INFORMATICS PROFESSIONALS. LEADING THE WAY.

4E – Natural Language Processing

Thomas Payne, MD, FACMI, FAMIA

University of Washington

Clinical Informatics
Board Review Course

Clinical Informatics Subspecialty Delineation of Practice (CIS DoP)

Domain 1: Fundamental Knowledge and Skills (no Tasks are associated with this Domain which is focused on fundamental knowledge and skills)

Clinical Informatics

K001. The discipline of informatics (e.g., definitions, history, careers, professional organizations)
K002. Fundamental informatics concepts, models, and theories
K003. Core clinical informatics literature (e.g., foundational literature, principle journals, critical analysis of literature, use of evidence to inform practice)
K004. Descriptive and inferential statistics
K005. Health Information Technology (HIT) principles and science
K006. Computer programming fundamentals and computational thinking
K007. Basic systems and network architectures
K008. Basic database structure, data retrieval and analytics techniques and tools
K009. Development and use of interoperability/exchange standards (e.g., Fast Health Interoperability Resources [FHIR], Digital Imaging and Communications in Medicine [DICOM])
K010. Development and use of transaction standards (e.g., American National Standards Institute X12)
K011. Development and use of messaging standards (e.g., Health Level Seven [HL7] v2)
K012. Development and use of ancillary data standards (e.g., imaging and Laboratory Information System[LIS])
K013. Development and use of data model standards
K014. Vocabularies, terminologies, and nomenclatures (e.g., Logical Observation Identifiers Names and Codes [LOINC], Systematized Nomenclature of Medicine --Clinical Terms [SNOMED-CT], RxNorm, International Classification Of Diseases[ICD], Current Procedural Terminology [CPT])
K015. Data taxonomies and ontologies
K016. Security, privacy, and confidentiality requirements and practices
K017. Legal and regulatory issues related to clinical data and information sharing
K018. Technical and non-technical approaches and barriers to interoperability
K019. Ethics and professionalism

The Health System

K020. Primary domains of health, organizational structures, cultures, and processes (e.g., health care delivery, public health, personal health, population health, education of health professionals, clinical research)
K021. Determinants of individual and population health
K022. Forces shaping health care delivery and considerations regarding health care access
K023. Health economics and financing
K024. Policy and regulatory frameworks related to the healthcare system
K025. The flow of data, information, and knowledge within the health system

Domain 2: Improving Care Delivery and Outcomes

K026. Decision science (e.g., Bayes theorem, decision analysis, probability theory, utility and preference assessment, test characteristics)
K027. Clinical decision support standards and processes for development, implementation, evaluation, and maintenance
K028. Five Rights of clinical decision support (i.e., information, person, intervention formats, channel, and point/time in workflow)
K029. Legal, regulatory, and ethical issues regarding clinical decision support
K030. Methods of workflow analysis
K031. Principles of workflow re-engineering
K032. Quality improvement principles and practices (e.g., Six Sigma, Lean, Plan-Do-Study-Act [PDSA] cycle, root cause analysis)
K033. User-centered design principles (e.g., iterative design process)
K034. Usability testing
K035. Definitions of measures (e.g., quality performance, regulatory, pay for performance, public health surveillance)
K036. Measure development and evaluation processes and criteria
K037. Key performance indicators (KPIs)
K038. Claims analytics and benchmarks
K039. Predictive analytic techniques, indications, and limitations
K040. Clinical and financial benchmarking sources (e.g., Gartner, Healthcare Information and Management Systems Society [HIMSS] Analytics, Centers for Medicare and Medicaid Services [CMS], Leapfrog)
K041. Quality standards and measures promulgated by quality organizations (e.g., National Quality Forum [NQF], Centers for Medicare and Medicaid Services [CMS], National Committee for Quality Assurance [NCQA])
K042. Facility accreditation quality and safety standards (e.g., The Joint Commission, Clinical Laboratory Improvement Amendments [CLIA])
K043. Clinical quality standards (e.g., Physician Quality Reporting System [PQRS], Agency for Healthcare Research and Quality [AHRQ], National Surgical Quality Improvement Program [NSQIP], Quality Reporting Document Architecture [QRDA], Health Quality Measure Format [HQMF], Council on Quality and Leadership [CQL], Fast Health Interoperability Resources [FHIR] Clinical Reasoning)
K044. Reporting requirements
K045. Methods to measure and report organizational performance
K046. Adoption metrics (e.g., Electronic Medical Records Adoption Model [EMRAM], Adoption Model for Analytics Maturity [AMAM])
K047. Social determinants of health
K048. Use of patient-generated data
K049. Prediction models
K050. Risk stratification and adjustment
K051. Concepts and tools for care coordination
K052. Care delivery and payment models

Domain 3: Enterprise Information Systems

K053. Health information technology landscape (e.g., innovation strategies, emerging technologies)
K054. Institutional governance of clinical information systems
K055. Information system maintenance requirements
K056. Information needs analysis and information system selection
K057. Information system implementation procedures
K058. Information system evaluation techniques and methods
K059. Information system and integration testing techniques and methodologies
K060. Enterprise architecture (databases, storage, application, interface engine)
K061. Methods of communication between various software components
K062. Network communications infrastructure and protocols between information systems (e.g., Transmission Control Protocol/Internet Protocol [TCP/IP], switches, routers)
K063. Types of settings (e.g., labs, ambulatory, radiology, home) where various systems are used
K064. Clinical system functional requirements
K065. Models and theories of human-computer (machine) interaction (HCI)
K066. HCI evaluation, usability engineering and testing, study design and methods
K067. HCI design standards and design principles
K068. Functionalities of clinical information systems (e.g., Electronic Health Records [EHR], Laboratory Information System [LIS], Picture Archiving and Communication System [PACS], Radiology Information System [RIS] vendor-neutral archive, pharmacy, revenue cycle)
K069. Consumer-facing health informatics applications (e.g., patient portals, mobile health apps and devices, disease management, patient education, behavior modification)
K070. User types and roles, institutional policy and access control
K071. Clinical communication channels and best practices for use (e.g., secure messaging, closed loop communication)
K072. Security threat assessment methods and mitigation strategies
K073. Security standards and safeguards
K074. Clinical impact of scheduled and unscheduled system downtimes
K075. Information system failure modes and downtime mitigation strategies (e.g., replicated data centers, log shipping)
K076. Approaches to knowledge repositories and their implementation and maintenance
K077. Data storage options and their implications
K078. Clinical registries
K079. Health information exchanges
K080. Patient matching strategies
K081. Master patient index
K082. Data reconciliation
K083. Regulated medical devices (e.g., pumps, telemetry monitors) that may be integrated into information systems
K084. Non-regulated medical devices (e.g., consumer devices)
K085. Telehealth workflows and resources (e.g., software, hardware, staff)

Domain 4: Data Governance and Data Analytics

K086. Stewardship of data
K087. Regulations, organizations, and best practice related to data access and sharing agreements, data use, privacy, security, and portability
K088. Metadata and data dictionaries
K089. Data life cycle
K090. Transactional and reporting/research databases
K091. Techniques for the storage of disparate data types
K092. Techniques to extract, transform, and load data
K093. Data associated with workflow processes and clinical context
K094. Data management and validation techniques
K095. Standards related to storage and retrieval from specialized and emerging data sources
K096. Types and uses of specialized and emerging data sources (e.g., imaging, bioinformatics, internet of things (IoT), patient-generated, social determinants)
K097. Issues related to integrating emerging data sources into business and clinical decision making
K098. Information architecture
K099. Query tools and techniques
K100. Flat files, relational and non-relational/NoSQL database structures, distributed file systems
K101. Definitions and appropriate use of descriptive, diagnostic, predictive, and prescriptive analytics
K102. Analytic tools and techniques (e.g., Boolean, Bayesian, statistical/mathematical modeling)
K103. Advanced modeling and algorithms
K104. Artificial intelligence
K105. Machine learning (e.g., neural networks, support vector machines, Bayesian network)
K106. Data visualization (e.g., graphical, geospatial, 3D modeling, dashboards, heat maps)
K107. Natural language processing
K108. Precision medicine (customized treatment plans based on patient-specific data)
K109. Knowledge management and archiving science
K110. Methods for knowledge persistence and sharing
K111. Methods and standards for data sharing across systems (e.g., health information exchanges, public health reporting)

Domain 5: Leadership and Professionalism

K112. Environmental scanning and assessment methods and techniques
K113. Consensus building, collaboration, and conflict management
K114. Business plan development for informatics projects and activities (e.g., return on investment, business case analysis, pro forma projections)
K115. Basic revenue cycle
K116. Basic managerial/cost accounting principles and concepts
K117. Capital and operating budgeting
K118. Strategy formulation and evaluation
K119. Approaches to establishing Health Information Technology (HIT) mission and objectives
K120. Communication strategies, including one-on-one, presentation to groups, and asynchronous communication
K121. Effective communication programs to support and sustain systems implementation
K122. Writing effectively for various audiences and goals
K123. Negotiation strategies, methods, and techniques
K124. Conflict management strategies, methods, and techniques
K125. Change management principles, models, and methods
K126. Assessment of organizational culture and behavior change theories
K127. Theory and methods for promoting the adoption and effective use of clinical information systems
K128. Motivational strategies, methods, and techniques
K129. Basic principles and practices of project management
K130. Project management tools and techniques
K131. Leadership principles, models, and methods
K132. Intergenerational communication techniques
K133. Coaching, mentoring, championing and cheerleading methods
K134. Adult learning theories, methods, and techniques
K135. Teaching modalities for individuals and groups
K136. Methods to assess the effectiveness of training and competency development
K137. Principles, models, and methods for building and managing effective interdisciplinary teams
K138. Team productivity and effectiveness (e.g., articulating team goals, defining rules of operation, clarifying individual roles, team management, identifying and addressing challenges)
K139. Group management processes (e.g., nominal group, consensus mapping, Delphi method)

Knowledge Statements from the DoP

K107 Natural language processing

Definition of natural language processing

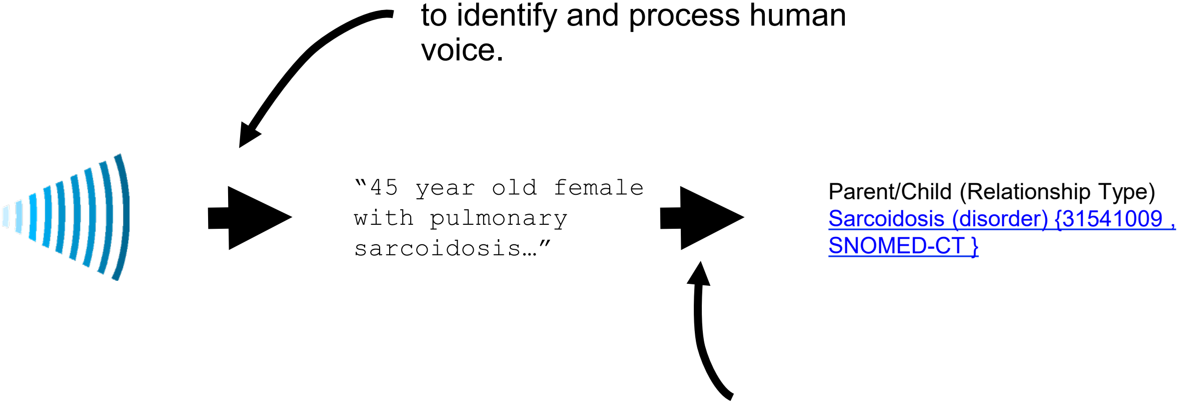
Natural language processing (NLP) systems are automated methods containing some linguistic knowledge that aim to improve the management of information in text.

NLP systems have been shown to be successful for realistic clinical applications, such as decision support, surveillance of infectious diseases, research studies, automated encoding, quality assurance, indexing patient records, and tools for billing.

Friedman 2005

Sound to meaning - Definitions

Automated speech recognition
The use of computer hardware and software-based techniques to identify and process human voice.



"45 year old female with pulmonary sarcoidosis…"

Parent/Child (Relationship Type)
Sarcoidosis (disorder) {31541009 , SNOMED-CT }

Natural language processing
A subfield of artificial intelligence and computational linguistics. It studies the problems of automated generation and understanding of natural human languages

NLP use cases in clinical computing

- Clinical decision support
- Findings in clinical notes radiology reports
- Detection of adverse medication events, social determinants of health, smoking status
- Speech recognition
- Computer-assisted coding
- Research
- Computational phenotyping

More broadly:

- Named entity recognition
  - Diseases
  - Medications
  - ADEs
- Relation extraction
  - Medication attribute relations (dose, sig, route)
  - Drug-drug interaction

Low-level NLP tasks:

1. Sentence boundary detection: abbreviations and titles (‘m.g.,’‘Dr.’) complicate this task, as do items in a list or templated utterances (eg, ‘MI [x], SOB[]’).

2. Tokenization: identifying individual tokens (word, punctuation) within a sentence. A lexer plays a core role for this task and the previous one. In biomedical text, tokens often contain characters typically used as token boundaries, for example, hyphens, forward slashes (‘10 mg/day,’ ‘N-acetyl-cysteine’).

3. Part-of-speech assignment to individual words in English, homographs (‘set’) and gerunds (verbs ending in ‘ing’ that are used as nouns) complicate this task.

Nadkarni, JAMIA 2011

Low-level NLP tasks, continued:

4. Morphological decomposition of compound words: many medical terms, for example, ‘nasogastric,’ need decomposition

5. Shallow parsing (chunking): identifying phrases from constituent part-of-speech tagged tokens. For example, a noun phrase may comprise an adjective sequence followed by a noun.

6. Problem-specific segmentation: segmenting text into meaningful groups, such as sections, including Chief Complaint, Past Medical History, HEENT, etc.

Nadkarni, JAMIA 2011

Higher-level NLP tasks:

1. Spelling/grammatical error identification and recovery:.

2. Named entity recognition: identifying specific words or phrases (‘entities’) and categorizing them for example, as persons, locations, diseases, genes, or medication.

3. Word sense disambiguation: determining a homograph’s correct meaning.

4. Negation and uncertainty detection: inferring whether a named entity is present or absent, and quantifying that inference’s uncertainty.

5. Relationship extraction: determining relationships between entities or events, such as ‘treats,’ ‘causes,’ and ‘occurs with.’

Nadkarni, JAMIA 2011

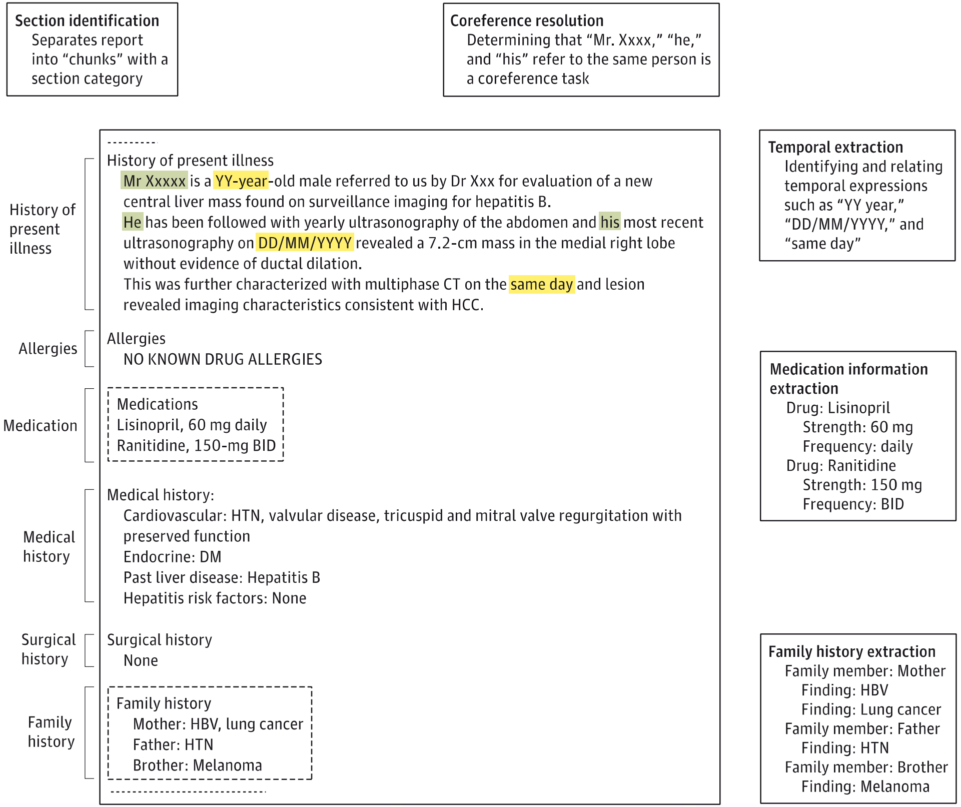
High-level NLP tasks, continued:

6. Temporal inferences/relationship extraction: E.g, medication X was prescribed after symptoms began.

7. Information extraction: the identification of problem- specific information and its transformation into (problem- specific) structured form. Tasks 1-6 are often part of the larger information task.)

Nadkarni, JAMIA 2011

Examples of NLP tasks when applied to notes



Section identification
Separates report into "chunks" with a section category

Coreference resolution
Determining that "Mr. Xxxx," "he," and "his" refer to the same person is a coreference task

History of present illness

History of present illness
Mr Xxxxx is a YY-year-old male referred to us by Dr Xxx for evaluation of a new central liver mass found on surveillance imaging for hepatitis B.
He has been followed with yearly ultrasonography of the abdomen and his most recent ultrasonography on DD/MM/YYYY revealed a 7.2-cm mass in the medial right lobe without evidence of ductal dilation.
This was further characterized with multiphase CT on the same day and lesion revealed imaging characteristics consistent with HCC.

Allergies

Allergies
NO KNOWN DRUG ALLERGIES

Medication

Medications
Lisinopril, 60 mg daily
Ranitidine, 150-mg BID

Medical history

Medical history:
Cardiovascular: HTN, valvular disease, tricuspid and mitral valve regurgitation with preserved function
Endocrine: DM
Past liver disease: Hepatitis B
Hepatitis risk factors: None

Surgical history

Surgical history
None

Family history

Family history
Mother: HBV, lung cancer
Father: HTN
Brother: Melanoma

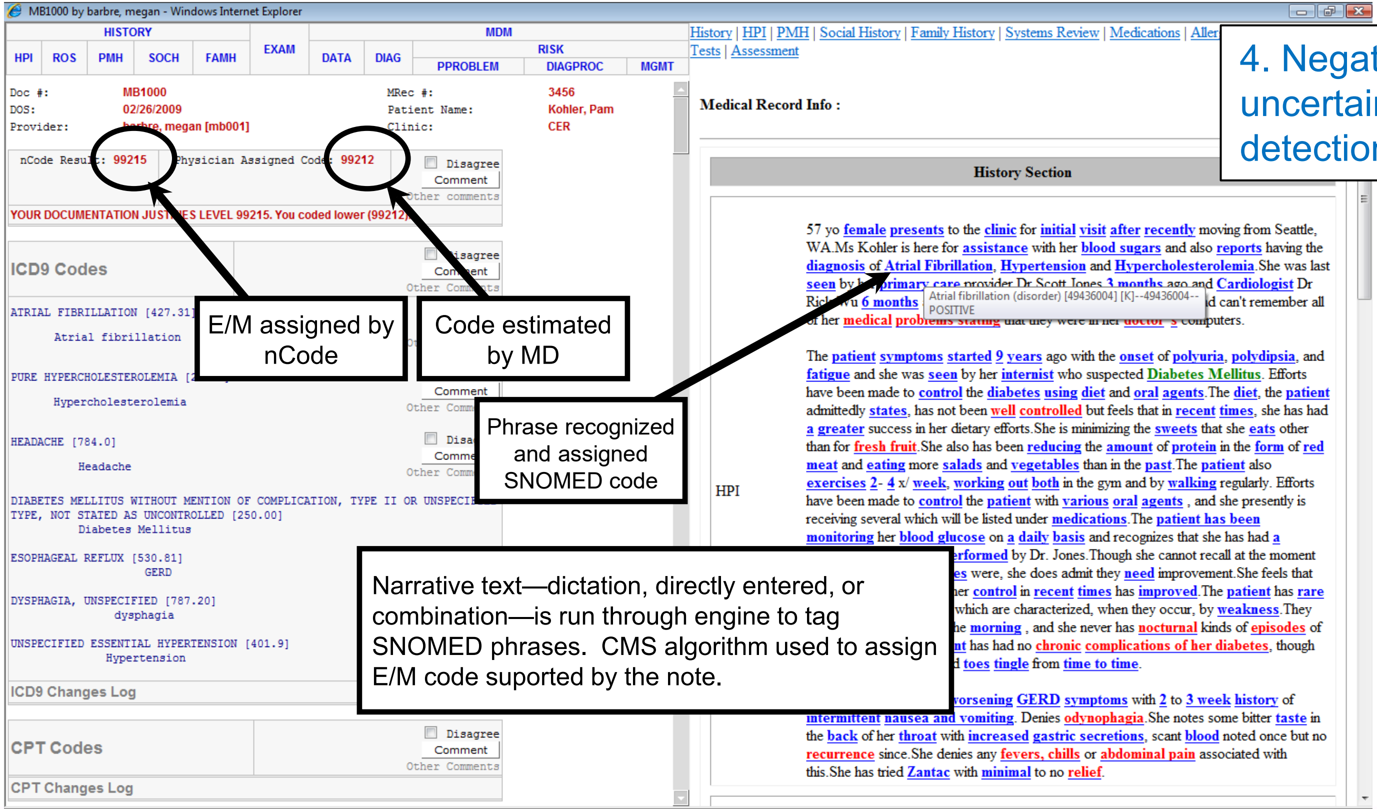
Temporal extraction
Identifying and relating temporal expressions such as "YY year," "DD/MM/YYYY," and "same day"

Medication information extraction
Drug: Lisinopril
Strength: 60 mg
Frequency: daily
Drug: Ranitidine
Strength: 150 mg
Frequency: BID

Family history extraction
Family member: Mother
Finding: HBV
Finding: Lung cancer
Family member: Father
Finding: HTN
Family member: Brother
Finding: Melanoma

Yim JAMA 2016

Example: Computer-assisted coding using NLP



MB1000 by barbre, megan - Windows Internet Explorer
HISTORY
HPI
ROS
PMH
SOCH
FAMH
EXAM
MDM
DATA
DIAG
RISK
PPROBLEM
DIAGPROC
MGMT
History | HPI | PMH | Social History | Family History | Systems Review | Medications | Aller
Tests | Assessment

Doc #: MB1000
DOS: 02/26/2009
Provider: barbre, megan [mb001]
MRec #: 3456
Patient Name: Kohler, Pam
Clinic: CER

nCode Result: 99215 Physician Assigned Code: 99212 □ Disagree Comment Other comments

YOUR DOCUMENTATION JUSTIFIES LEVEL 99215. You coded lower (99212)

ICD9 Codes □ Disagree Comment Other Comments

ATRIAL FIBRILLATION [427.31]
Atrial fibrillation

PURE HYPERCHOLESTEROLEMIA [2
Hypercholesterolemia

HEADACHE [784.0]
Headache

DIABETES MELLITUS WITHOUT MENTION OF COMPLICATION, TYPE II OR UNSPECIFIED TYPE, NOT STATED AS UNCONTROLLED [250.00]
Diabetes Mellitus

ESOPHAGEAL REFLUX [530.81]
GERD

DYSPHAGIA, UNSPECIFIED [787.20]
dysphagia

UNSPECIFIED ESSENTIAL HYPERTENSION [401.9]
Hypertension

ICD9 Changes Log

CPT Codes □ Disagree Comment Other Comments

CPT Changes Log

Medical Record Info :

History Section

HPI

57 yo female presents to the clinic for initial visit after recently moving from Seattle, WA.Ms Kohler is here for assistance with her blood sugars and also reports having the diagnosis of Atrial Fibrillation, Hypertension and Hypercholesterolemia.She was last seen by h primary care provider Dr Scott Jones 3 months ago and Cardiologist Dr Rick Wu 6 months d can't remember all of her medical problems stating that they were in her doctor 's computers.

Atrial fibrillation (disorder) [49436004] [K]--49436004--
POSITIVE

The patient symptoms started 9 years ago with the onset of polyuria, polydipsia, and fatigue and she was seen by her internist who suspected Diabetes Mellitus. Efforts have been made to control the diabetes using diet and oral agents.The diet, the patient admittedly states, has not been well controlled but feels that in recent times, she has had a greater success in her dietary efforts.She is minimizing the sweets that she eats other than for fresh fruit.She also has been reducing the amount of protein in the form of red meat and eating more salads and vegetables than in the past.The patient also exercises 2- 4 x/ week, working out both in the gym and by walking regularly. Efforts have been made to control the patient with various oral agents , and she presently is receiving several which will be listed under medications.The patient has been monitoring her blood glucose on a daily basis and recognizes that she has had a
erformed by Dr. Jones.Though she cannot recall at the moment
es were, she does admit they need improvement.She feels that
er control in recent times has improved.The patient has rare
which are characterized, when they occur, by weakness.They
he morning , and she never has nocturnal kinds of episodes of
nt has had no chronic complications of her diabetes, though
d toes tingle from time to time.

orsening GERD symptoms with 2 to 3 week history of intermittent nausea and vomiting. Denies odynophagia.She notes some bitter taste in the back of her throat with increased gastric secretions, scant blood noted once but no recurrence since.She denies any fevers, chills or abdominal pain associated with this.She has tried Zantac with minimal to no relief.

4. Negation (red) and uncertainty (green) detection

E/M assigned by nCode

Code estimated by MD

Phrase recognized and assigned SNOMED code

Narrative text—dictation, directly entered, or combination—is run through engine to tag SNOMED phrases. CMS algorithm used to assign E/M code suported by the note.

Challenges in clinical NLP

Component	Problems	Examples
Named entity recognition	• Linguistic variation—different words with same meaning • Polysemy—one word with multiple meanings • Finding validation • Implication	APC: Activated protein C, adenomatosis polyposis coli, atrial premature complex
Contextual attribute assignment	• Negation • Uncertainty • Temporality	The mediastinum is not widened. Treated for presumptive sinusitis.
Discourse processing	• Report structure • Coreference	Cardiovascular: [ ] Angina [ ] MI [x] HTN [ ] CHF [ ] PVD [ ] DVT [ ] Arrhythmias [ ] Previous PTCA [ ] Previous Cardiac Surgery [ ] Negative - Denies CV problems

Key Readings

Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. doi: 10.1136/amiajnl-2011-000464. PMID: 21846786; PMCID: PMC3168328.

Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 Jun 1;2(6):797-804. doi: 10.1001/jamaoncol.2016.0213. PMID: 27124593.

K107 Natural Language Processing References

Friedman C. Semantic text parsing for patient records. In: Chen H, Fuller SS, Friedman C, Hersh W. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer, 2005.

Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearb Med Inform. 2020 Aug;29(1):208-220. doi: 10.1055/s-0040-1702001. Epub 2020 Aug 21. PMID: 32823318; PMCID: PMC7442512. [Article]

Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. doi: 10.1136/amiajnl-2011-000464. PMID: 21846786; PMCID: PMC3168328. [Article]

Payne TH, Garver-Hume A, Kirkegaard S, Sweeney J, Ash M, Kailasam KK, Hall CL, Sinanan MN. Group improves coding with natural language processing. MGMA Connex. 2011 Oct;11(9): 15-7. PMID: 22375458

Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 Jun 1;2(6):797-804. doi: 10.1001/jamaoncol.2016.0213. PMID: 27124593. [Abstract]