

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



First International Conference on Research Methods for Standardized Terminologies



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



First International Conference on Research Methods for Standardized Terminologies




NLP in Health Sciences



Clinical Notes


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

NLP




Biomedical Literature


Biomedical knowledge (structured)



Health care providers, clinical researchers



First International Conference on Research Methods for Standardized Terminologies



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

First International Conference on Research Methods for Standardized Terminologies

Unified Medical Language System® (UMLS®)

Metathesaurus

- Over 1 million biomedical concepts
- 100 vocabularies (SNOMED CT, MeSH, RxNorm, LOINC, Omaha System, etc.)

Semantic Network

- 133 semantic types
- 54 relationships between types

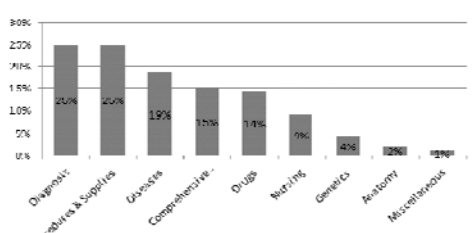
SPECIALIST Lexicon & Lexicon Tools

- Over 200,000 terms
- Syntactic, morphological, orthographic information
- LVG, Norm, Wordind

UMLS Knowledge Sources

http://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.htm
 First International Conference on Research Methods for Standardized Terminologies

UMLS-Metathesaurus



Category	Percentage
Diagnosis	21%
Procedures & Supplies	20%
Diseases	19%
Comprehensive	17%
DUIDS	14%
RxNorm	7%
Gene/Co	4%
Anatomic	2%
Miscellaneous	2%

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
 First International Conference on Research Methods for Standardized Terminologies

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

First International Conference on Research Methods for Standardized Terminologies

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)

First International Conference on Research Methods for Standardized Terminologies

An Example

cyakes

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 diseases .

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

Named Entity Recognition - 5 Named Entities found:
fx of obesity but no fx of coronary artery diseases .
obesity (type=disease/disorders, SNLS CUI=C0028794, SDRMSO-CT c028794134008 and 5474004)
coronary artery diseases (type=disease/disorders, CUI=C0010054, SDRMSO-CT=8957000)
coronary artery (type=anatomy, CUI(U) and SDRMSO-CT c008481004)
artery (type=anatomy, CUI(U) and SDRMSO-CT c008481004)
diseases (type=disease/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 = not_negated)

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

First International Conference on Research Methods for Standardized Terminologies

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 po: per oral/ by mouth
- Form: null
- Status: current
- Change Status: no change



First International Conference on Research Methods for Standardized Terminologies



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
First International Conference on Research Methods for Standardized Terminologies



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



First International Conference on Research Methods for Standardized Terminologies