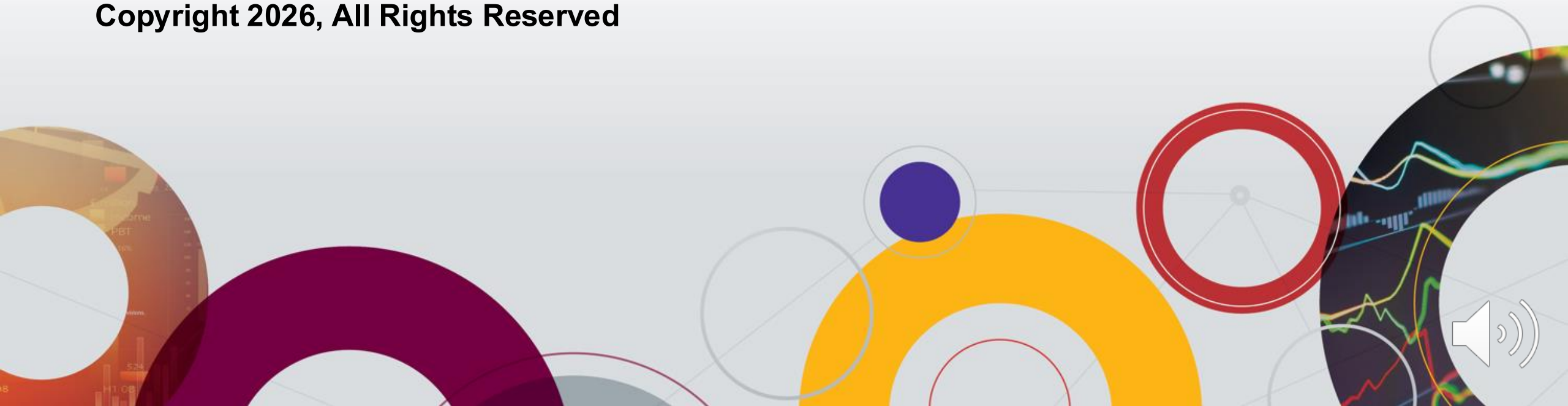


6.5a Natural Language Processing

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Natural language processing (NLP)

Definitions and overviews

Motivations and uses

Challenges for processing clinical narratives

Methods and results

Challenges for clinical NLP

Future directions

Definitions and overviews

NLP is computational understanding of natural language; other names include

- Computational linguistics
- Text mining
- Better name might be “natural language understanding?”

Why in biomedicine and health?

- Much clinical information is “locked” in text (Hripcsak, 1995; Hripcsak, 2013)
- Estimated that 70-80% of medical record is text (Murdoch, 2013; Roberts, 2017)
- Current coded data does not cover complexity of clinical narrative (Jollis, 1993; O’Malley, 2005)
- Clinicians prefer to read and write patient story (Vigilante, 2018; Gantzer, 2020)

Overviews

- General (Lane, 2025; Jurafsky, 2026)
- Biomedical and health (Xu, 2024)

Overview of clinical NLP (Demner-Fushman, 2021)

Information extraction – process narrative to structured data

Summarization – summarizing from larger amount of content

Text classification – categorizing text into known types

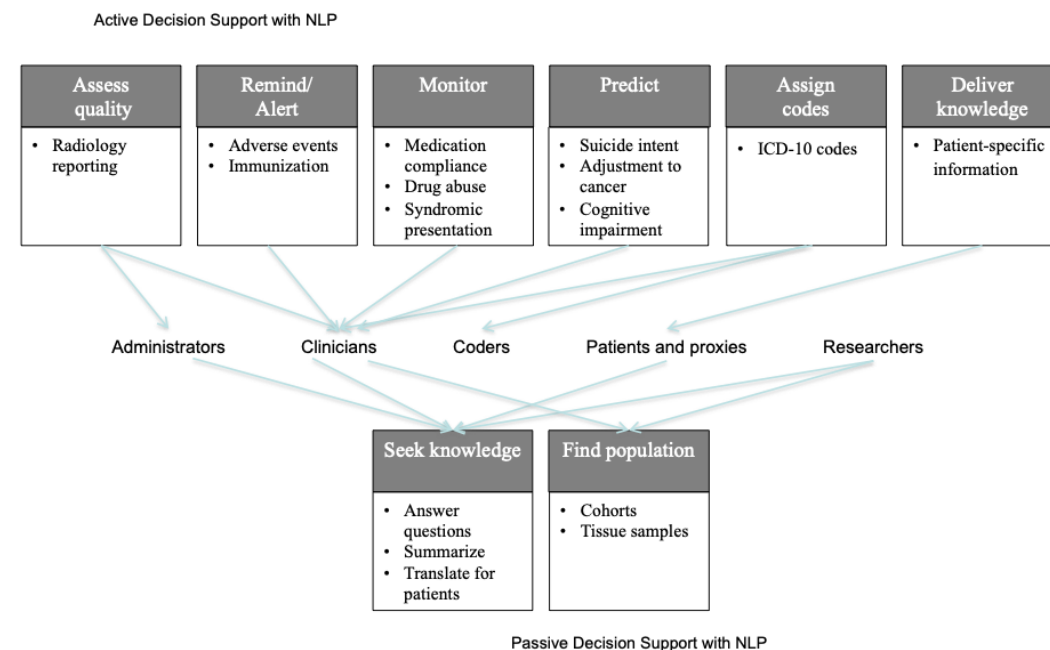
Information retrieval – retrieve documents and other text

Question-answering – finding answers in text

Machine translation

Conversational agents

Sentiment analysis



Uses of clinical NLP (Percha, 2021)

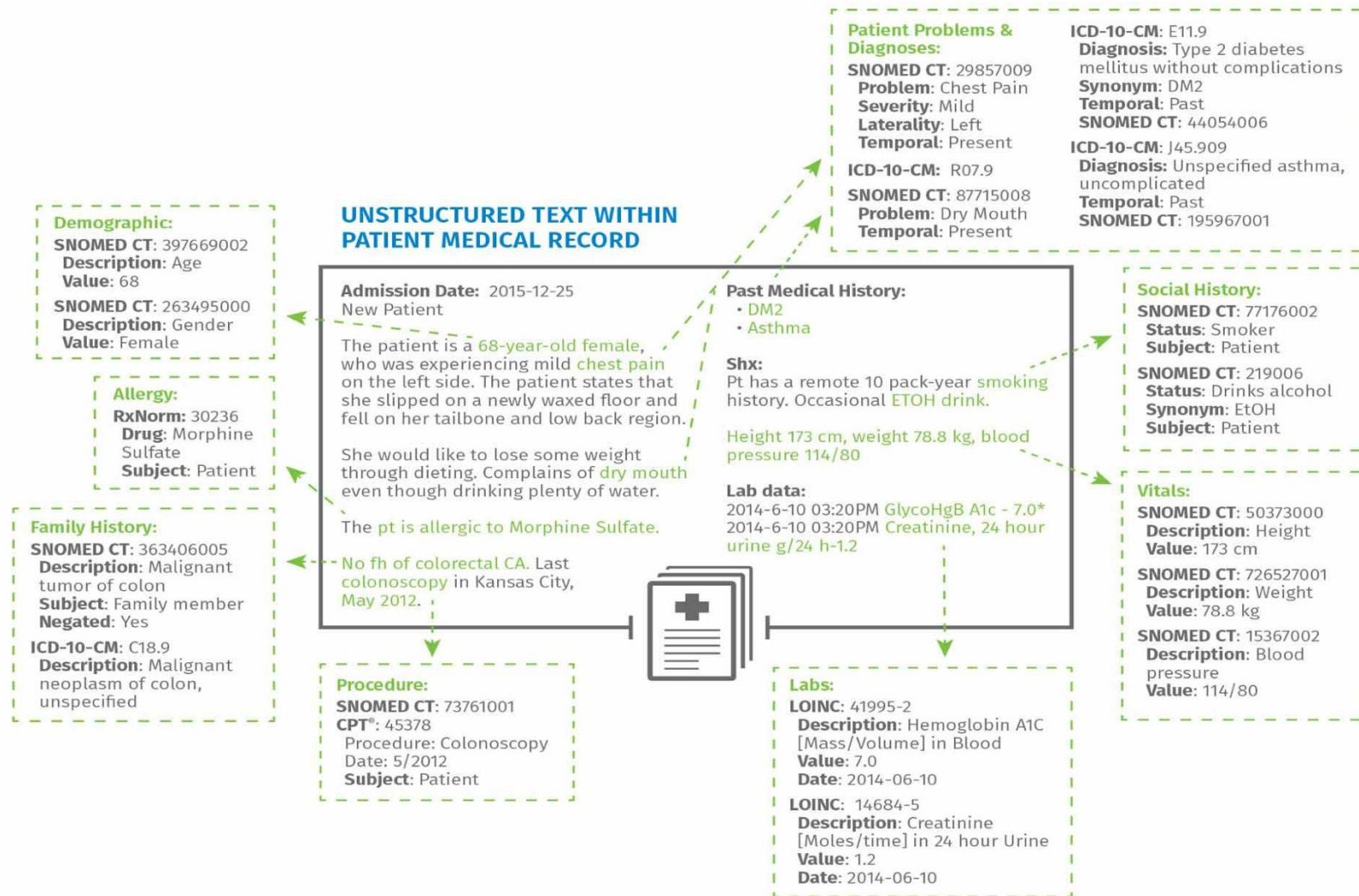
Tasks

- Extraction – identify important concepts and their relationships in text
- Modeling – use text, with or without extraction, to create predictive models

Extraction methods

- Named-entity recognition (NER) – identify entity and (usually) type
 - E.g., recognize presence of *hypertension* (type of disease)
- Concept normalization – map concept to normalized form (code)
 - E.g., recognize *hypertension*, *HTN*, and *high blood pressure* refer to same concept
- Relation extraction – structure concepts and relationships
 - E.g., recognize *chest pain* is a symptom of some types of *heart disease* or *pneumonia*

Use of NLP for EHR text



Levels of human language

Phonology – sound units that make up language, phonemes

Morphology – analysis of parts of words, morphemes

- e.g. *appendic-, pharyng-, -itis, -ectomy*

Syntax – rules governing construction of language, grammar

Semantics – meaning of words, phrases, and sentences

Pragmatics – how context effects meaning of sentences, discourse

World knowledge – general knowledge necessary to understand language

Phases of NLP

Three major phases in classical NLP

- Syntax – recognition of grammatical constituents of language
- Semantics – recognition of meaning
- Context – larger framing of content

Each level is successively harder and requires more knowledge engineering, but would add more value if problems could be solved

Challenges in processing clinical narrative

Clinical narratives are more difficult to process than other types of text for many reasons

- Written in telegraphic, elliptical style
- Often has spelling and/or grammatical errors
- Physicians and others may take “license” with language
- Important information can be buried among “normal” information

Challenges can be

- Syntactic
- Semantic
- Contextual

Syntactic challenges

Clinical narrative text syntactically incomplete (Sager, 1987)

- Half of all sentences incomplete
- Minimal English sentence is subject-verb-object
- Examples in order of frequency
 - Deleted verb and object: *Stiff neck and fever*
 - Deleted verb: *Brain scan negative*
 - Deleted subject and verb: *Positive for heart disease*
 - Deleted subject: *Was seen by local doctor*

Semantic challenges

Words have different senses and meanings

- *Murmur is appreciated*
- *Eye drops*
- *Mass at 3 o'clock*

Synonymy – different words/phrases having same meaning

- *Epigastric pain after eating vs. postprandial stomach discomfort*

Polysemy – same words/phrases having different meaning

- *The PCP of the patient with PCP advised him to stop using PCP*

Semantic challenges (cont.)

Negation – common in medical text (Wu, 2014)

- *Patient does not have any chest pain*

Uncertainty

- *Patient treated for possible pneumonia*

Temporality

- *Patient has history of pneumonia*
- *Chest pain resolved after administration of nitroglycerin*

Challenges for numerical data in clinical notes (Hanauer, 2019)

Spelled out – negatives (*minus*), fractions (*one-half*), dimensions (*two by two*), and ranges (*one to five*)

Roman numbers – *IV*, *type II*, *stage 3*

Invalid dates

Biologically implausible ages

Decades – *octogenarian*

Ranking correct (*1st*) and incorrect (*3st*), including with dates

Imprecision – *a few*, *a million*

Units – *lbs*, *pounds*



Contextual challenges

Coreference – relation between linguistic expressions that refer to same real-world entity (Zheng, 2011; Kilicoglu, 2016)

- *Chest x-ray shows nodule in left upper lobe. The tumor has increased in size to 2 cm.*
- One particular type of coreference is anaphora – use of pronouns
 - *He complains of chest pain. It awakens him at night.*

Ellipsis – deletion of subjects is common in clinical narratives

- *Complains of chest pain. Increasing frequency. Worse in the morning.*

Practical challenges

“Note bloat” and redundancy in clinical notes reduce NLP model performance (Liu, 2022)

- Growing note length and redundancy over years (Rule, 2021)
- Half of discharge summary content emanates from outside records of hospital stay – 39% from other records and 11% from no records at all (Ando, 2022)

Benchmark data sets tend to focus on information task and not clinical need (Blagec, 2023)

In voice recognition, non-lexical conversational sounds (*mm-hm, uh-huh, etc.*) degrade performance (Tran, 2023)