Natural Language Processing and Standardized Terminologies

Rui Zhang, MS
Institute for Health Informatics, University of Minnesota, Twin Cities

Genevieve B Melton, MD, MA, FACS, FASCRS
Department of Surgery & Institute for Health Informatics
University of Minnesota, Twin Cities

Natural Language Processing (NLP)

- Techniques to automatically analyze natural language (free text written by people)
- MRI revealed a lacunar infarction in the internal capsule.

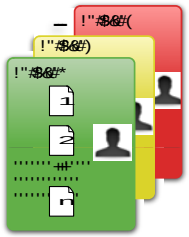
↓ Parsing, Named entity recognition (NER), etc.

MRI ***revealed*** a lacunar infarction ***in*** the internal capsule.

↓ Mapping, Acronym detection, Relationship extraction, etc.

Subject	Predicate (Indicator)	Object
Magnetic Resonance Imaging (MRI)	DIAGNOSES	Infarction, Lacunar
Internal Capsule	LOCATION_OF	Infarction, Lacunar

NLP in Health Sciences



Clinical Notes

Biomedical Literature

NLP

- Medication
- Problem list
- Medical history
- Smoking status
- …

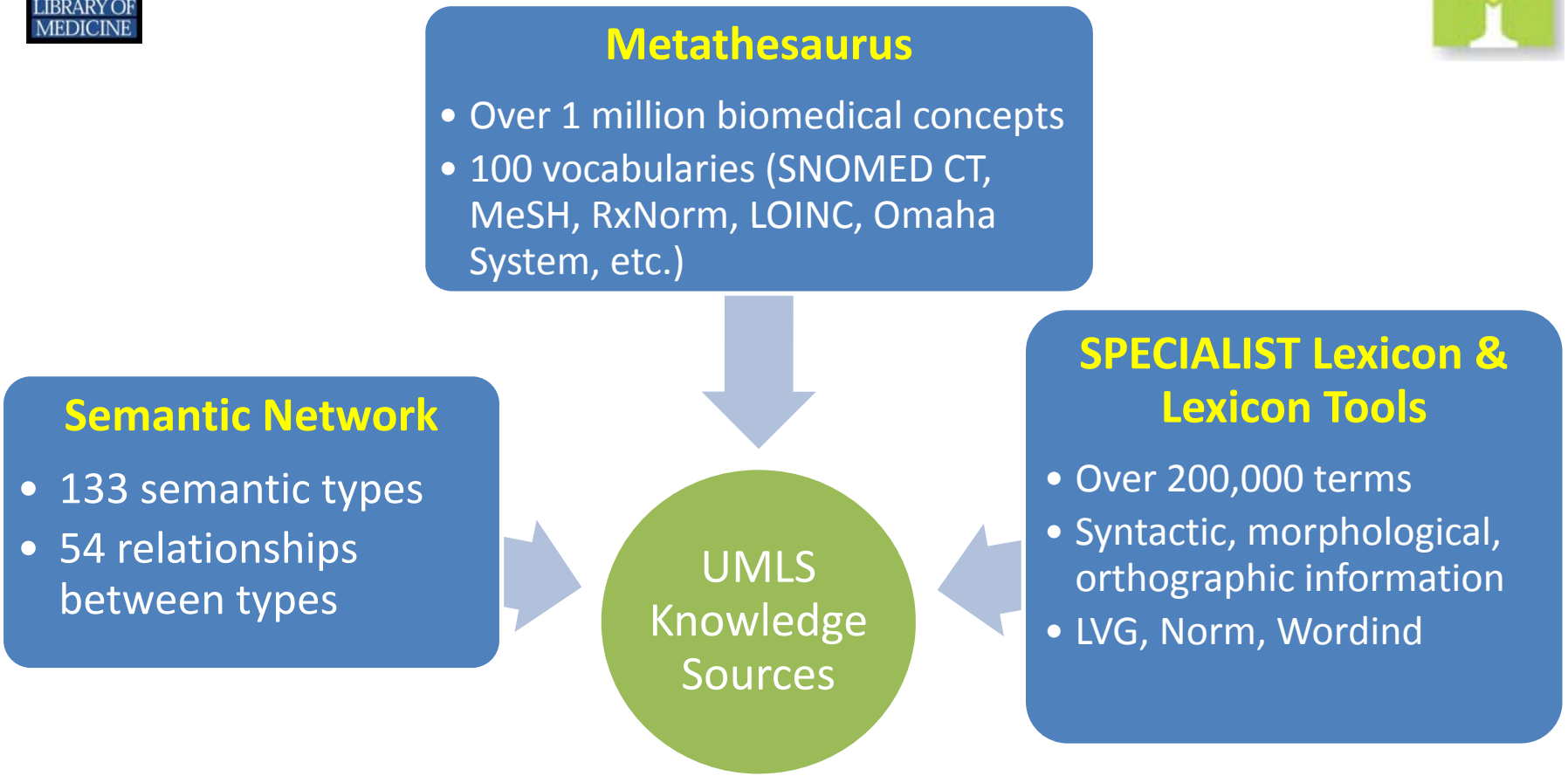
Biomedical knowledge (structured)

Health care providers, clinical researchers

Clinical NLP and Standardized Terminologies

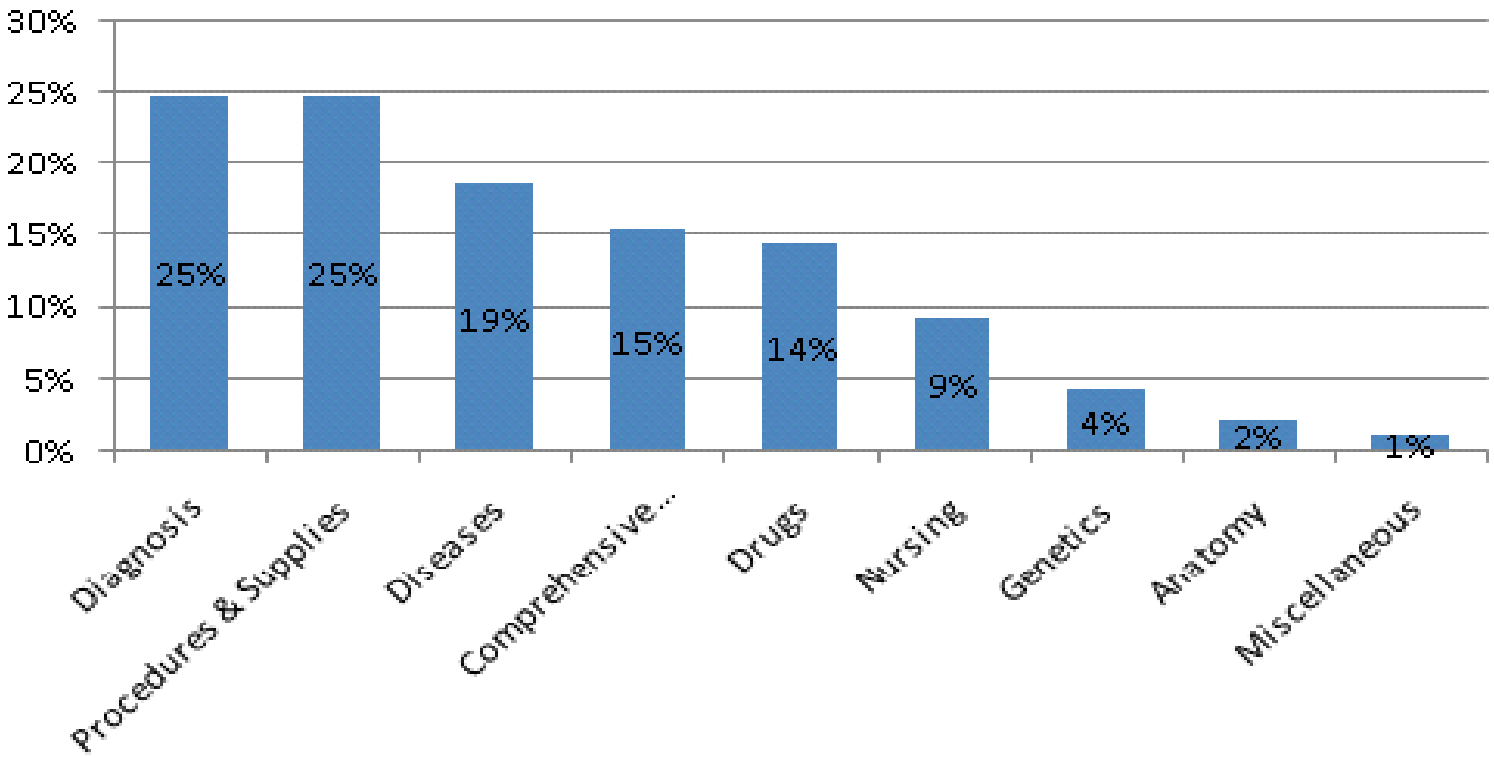
- Linguistic and medical knowledge are necessary to implement clinical NLP tasks
- Linguistic knowledge provides
  - Lexical information
  - Syntactic structure
- Medical knowledge provides
  - Standardized terminologies
  - Semantic network

Unified Medical Language System® (UMLS®)



http://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.htm

UMLS-Metathesaurus



Diagnosis: Logic Observation Identifier Names and Codes (LOINC)
Procedures & Supplies: Current Procedural Terminology (CPT)
Diseases: International Classification of Diseases and Related Health Problems (ICD-10)
Comprehensive: Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)

http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_002.htm

MetaMap

- **Map** biomedical text to the UMLS **Meta**thesaurus

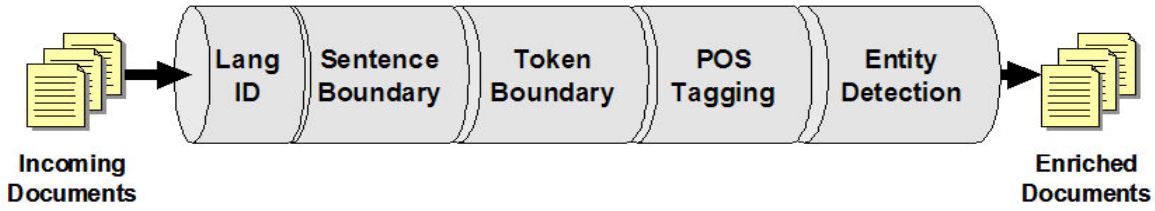
Phrase: "obstructive sleep apnea"
Meta Candidates
1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or Syndrome]
901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]
827 Apnea [Pathologic Function]
827 Sleep [Organism Function]
827 Obstructive (Obstructed) [Functional Concept]
827 Apnea (Apnea Adverse Event) [Finding]
793 E Sleeping (Asleep) [Finding]
755 E Sleepy (Drowsiness) [Finding]
727 E Sleeplessness [Sign or Symptom]
Meta Mapping (1000):
1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or Syndrome]

Aronson AR, Lang FM. *J Am Med Inform Assoc* 2010;17(3):229-236

Chaining NLP tasks: pipelines

- Any practical NLP task must perform sub-tasks (low-level tasks must execute sequentially)
- Pipelined system enables applications to be decomposed into components
- Each component does the actual work of analyzing the unstructured information
- Unstructured information management architecture (UIMA)

Unstructured Information Management Architecture
An Apache Project
TM



An Example

An example of a sentence discovered by the sentence boundary detector:

Fx of obesity but no fx of coronary artery diseases.

Tokenizer output – 11 tokens found:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
```

Normalizer output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  disease   .
```

Part-of-speech tagger output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NN  IN  NN       CC   DT  NN  IN  JJ         NN       NNS       .
```

Shallow parser output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NP  PP  \NP/          \NP/  PP    \_____  NP  _____/
```

NN: noun
JJ: adjective
IN: preposition
CC: coordinating conjunction
DT: determiner
NNS: plural noun

NP: noun phrase
PP: preposition phrase

Named Entity Recognition – 5 Named Entities found:

```
Fx of obesity but no fx of coronary artery diseases .
      obesity  (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
                        coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
                        coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                diseases (type=diseases/disorders, CUI = C0010054)
```

Status and Negation attributes assigned to Named Entities:

```
Fx of obesity but no fx of coronary artery diseases .
      obesity (status = family_history_of; negation = not_negated)
                        coronary artery diseases (status = family_history_of, negation = is_negated)
```

Savova GK et al. *J Am Med Inform Assoc* 2010;17(5):507-513

Output Example: Drug Object

"Tamoxifen 20 mg po once daily started on March 1, 2005."

- Drug
  - Text: Tamoxifen
  - Associated code: C0351245
  - Strength: 20 mg
  - Start date: March 1, 2005
  - End date: null
  - Frequency: 1.0
  - Frequency unit: daily
  - Duration: null
  - Route: Enteral Oral
  - Form: null
  - Status: current
  - Change Status: no change

po: per oral/ by mouth

NLP of Nursing Narratives

- To compare the semantic categories of MedLEE and ISO reference terminology models for nursing diagnoses and actions
- In aspects of site or location, MedLEE was more granular than ISO models
- In clinical procedure, two ISO components (action and target) mapped to one MedLEE semantic category
- The ISO models requires additional specification of selected semantic categories
- Analysis also suggested areas for extension of MedLEE

MedLEE: Medical Language Extraction and Encoding system, Columbia University
ISO: International Standards Organization
Bakken S, Hyun S, Friedman C, Johnson SB, Int J Med Info 2005 74, 615-622.

Analysis of Free Text to Inform Terminology Development

- Analyze text associated with “other” targets within Omaha system interventions
- To understand the clinicians’ information needs
- To identify additional suggested and new targets
- In particular, new targets were suggested for:
  - Daily living
  - Disease pathophysiology
  - Pain management

Melton GB, Westra BL, Raman N, Monsen KA, et al. *Proc AMIA* 2010, 512-516.
Farri O, Monsen KA, Westra BL, Melton GB. *Appl Clin Inf* 2011; 2: 304–316

Summary

- Linguistic and medical knowledge are needed to implement clinical NLP tasks
- UMLS provides useful standardized terminologies for clinical NLP applications
- UIMA provides pipelined framework to analyze clinical texts
- Analysis of NLP systems and free texts can inform the development of terminologies