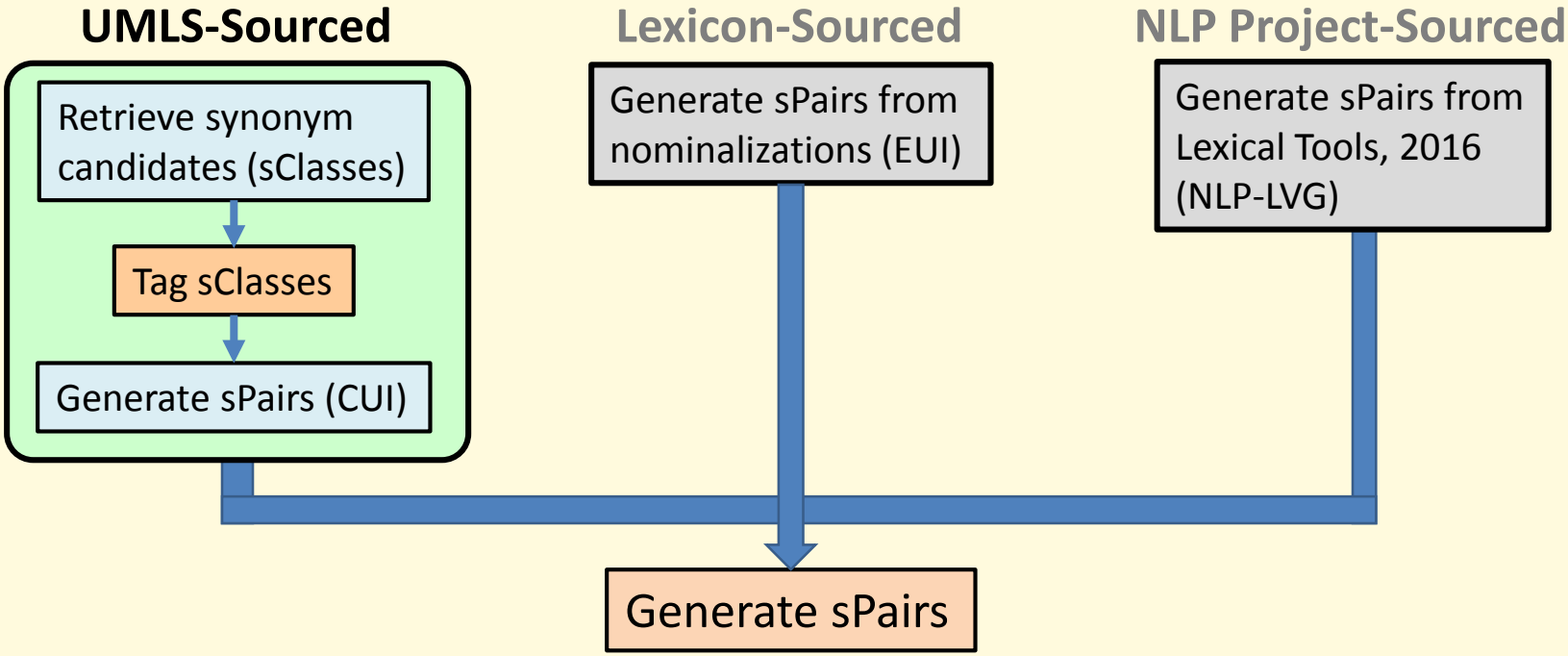
Computer-aided System



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
…

sPairs Generation



Synonym-1	POS-1	Synonym-2	POS-2	Source
mug	verb	assault	noun	C0004063
assault	noun	mug	verb	C0004063
…	…	…	…	…

Results – 2017 Release

- 2017 LexSynonyms

	Candidates	Tagged	Completion (%)
sClass	22,779	7,686	33.74%
Synonyms	80,913	29,990	37.06%

- Synonyms (sPairs):

Year	CUI	EUI	NLP	Total
2016	0 (0%)	0 (0%)	5,198 (100%)	5,198
2017	118,468 (62%)	67,584 (35%)	4,792 (3%)	190,844

36.71 growth

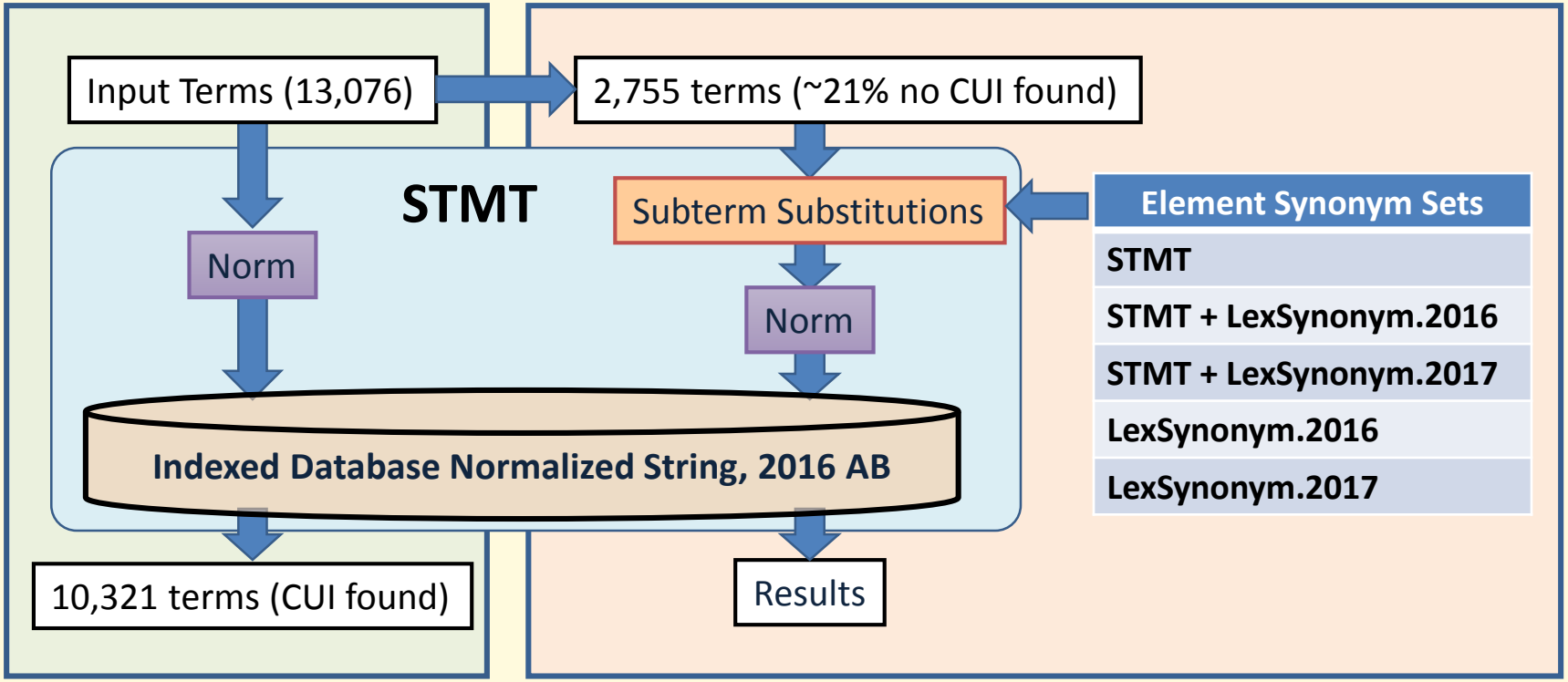
- Format:

Synonym-1	POS-1	Synonym-2	POS-2	Source

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 form 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



Evaluation 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.

Element Synonym Set	N. Size	T.P.	F.P.	F.N.	Precision	Recall	F1	Time
STMT [6]	7,873	690	353	2,066	66.16%	25.04%	0.3633	7:57
STMT + LexSynonym.2016	12,681	691	358	2,065	65.87%	25.07%	0.3632	5:31
STMT + LexSynonym.2017	151,913	828	424	1,928	**66.13%**	**30.04%**	**0.4132**	9:18
Element Synonym Set	**N. Size**	**T.P.**	**F.P.**	**F.N.**	**Precision**	**Recall**	**F1**	**Time**
LexSynonym.2016	5,070	9	12	2,747	42.86%	0.33%	0.0065	0:16
LexSynonym.2017	149,912	287	117	2,469	71.04%	10.41%	0.1816	3:19

Summary & Conclusion

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

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

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
- Chris Lu (E): chlu@mail.nih.gov