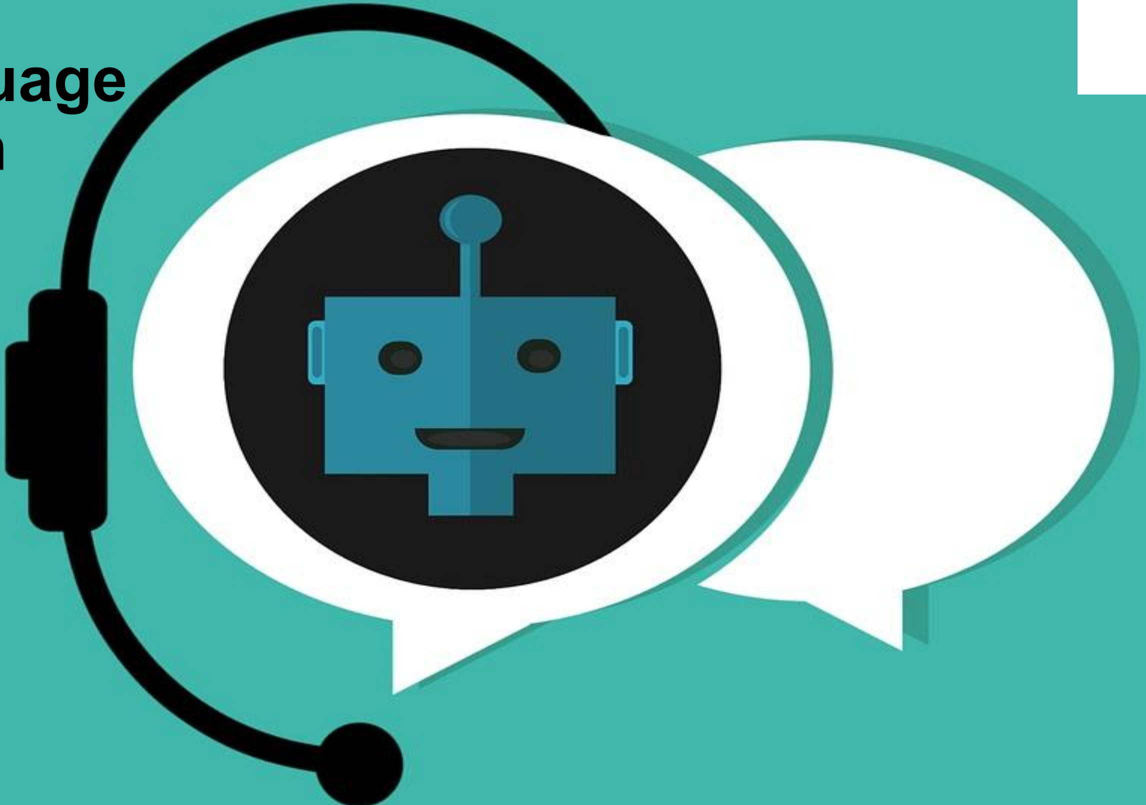
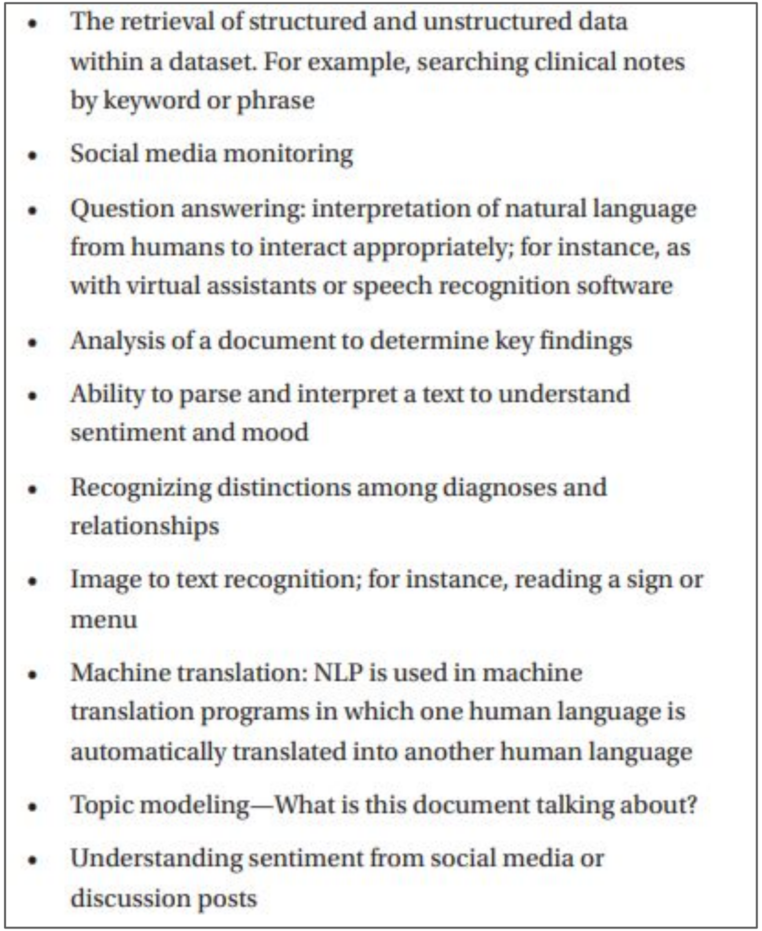
Module 4

Natural Language Processing in Healthcare



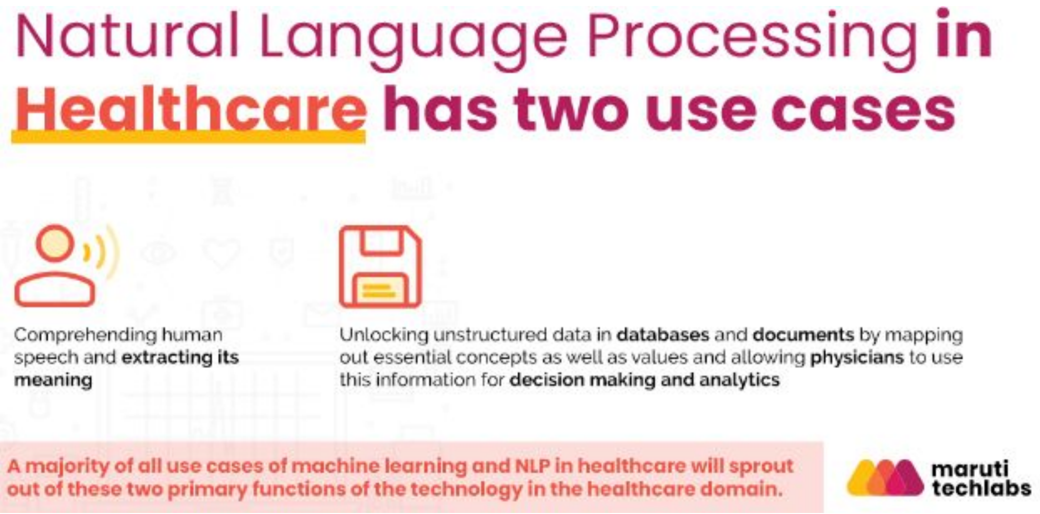
4.1 NLP tasks in Medicine, Low-level NLP components, High level NLP components, NLP Methods.
4.2 Clinical NLP resources and Tools, NLP Applications in Healthcare. Model Interpretability using Explainable AI for NLP applications. link

NLP tasks in Medicine



- The retrieval of structured and unstructured data within a dataset. For example, searching clinical notes by keyword or phrase
- Social media monitoring
- Question answering: interpretation of natural language from humans to interact appropriately; for instance, as with virtual assistants or speech recognition software
- Analysis of a document to determine key findings
- Ability to parse and interpret a text to understand sentiment and mood
- Recognizing distinctions among diagnoses and relationships
- Image to text recognition; for instance, reading a sign or menu
- Machine translation: NLP is used in machine translation programs in which one human language is automatically translated into another human language
- Topic modeling—What is this document talking about?
- Understanding sentiment from social media or discussion posts

NLP is an aspect of computational linguistics (studying linguistics using computer science)

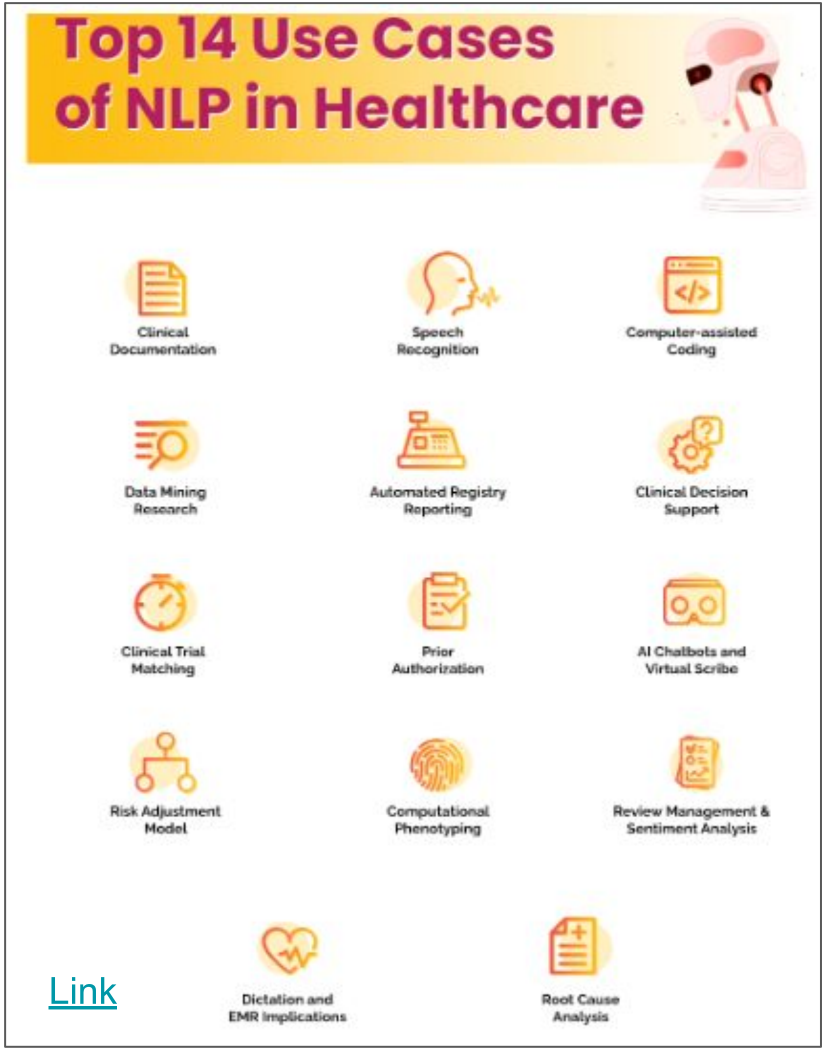


Natural Language Processing **in Healthcare has two use cases**

Comprehending human speech and **extracting its meaning**

Unlocking unstructured data in **databases** and **documents** by mapping out essential concepts as well as values and allowing **physicians** to use this information for **decision making and analytics**

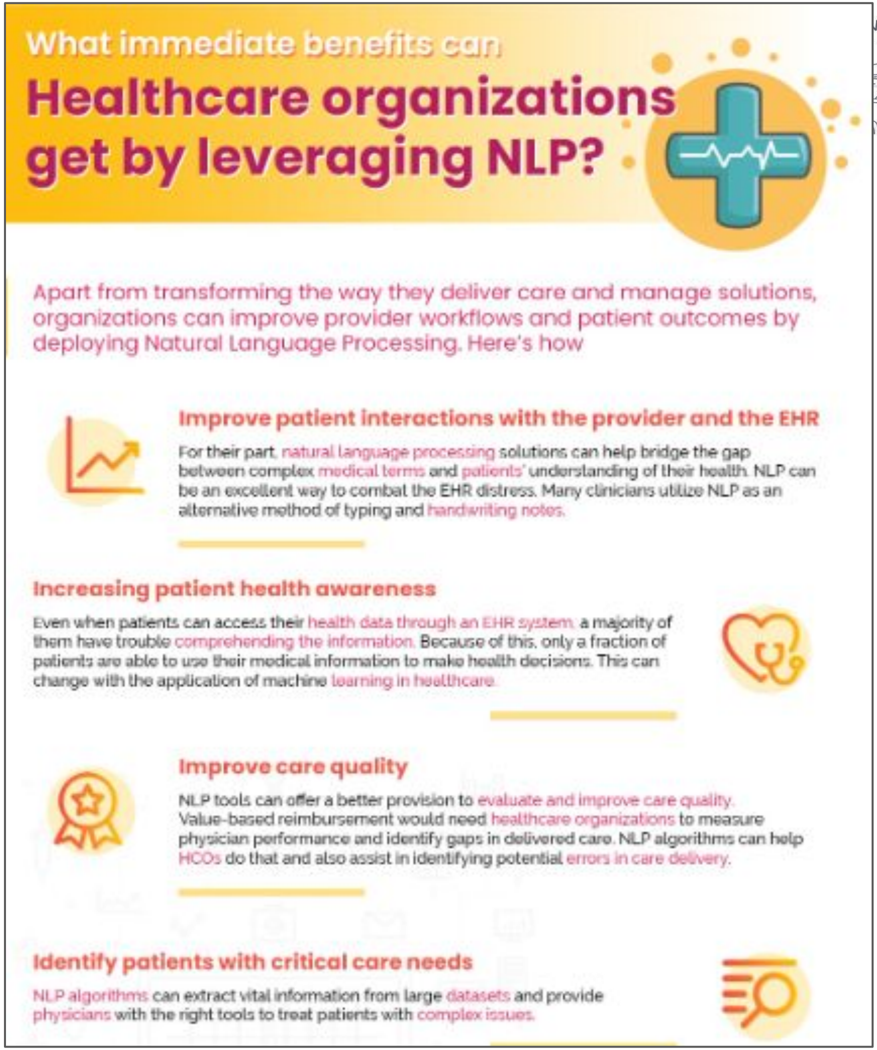
A majority of all use cases of machine learning and NLP in healthcare will sprout out of these two primary functions of the technology in the healthcare domain.



Top 14 Use Cases of NLP in Healthcare

- Clinical Documentation
- Speech Recognition
- Computer-assisted Coding
- Data Mining Research
- Automated Registry Reporting
- Clinical Decision Support
- Clinical Trial Matching
- Prior Authorization
- AI Chatbots and Virtual Scribe
- Risk Adjustment Model
- Computational Phenotyping
- Review Management & Sentiment Analysis
- Dictation and EMR Implications
- Root Cause Analysis

[Link]



What immediate benefits can

Healthcare organizations get by leveraging NLP?

Apart from transforming the way they deliver care and manage solutions, organizations can improve provider workflows and patient outcomes by deploying Natural Language Processing. Here's how

Improve patient interactions with the provider and the EHR

For their part, natural language processing solutions can help bridge the gap between complex medical terms and patients' understanding of their health. NLP can be an excellent way to combat the EHR distress. Many clinicians utilize NLP as an alternative method of typing and handwriting notes.

Increasing patient health awareness

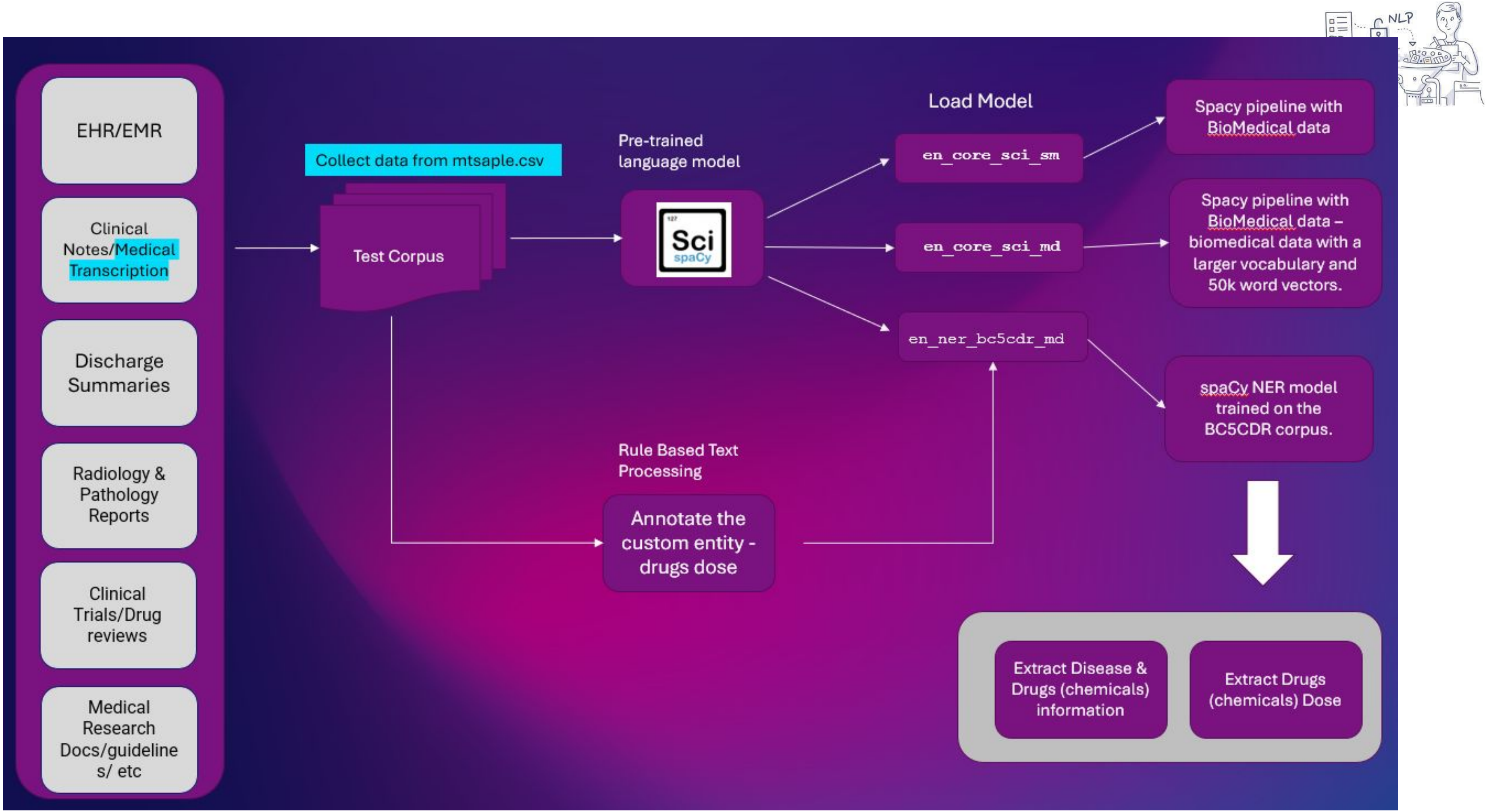
Even when patients can access their health data through an EHR system, a majority of them have trouble comprehending the information. Because of this, only a fraction of patients are able to use their medical information to make health decisions. This can change with the application of machine learning in healthcare.

Improve care quality

NLP tools can offer a better provision to evaluate and improve care quality. Value-based reimbursement would need healthcare organizations to measure physician performance and identify gaps in delivered care. NLP algorithms can help HCOs do that and also assist in identifying potential errors in care delivery.

Identify patients with critical care needs

NLP algorithms can extract vital information from large datasets and provide physicians with the right tools to treat patients with complex issues.



EHR/EMR

Clinical Notes/Medical Transcription

Discharge Summaries

Radiology & Pathology Reports

Clinical Trials/Drug reviews

Medical Research Docs/guidelines/ etc

Collect data from mtsaple.csv

Test Corpus

Pre-trained language model

Sci spaCy

Load Model

`en_core_sci_sm`

`en_core_sci_md`

`en_ner_bc5cdr_md`

Spacy pipeline with BioMedical data

Spacy pipeline with BioMedical data – biomedical data with a larger vocabulary and 50k word vectors.

spaCy NER model trained on the BC5CDR corpus.

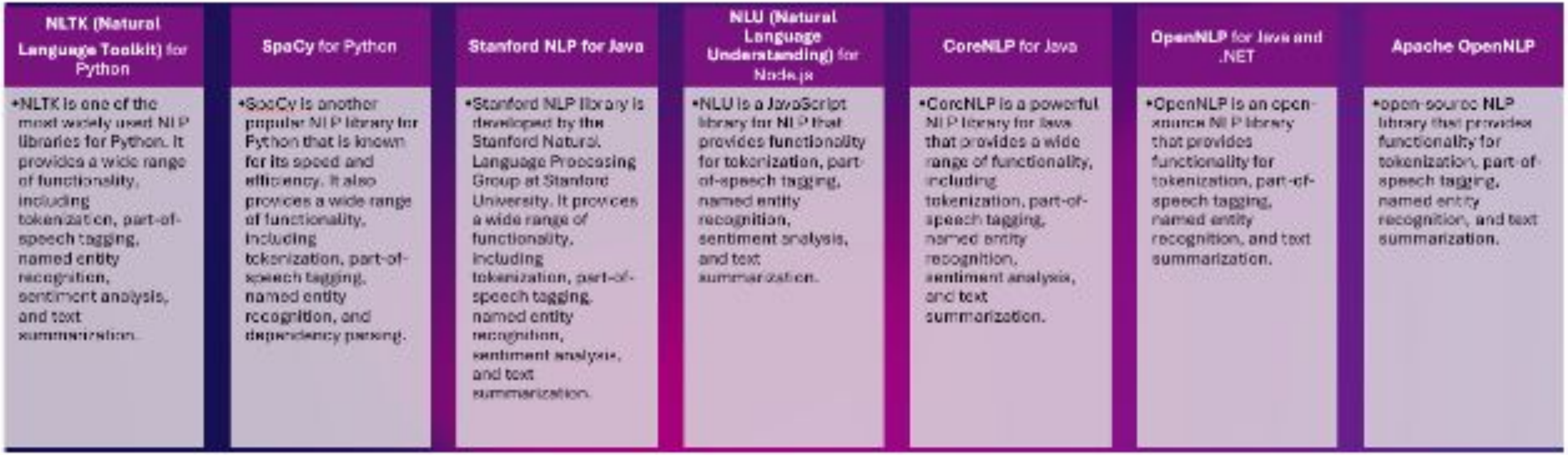
Rule Based Text Processing

Annotate the custom entity - drugs dose

Extract Disease & Drugs (chemicals) information

Extract Drugs (chemicals) Dose

Various libraries provide a wide range of NLP functionalities. Such as :



NLTK (Natural Language Toolkit) for Python	**SpaCy** for Python	**Stanford NLP for Java**	**NLU (Natural Language Understanding)** for Node.js	**CoreNLP** for Java	**OpenNLP** for Java and .NET	**Apache OpenNLP**
•NLTK is one of the most widely used NLP libraries for Python. It provides a wide range of functionality, including tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and text summarization.	•SpaCy is another popular NLP library for Python that is known for its speed and efficiency. It also provides a wide range of functionality, including tokenization, part-of-speech tagging, named entity recognition, and dependency parsing.	•Stanford NLP library is developed by the Stanford Natural Language Processing Group at Stanford University. It provides a wide range of functionality, including tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and text summarization.	•NLU is a JavaScript library for NLP that provides functionality for tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and text summarization.	•CoreNLP is a powerful NLP library for Java that provides a wide range of functionality, including tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and text summarization.	•OpenNLP is an open-source NLP library that provides functionality for tokenization, part-of-speech tagging, named entity recognition, and text summarization.	•open-source NLP library that provides functionality for tokenization, part-of-speech tagging, named entity recognition, and text summarization.

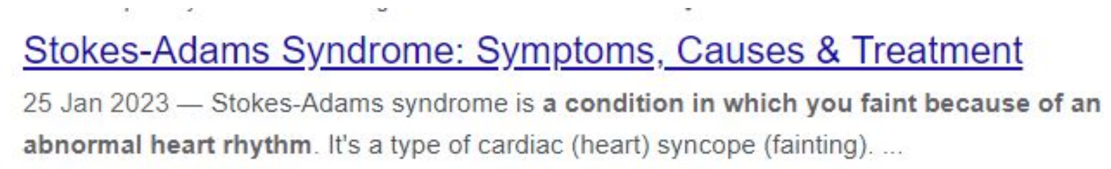
Low-level NLP components

Tokenization is an initial step of automated processing of a text. It is the task of identifying **boundaries that separate semantic units, which include morphemes, words, dates, and symbols within a text.** The primary indication of such semantic units, also called tokens, in general English is **white space** that occurs before and after a word.

A token may also be separated by punctuation marks instead of a word space, such as by a period, comma, semicolon, or question mark.

Some of the **difficulties** that occur with tokenization stem from **ambiguous punctuation,** such as the colon in **"2:30am" or the periods in "M.D."**

The biomedical literature will also have certain **technical terms and heterogeneous orthographics**, such as **"Adams Stokes" and "Adams-Stokes,"** which add additional difficulties in tokenization.



Stokes-Adams Syndrome: Symptoms, Causes & Treatment

25 Jan 2023 — Stokes-Adams syndrome is **a condition in which you faint because of an abnormal heart rhythm**. It's a type of cardiac (heart) syncope (fainting). ...

Tokenization

For this reason, a simple tokenizer for general English text will typically not work well in biomedical text.

Therefore, tokenization algorithms often **need new heuristics and domain-specific training corpora** to accommodate the distinct features of medical sublanguages.

Medical corpus

https://link.springer.com/article/10.1007/s10579-022-09596-2

https://www.sketchengine.eu/guide/text-type-analysis/

https://github.com/adahealth/medical_case_report_corpus/blob/master/docs/Medical_Entities_in_Case_Reports.pdf

https://github.com/adahealth/medical_case_report_corpus/tree/master

Low-level NLP components

Sentence boundary detection (SBD), also called sentence boundary disambiguation or *sentence breaking*, can also be a challenging NLP component, particularly for clinical documents. This task aims to detect where sentences start and end.

A simple SBD system can identify sentence boundaries using a small set of rules. However, t**he task can be complicated by the fact that punctuation marks such as question marks, semicolons, and periods are often ambiguous and need more complex logic in special cases.**

In addition to rulebased systems, AI methods such as decision trees, neural networks, and hidden Markov models (HMMs) are frequently used for SBD.

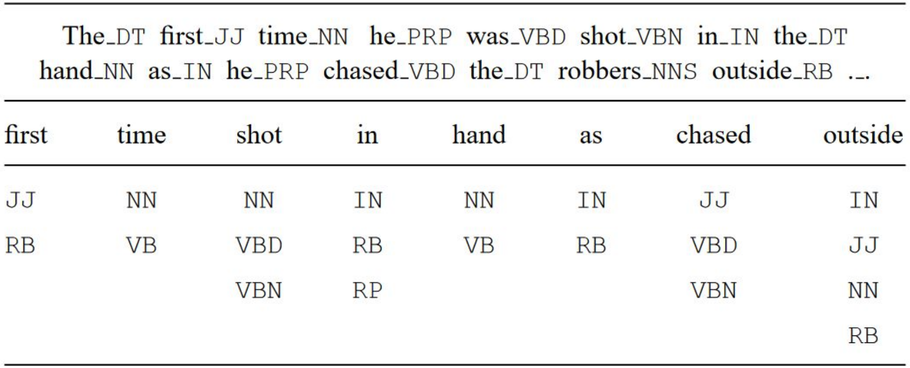
Also, the **biomedical literature and clinical documents are full of abbreviations (e.g., “q.i.d.,” “p.r.n.”), acronyms (e.g., “OD,” “OS”), and symbolic constructions (e.g., “blood pressure: 130/67”) that add difficulty to SBD** https://www.nature.com/articles/s41467-022-35007-9

For medical NLP systems, one frequent approach for SBD includes the use of **domain lexical resources such as the National Library of Medicine’s (NLM’s) SPECIALIST Lexicon (McCray et al. 1994) and annotated domain corpora to ensure satisfactory SBD performance**

Low-level NLP components(LINK)

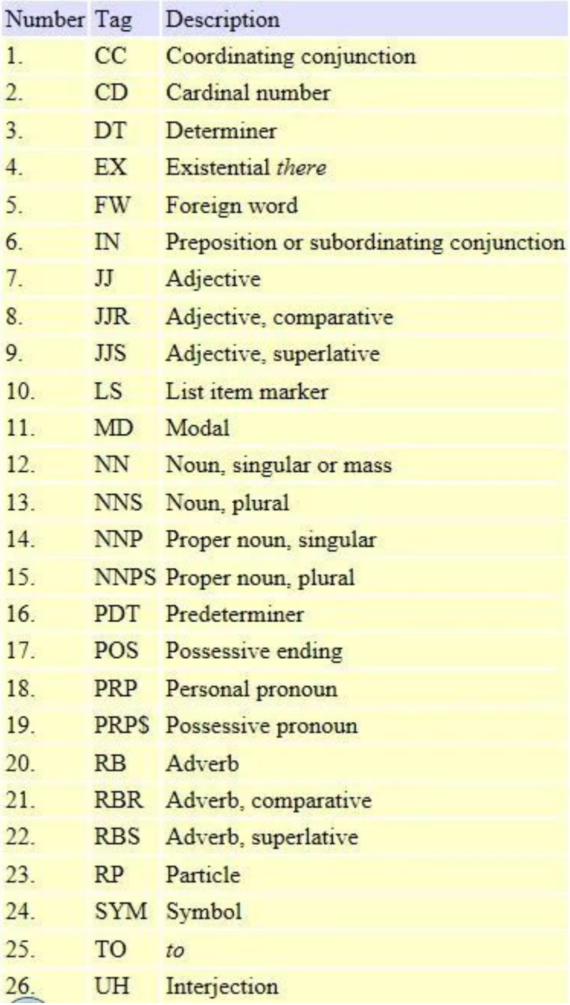
Part-of-speech (POS) tagging is the process for **determining the part of speech of words in a piece of text, based on both definition as well as local context.**
The example below shows the tagging output of the following sentence using the Penn Treebank tag set
"The cystic duct was triply clipped distally and singly proximally and transected."
"The/DT cystic/JJ duct/NN was/VBD triply/RB clipped/VBN distally/RB and/CC singly/ RB proximally/RB and/CC transected/VBN."
POS tagging is an essential step of NLP systems **where errors can propagate upward to the syntactic processing level and produce more errors in the syntactic output**, which provides important information necessary for text understanding. Therefore, having **reliable POS information is critical to successful implementation of various NLP applications.**
POS taggers trained merely on general English **do not usually achieve state-of-the-art performance on medical tex**t.
A number of POS taggers have been developed specifically for the medical domain, such as **the adapted Trigrams'n'Tags (TnT) tagger which is a TnT tagger (http://www.coli.uni-saarland.de/~thorsten/tnt/) trained on a relatively small set of clinical notes, and the MedPost tagger (Smith et al. 2004), a POS tagger based on an HMM and trained on manually tagged sentences in medical text**

Low-level NLP components



The_DT first_JJ time_NN he_PRP was_VBD shot_VBN in_IN the_DT hand_NN as_IN he_PRP chased_VBD the_DT robbers_NNS outside_RB ._.

first	time	shot	in	hand	as	chased	outside
JJ	NN	NN	IN	NN	IN	JJ	IN
RB	VB	VBD	RB	VB	RB	VBD	JJ
		VBN	RP			VBN	NN
							RB



Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection

Low-level NLP components

Shallow parsing, also called chunking, is the process of identifying constituents (**syntactically correlated parts of words like noun groups, verb groups, etc.)** in a sentence.

As an intermediate step toward deep parsing, shallow parsing produces a **limited** amount of syntactic information from sentences and does not specify internal structures or roles of each constituent in the main sentence.

The sentence below exemplifies shallow parsing output:

"[NP The cystic duct] [VP was triply clipped] [ADVP distally and singly] [ADVP proximally] and [UCP transected]"

In the medical domain, shallow parsing is used in a wide range of tasks such as **drug– drug interaction (DDI) detection, medical problem assertion detection, biological entity relation extraction, and medical information extraction (IE).**

Several shallow parsers have been built for medical text processing, such as the **SPECIALIST minimal commitment parser (McCray et al. 1993), which produces high-level syntactic information rather than the traditional full syntactic information for better noun phrase discovery in medical text.**

Low-level NLP components(link)

Deep parsing is the process **to produce an ordered, rooted tree** that represents the **syntactic structure of a string according to some formal grammar such as constituency grammars and dependency grammars**

Full syntactic parsing of text can provide a large amount of **deep linguistic information such as sentence voice, phrase type, and POS tags**, which are shown to perform considerably better than **surface-oriented features (e.g., pattern matching) for many NLP tasks.**

Because of the special features of medical sublanguage (e.g., **domain vocabulary, telegraphic text, special grammar),** parsers trained on general English corpus like the Wall Street Journal only have limited performance on medical text.

New entries can be imported from domain resources to **existing parser lexicons using morphological clues, heuristic mapping, and direct expansion (Szolovits 2003). POS tag information of domain-specific lexical elements can also be provided to a parser to avoid inconsistencies between domain POS tags and parser lexicon POS tags (Rimell and Clark 2009)**.

Moreover, better parsing performance can also be acquired by adjusting the syntactical category statistics for important domain lexical elements like **verbs and other lexical elements that have unusual usage in a particular domain (Huang et al. 2005).**

Low-level NLP components

VBD: Verb, past tense. VBG: Verb, gerund or present participle. VBN: Verb, past participle. PRP. Personal pronoun (PP)

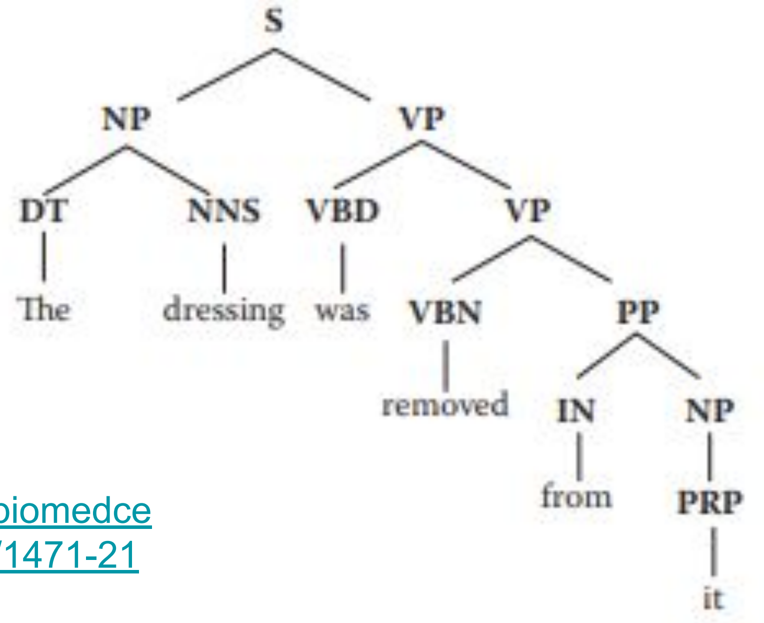


FIGURE 23.1
A constituent (phrase structure) tree for "The dressing was removed from it."

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S3-S2

Shallow Parsing (Chunking)
Shallow parsing identifies **chunks** of phrases (like noun phrases, verb phrases) but does not show the complete hierarchical or dependency structure.
Output (chunks):

- [NP The doctor]
- [VP treated]
- [NP the patient]
- [PP with [NP antibiotics]]
- [PP for [NP pneumonia]]

We only know that “The doctor” is a noun phrase, “treated” is a verb, etc., but we don’t know **who treated whom**.

Deep Parsing (Full Parsing)
Deep parsing builds a **full syntactic structure** (dependency or constituency tree) and explicitly shows relationships.
Dependency parse (simplified):

- **treated** (root)

doctor → subject (who did the action)
patient → object (who was treated)
antibiotics → instrument (used for treatment)
pneumonia → reason (why treatment given)
we know **doctor** → **treated** → **patient** using **antibiotics** for **pneumonia**.

High level NLP components

- **Negation Detection:** Many medical documents such as **discharge summaries and radiology reports contain large amounts of important information of patients, like conditions, findings, and diseases, that can be used for a wide range of secondary applications.**

In these reports, about half of the described findings and diseases are actually absent in a given patient.

For example: **“They have not noticed any abnormal behaviors, movements, or rash anywhere else on his body” or “no significant complications of bleeding.”** As a result, negation detection is a critical component in medical NLP systems.

In medical text, negation detection is not an easy task as negation can be **explicit (e.g., “Patient denies any fevers, emesis, or diarrhea”) or implied (e.g., “Chest x-ray is clear upon my read”).** The scope of negation is another challenge for the negation detection process. Consider two sentences: **“The child is not tired” and “The child is not very tired.”** In the first sentence, the word “not” scopes over “tired,” while in the second sentence, the word “very” redirects the scope of “not” to itself and away from “tired.**” The patient is free of cancer(**Output: **Cancer** → **Negated) The patient has asthma but no history of diabetes.”** Concept: *asthma* → Affirmed, Concept: *diabetes* → Negated (trigger = *no history of*)

High level NLP

- **Relation extraction** aims to determine or discover relationships between **entities (e.g., drugs, diseases, findings, genes) in medical texts.** Relations among these entities, in their simplest form, are binary, involving only two entities.

A large variety of relations have been investigated, such as interactions between drugs, genes, associations between diseases and symptoms, and relations between patient problems and treatments.

The hypothesis behind this approach is that an entity and its related entities are more likely to appear together than random combinations of entities. Thus, if entities are repeatedly mentioned together, then there is a good chance that they may be related. **Rule-based approaches for relation extraction work by exploiting the particular linguistic patterns exhibited by relations**. Rules used can be manually defined by domain experts or derived from annotated corpora. Machine learning-based systems rely on machine learning techniques along with a variety of features based on the nature of the relationship, such as lexical, syntactic, semantic, and dependency features. Several important **challenges** are associated with relation extraction in the medical domain. First, in the medical domain, annotation of relations can be complicated because **relations are often expressed across discontinuous spans of text**. Secondly, there can be **a lack of consensus on how to best annotate a particular type of relation**. As a result, annotation resources between research groups can be largely incompatible and the quality of systems constructed based upon these resources can be difficult to evaluate.

High level NLP

- Named entity recognition (NER) aims to identify and classify elements into named entities, which are predefined categories such as **names (e.g., drugs, genes, person), findings, diseases, and medications.**

The biomedical literature is full of terms particular to the biomedical domain that are typically **not detected by conventional general English NLP systems**.

In order to extract relations between entities, **it is crucial for the system to be able to detect unknown nouns or named entities.** Some named entities can be effectively identified solely through **surface patterns** (e.g., phone number: xxx-xxx-xxxx, person: Carole Green MD).

Machine learning is another effective approach for NER. Compared with rule-based systems, it requires **less human intuition with rules,** and this approach can easily be adapted to new domains. The disadvantage of machine learning approaches is the **large annotation corpus required for model training**

https://medium.com/@sowmyavivek/named-entity-recognition-ner-to-extract-enrich-content-from-the-healthcare-ecosystem-a5adcc7f7e2d

High level NLP

- **Word Sense Disambiguation** Ambiguity is a problem inherent to natural language, where a term can have more than one meaning depending upon the context or use of the term in a particular text.

It is the process of understanding which sense of a term, including **single words, abbreviations, or acronyms, is being used in a particular context a**mong a list of predefined sense candidates.

In the medical domain, researchers have suggested that the problem might be more restricted compared to general English based upon the idea that since medicine is scientific, it might be more specific than general English. Instead, the problem is more extensive in the medical domain due to the high use of abbreviations and acronyms in medical documents and the biomedical literature.

- **Semantic role labeling (SRL)** is the task of detecting **semantic roles associated with predicates, which are mainly verbs, such as “hit” and “move,” in a sentence.**

For example, in the sentence **“He placed the ball beside the couch,” the predicate is the verb “place.” The semantic roles associated with “place” include “placer”—who placed; “thing placed”— what is placed; and “location”—where it (the ball) is placed.**

contd…

High level NLP

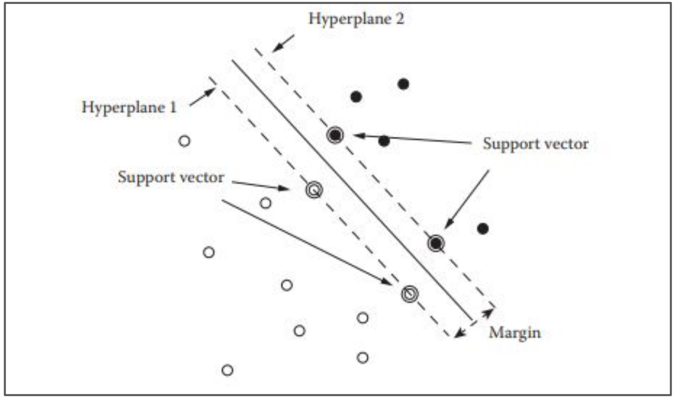
Labeling semantic roles like above for predicates in a given text answers questions such as **“who,” “when,” “what,” “where,” and “why.” SRL can be used for IE, question answering (QA), text summarization, and other NLP tasks that require some kind of semantic interpretation.**SRL USES
Extracts structured meaning from clinical notes (who did what, to whom, for what condition).
Supports **clinical decision support systems** (linking treatment to diagnosis).
Improves **EHR information retrieval** by capturing relationships, not just keywords.

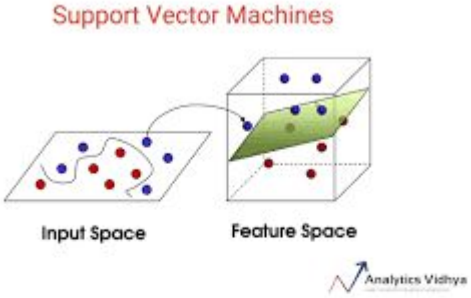
- **IE(Information Extraction)** is a task that involves **extracting problem-specific information from the text of interest and then transforming this information into structured form.**

For example, **vaccination reactions can be extracted from medical reports, and relationships between genes and diseases from the biomedical literature are all cases of IE.**
Most early and straightforward IE systems were built mostly using pattern matching techniques such as regular expressions over **features such as text strings, syntactic structure, semantic type, and dictionary entries**

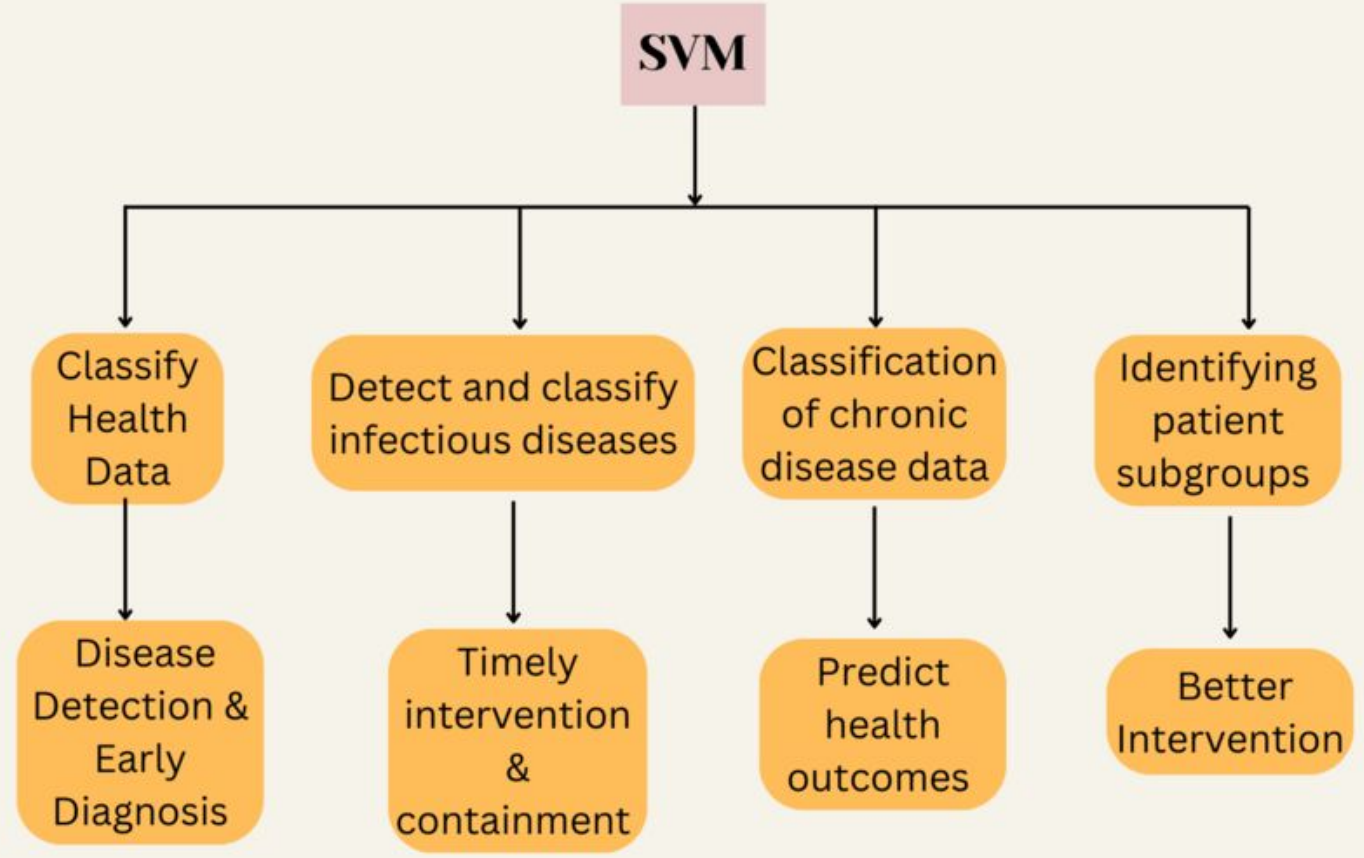
NLP Methods:SVM

The prediction accuracy of **SVM** is generally high because of the sound mathematical theory behind it and the robustness of the method. It generally works well when training examples contain errors, as well, because of its use of a separation process. On the downside, SVM is computationally expensive. Its training process is a convex optimization problem that requires at least quadratic time with respect to the number of training examples. In the medical domain, SVM has been shown to perform well on many classification tasks such as smoking status classification (Cohen 2008), SRL for protein transport predicates (Bethard et al. 2008), and disease comorbidity status classification (Ambert and Cohen 2009)





Applications of SVM in Public Health Data Analysis



SVM

Classify Health Data

Disease Detection & Early Diagnosis

Detect and classify infectious diseases

Timely intervention & containment

Classification of chronic disease data

Predict health outcomes

Identifying patient subgroups

Better Intervention

NLP Methods: Maximum Entropy Modeling

Maximum Entropy (MaxEnt) is a **probabilistic classification method** often used in NLP tasks like part-of-speech tagging, text classification, machine translation, and named entity recognition.

It is based on the **Maximum Entropy Principle**:

> *When estimating a probability distribution, choose the one with the **highest entropy** (most uniform / least biased), subject to the known constraints.*

This ensures that we don’t make any assumptions beyond what is supported by the training data.

NLP Methods: Maximum Entropy Modeling

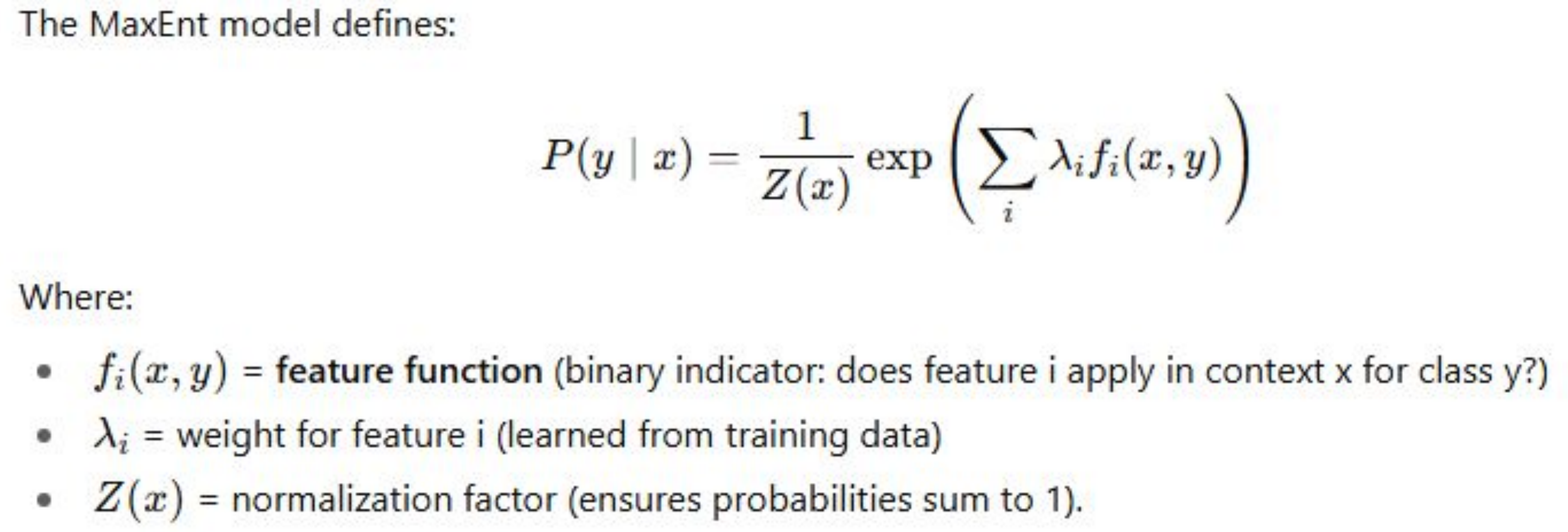
Why “Maximum Entropy”?

- Entropy measures **uncertainty** in a probability distribution.
- If multiple probability distributions satisfy the training constraints, we prefer the one with **maximum entropy** → meaning we are making the least amount of assumptions.

In simple words: *“Don’t assume anything extra about the data. Stay as unbiased as possible while still fitting what we know.”*

NLP Methods: Maximum Entropy Modeling

Suppose we want to predict **class y** (e.g., a POS tag, a topic label) given **context x** (e.g., a word and its features).



The MaxEnt model defines:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x, y)\right)$$

Where:

- $f_i(x, y)$ = **feature function** (binary indicator: does feature i apply in context x for class y?)
- λ_i = weight for feature i (learned from training data)
- $Z(x)$ = normalization factor (ensures probabilities sum to 1).

NLP Methods: Maximum Entropy Modeling

Example (POS Tagging)

Imagine we want to classify the word *“run”* into a POS tag: {Verb, Noun}.

Features could be:

- $f_1(x,y)$ = 1 if y = Verb and previous word is “to”.
- $f_2(x,y)$ = 1 if y = Noun and previous word is “the”.
- $f_3(x,y)$ = 1 if word ends with “ing” and y = Verb.

The MaxEnt model will combine these features with weights to compute probabilities like:

- P(Verb | “to run”) = 0.92
- P(Noun | “to run”) = 0.08

n-Gram Model

Statistical language modeling (SLM) is widely used for many NLP tasks, such as **POS tagging, parsing, information retrieval, and machine translation**.

SLM assigns a probability to a set of n words based on a probability distribution.

An n-gram model is a typical language method used in the field of computational linguistics and NLP (Manning and Schütze 2003). The word “n-gram” means consecutive items, such as words or terms. n-gram models (n = 2, 3, 4 …) are used to estimate the probability of the existence of the n-gram.

To simplify the calculation of the probability of the word, the Markov assumption states that the probability of the word is only based on the prior few words instead of all previous words. The probability of a word is then simplified as follows: An n-gram model (which checks the n-1 previous words) is an (n-1)th order Markov model

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

HMM

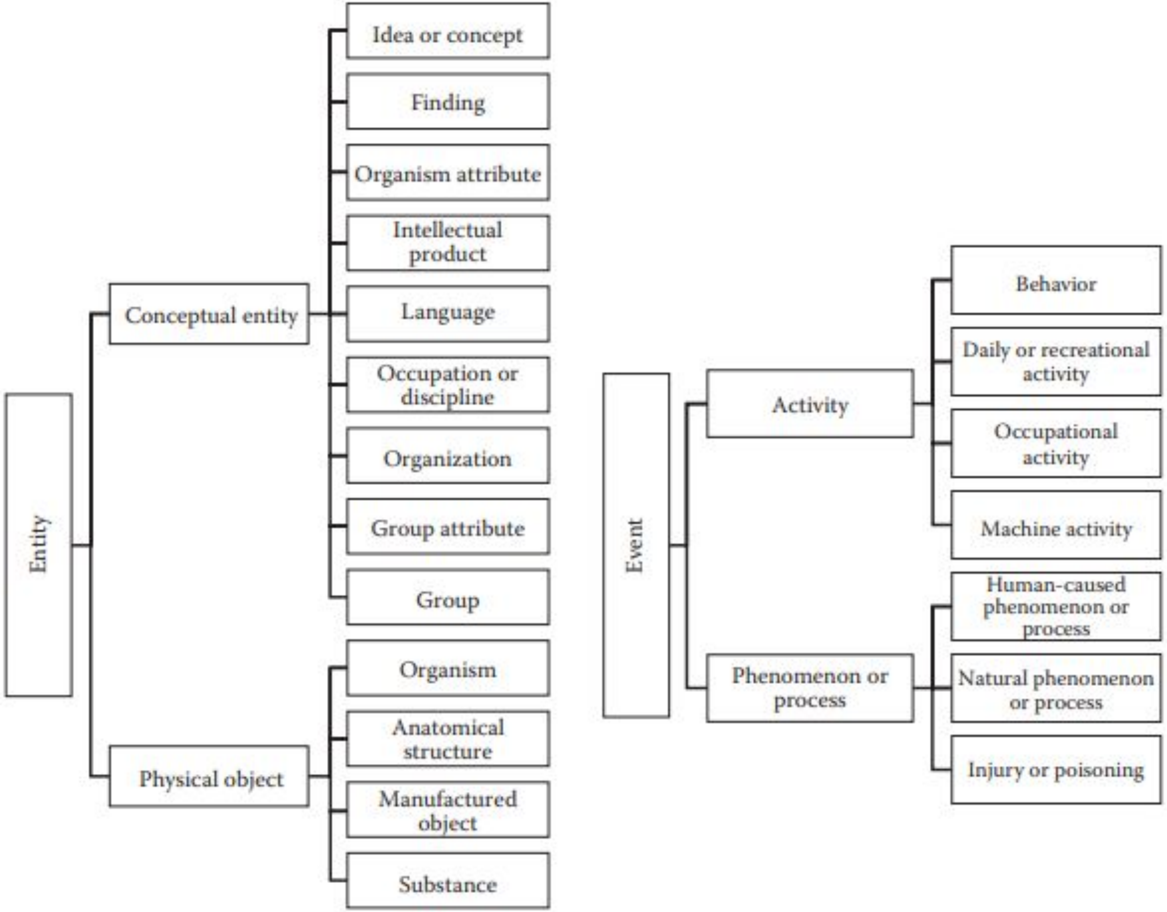
Markov models are built on the Markov assumption that the current state occurs based upon on the previous state(s).

For the simplest first-order Markov model, there are M square transitions between M states. Unlike deterministic models, where each state is dependent on another state, Markov models assign probability to each transition between two states.

In a visible Markov model, the state is visible, and state transition probabilities are the only parameters to calculate. In an HMM, hidden states have a probability contribution to the outputs. **For example, in a speech recognition system, the sound we hear is the output of hidden states, such as vocal chords, the size of the person's throat, the position of the person's tongue, and many other factors. Each sound of a word is generated from changes of these hidden factors.**

Clinical NLP Resources and Tools

UMLS (http://www.nlm.nih.gov/research/umls/) was developed by and is maintained by the NLM to provide health care professionals and researchers with a biomedical domain knowledge resource (Humphreys et al. 1998). **UMLS is a structured knowledge base that connects different biomedical sources and enables biomedical research application development.** UMLS contains three knowledge sources: Metathesaurus, Semantic Network (McCray 2003), and SPECIALIST Lexicon (McCray et al. 1994) and lexical tools



- Entity
  - Conceptual entity
    - Idea or concept
    - Finding
    - Organism attribute
    - Intellectual product
    - Language
    - Occupation or discipline
    - Organization
    - Group attribute
    - Group
  - Physical object
    - Organism
    - Anatomical structure
    - Manufactured object
    - Substance
- Event
  - Activity
    - Behavior
    - Daily or recreational activity
    - Occupational activity
    - Machine activity
  - Phenomenon or process
    - Human-caused phenomenon or process
    - Natural phenomenon or process
    - Injury or poisoning

FIGURE 23.4
Hierarchy structure of UMLS semantic types.

NLP tools

The **Semantic Network** is an upper-level ontology of biomedical knowledge.

It contains **135 semantic types and 54 relationships between semantic types**. Each concept is assigned at least 1 of 135 defined semantic types. All semantic types are hierarchically organized under two main topics: **Entity and Event**. The top level of semantic types in the UMLS Semantic Network is depicted in Figure 23.4.

The relationship "ISA" is used to link most concepts.

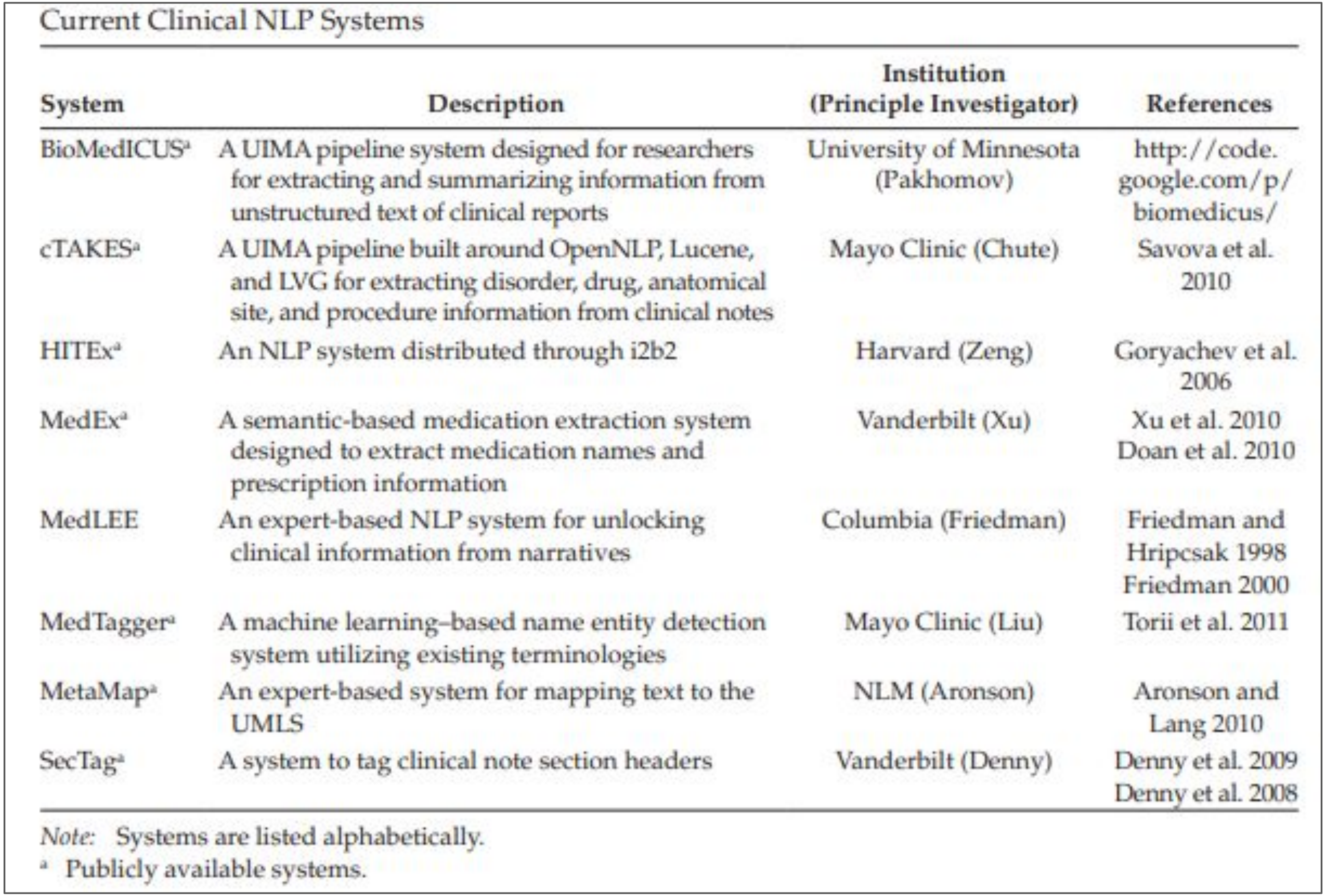
For example, "Carbohydrate" ISA "Chemical." There are also five major, nonhierarchical relationships: physical (e.g., PART_OF, BRANCH_OF); spatial (e.g., LOCATION_OF, ADJACENT_TO); temporal (e.g., CO-OCCURS_ TO, PRECEDES); functional (e.g., TREATS, CAUSES); and conceptual (e.g., EVALUATION_ OF, DIAGNOSES). Entity Event Conceptual entity Activity Phenomenon or process Injury or poisoning Natural phenomenon or process Human-caused phenomenon or process Machine activity Occupational activity Daily or recreational activity Behavior Idea o

The **MEDLINE database** is a collection of biomedical abstracts. It is maintained by the NLM and contains over 21 million reference from 1946 to the present.

The **GENIA corpus** (http:// www.nactem.ac.uk/genia/genia-corpus) collects 1999 MEDLINE abstracts, selected from a PubMed query for MeSH terms “human,” “blood cells,” and “transcription factors.” The corpus has been annotated with various levels of linguistic and semantic information covering POS, syntactic, term, event, relation, and coreference annotation (Kim et al. 2003).

SPECIALIST NLP tools (http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html) are computer programs developed by the NLM to aid in dealing with different biomedical NLP tasks. Tools include lexical tools such as **lexical variant generator (LVG)**, normalized string generator (Norm), word index generator (WordInd), dTagger POS tagger, subterm mapping tools (STMTs), and others. LVG contains a series of commands to perform lexical transformation of text. Norm provides a normalization process for those terms included in the SPECIALIST Lexicon.

MetaMap (http://metamap.nlm.nih.gov/) is a program developed by the NLM to map biomedical text to the UMLS Metathesaurus (Aronson 2001; Aronson and Lang 2010). MetaMap provides various options, including data option (choose specific vocabularies and data model); processing options (such as author-defined acronyms/abbreviations, negation detection, WSD:Word Sense Disambiguation) and output options (human readable, machine output, and XML). Released application programming interfaces (APIs) provide options to integrate MetaMap into other programs. **SemRep** is a rule-based, symbolic NLP program developed by NLM for semantic knowledge representation from biomedical literatures, mainly from titles and abstracts in MEDLINE



Current Clinical NLP Systems

System	Description	Institution (Principle Investigator)	References
BioMedICUS[a]	A UIMA pipeline system designed for researchers for extracting and summarizing information from unstructured text of clinical reports	University of Minnesota (Pakhomov)	http://code.google.com/p/biomedicus/
cTAKES[a]	A UIMA pipeline built around OpenNLP, Lucene, and LVG for extracting disorder, drug, anatomical site, and procedure information from clinical notes	Mayo Clinic (Chute)	Savova et al. 2010
HITEx[a]	An NLP system distributed through i2b2	Harvard (Zeng)	Goryachev et al. 2006
MedEx[a]	A semantic-based medication extraction system designed to extract medication names and prescription information	Vanderbilt (Xu)	Xu et al. 2010 Doan et al. 2010
MedLEE	An expert-based NLP system for unlocking clinical information from narratives	Columbia (Friedman)	Friedman and Hripcsak 1998 Friedman 2000
MedTagger[a]	A machine learning–based name entity detection system utilizing existing terminologies	Mayo Clinic (Liu)	Torii et al. 2011
MetaMap[a]	An expert-based system for mapping text to the UMLS	NLM (Aronson)	Aronson and Lang 2010
SecTag[a]	A system to tag clinical note section headers	Vanderbilt (Denny)	Denny et al. 2009 Denny et al. 2008

Note: Systems are listed alphabetically.

[a] Publicly available systems.

Medical Applications of NLP

NLP for Surveillance: Surveillance is a fundamental and important task in health care, especially surveillance of adverse events (AEs) based on the clinical texts. Hripcsak et al. (2003) developed a framework to discover AEs from clinical notes. They used MedLEE to parse the clinical narratives and generate a coded database, followed by query generation to detect and classify events.

NLP for Clinical Decision Support : An NLP system can transfer clinical texts to encoded information, which meets the needs of clinical decision support (CDS) (Demner-Fushman et al. 2009). For example, NLP systems can help to find patients who match certain criteria based on the information extracted from clinical texts. Jain et al. (1996) used MedLEE to encode the information in chest radiograph and mammogram reports and identified patients at risk of having tuberculosis (TB). Fiszman et al. (2000) found that an NLP system for automatic detection of acute bacterial pneumonia from chest x-ray reports performed similarly to physicians and better than lay persons and keyword searching. Day et al. (2007) have developed a daily program using the MPLUS NLP system and decision support technologies to automatically identify trauma patients. Compared with results with clinicians' judgments, the system performed well, with sensitivity of 71% and specificity of 99%.

Model Interpretability using Explainable AI for NLP applications

https://www.mdpi.com/1099-4300/23/1/18

https://deepsense.ai/overview-of-explainable-ai-methods-in-nlp/

NLP in Healthcare - One Page Summary with Mnemonics

🏥 NLP Tasks in Medicine

Natural Language Processing applies computational linguistics to healthcare texts, processing clinical notes, research papers, and patient records.

Mnemonic: “MEDICAL NLP” - Medicine Engages Data Intelligently Creating Automated Language Natural Learning Processing

🔧 Low-Level NLP Components

Mnemonic: “TSPSD” - Tokenization, Sentence boundary, POS tagging, Shallow parsing, Deep parsing

1. Tokenization - Breaking text into meaningful units (words, symbols)
   - Challenge: Medical terms like “Adams-Stokes” vs “Adams Stokes”
2. Sentence Boundary Detection (SBD) - Finding sentence start/end
   - Challenge: Medical abbreviations like “q.i.d.” or “p.r.n.”
3. Part-of-Speech (POS) Tagging - Identifying grammatical roles
   - Example: “The/DT cystic/JJ duct/NN was/VBD triply/RB clipped/VBN”
4. Shallow Parsing (Chunking) - Identifying phrase groups
   - Output: [NP The doctor] [VP treated] [NP the patient]
5. Deep Parsing - Complete syntactic structure showing relationships
   - Shows: doctor → treated → patient (with full dependency tree)

🎯 High-Level NLP Components

Mnemonic: “NERWS-RIE” - Negation Entity Recognition Word Sense, Relation Information Extraction

1. Negation Detection - Identifying absent conditions
   - Example: “Patient denies fever” → Fever: Negated
2. Named Entity Recognition (NER) - Classifying medical entities
   - Categories: Drugs, diseases, genes, symptoms
3. Relation Extraction - Finding entity relationships
   - Example: Drug-drug interactions, disease-symptom associations
4. Word Sense Disambiguation - Resolving ambiguous terms
   - Challenge: Medical abbreviations with multiple meanings
5. Semantic Role Labeling (SRL) - Identifying “who did what to whom”
   - Questions: Who, when, what, where, why
6. Information Extraction (IE) - Converting unstructured text to structured data

🤖 NLP Methods

Mnemonic: “SMHN” - SVM MaxEnt HMM N-gram

1. Support Vector Machine (SVM) - High accuracy classification
   - Uses: Smoking status, disease classification
2. Maximum Entropy (MaxEnt) - Probabilistic classification with least bias
   - Principle: “Don’t assume anything extra beyond training data”
3. Hidden Markov Models (HMM) - Probability-based state transitions
   - Example: Speech recognition with hidden vocal states
4. N-Gram Models - Statistical language modeling using consecutive words
   - Markov Assumption: Current word depends on previous n-1 words

📚 Clinical NLP Resources & Tools

Mnemonic: “UMLS-MSG” - UMLS MetaMap SPECIALIST GENIA

1. UMLS (Unified Medical Language System)
- Components: Metathesaurus, Semantic Network (135 types), SPECIALIST Lexicon

2. MetaMap - Maps biomedical text to UMLS concepts
3. SPECIALIST Tools - NLP utilities (LVG, Norm, WordInd, dTagger)
4. GENIA Corpus - Annotated biomedical abstracts for training

🏥 Medical Applications

Mnemonic: “SCD” - Surveillance Clinical Decision

1. Surveillance - Monitoring adverse events from clinical texts
2. Clinical Decision Support - Automated patient identification and risk assessment
3. Documentation - Converting clinical narratives to structured data

🔍 Key Challenges & Solutions

- Domain Specificity: Medical sublanguage requires specialized training
- Abbreviations: Extensive use of acronyms and abbreviations
- Negation Scope: Complex negation patterns in clinical text
- Explainable AI: Model interpretability for clinical decision-making

Memory Palace Technique: Imagine walking through a hospital where each room represents a component - Token-ization Room, POS-tagging Lab, NER-gency Department, and SVM-surgery Suite!