Module 3:

- **Data Science and Natural Language Processing (NLP) for Clinical Text**

Natural Language Processing

- Analyze free text to extract “information”
- Key challenges:
  - Ambiguity: *heart, ברק*
  - Variability: *diabetes, dm, diab.*
- Applications:
  - Search
  - Text Mining: information extraction, relations
  - Summarization

NLP for Medical Domain

Opportunity

- Availability of online textual documents
  - EHR: mostly textual (release notes)
  - Scientific literature (PubMed)

Challenge

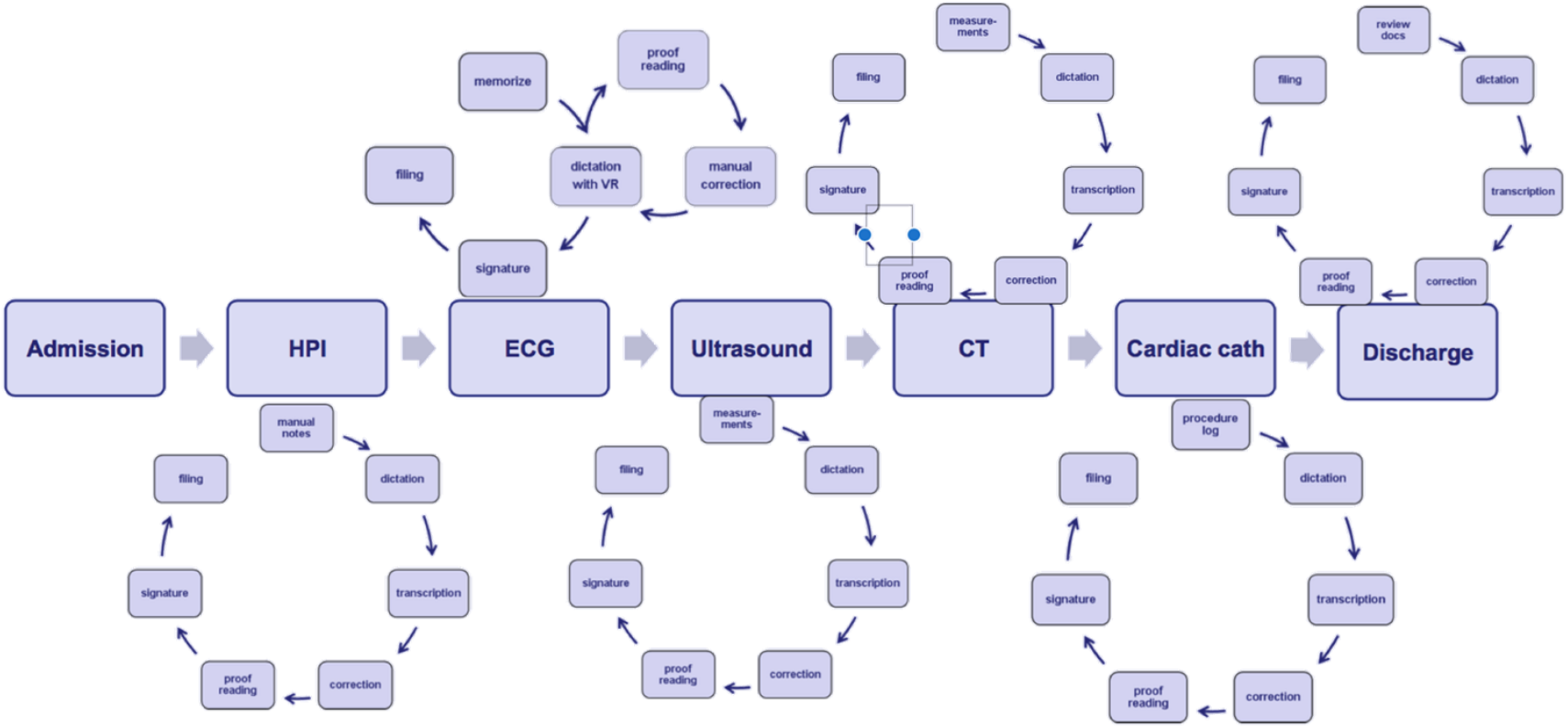
- Methods developed on “regular language” fail on “medical language”

EHR Data / Clinical Notes

Clinical observations captured in free-text notes in EHR

Image: https://www.emr-ehrs.com/gastrointestinal-electronic-health-records.php

Document Generation



Scientific Publications & Clinical Guidelines

Research findings are mainly disseminated via scientific publications (> 27 mio. indexed on PubMed) and synthesized in clinical practice guidelines to enable "evidence-based medicine"

NCBI Resources How To

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed

Advanced

Format: Abstract

Send to

Lancet Oncol. 2012 Mar;13(3):239-46. doi: 10.1016/S1470-2045(11)70393-X. Epub 2012 Jan 26.

Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial.

Rosell R[1], Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, Palmero R, Garcia-Gomez R, Pallares C, Sanchez JM, Porta R, Cobo M, Garrido P, Longo F, Moran T, Insa A, De Marinis F, Corre R, Bover I, Illiano A, Dansin E, de Castro J, Milella M, Reguart N, Altavilla G, Jimenez U, Provencio M, Moreno MA, Terrasa J, Muñoz-Langa J, Valdivia J, Isla D, Domine M, Molinier O, Mazieres J, Baize N, Garcia-Campelo R, Robinet G, Rodriguez-Abreu D, Lopez-Vivanco G, Gebbia V, Ferrera-Delgado L, Bombaron P, Bernabe R, Bearz A, Artal A, Cortesi E, Rolfo C, Sanchez-Ronco M, Drozdowskyj A, Queralt C, de Aguirre I, Ramirez JL, Sanchez JJ, Molina MA, Taron M, Paz-Ares L; Spanish Lung Cancer Group in collaboration with Groupe Français de Pneumo-Cancérologie and Associazione Italiana Oncologia Toracica.

Author information

1 Catalan Institute of Oncology, Badalona, Spain. rrosell@iconcologia.net

Abstract

BACKGROUND: Erlotinib has been shown to improve progression-free survival compared with chemotherapy when given as first-line treatment for Asian patients with non-small-cell lung cancer (NSCLC) with activating EGFR mutations. We aimed to assess the safety and efficacy of erlotinib compared with standard chemotherapy for first-line treatment of European patients with advanced EGFR-mutation positive NSCLC.

METHODS: We undertook the open-label, randomised phase 3 EURTAC trial at 42 hospitals in France, Italy, and Spain. Eligible participants were adults (> 18 years) with NSCLC and EGFR mutations (exon 19 deletion or L858R mutation in exon 21) with no history of chemotherapy for metastatic disease (neoadjuvant or adjuvant chemotherapy ending ≥ 6 months before study entry was allowed). We randomly allocated participants (1:1) according to a computer-generated allocation schedule to receive oral erlotinib 150 mg per day or 3 week cycles of standard intravenous chemotherapy of cisplatin 75 mg/m(2) on day 1 plus docetaxel (75 mg/m(2) on day 1) or gemcitabine (1250 mg/m(2) on days 1 and 8). Carboplatin (AUC 6 with docetaxel 75 mg/m(2) or AUC 5 with gemcitabine

Clinical trial report on PubMed

EGFR tyrosine kinase inhibitors (TKIs) are effective as first line treatment of advanced NSCLC in patients with sensitising *EGFR* mutations. The optimum treatment is orally delivered single agent therapy. TKIs significantly increased progression-free survival (PFS) (HR 0.45, 95% CI 0.36 to 0.58, P<0.0001) over SACT.[230] In a European trial, the median PFS was 9.4 months in the erlotinib (TKI) group and 5.2 months in the doublet SACT group, (HR 0.42, 95% CI 0.27 to 0.64), p<0.0001.[231] | 1+

Randomised evidence does not support the use of SACT in combination with a TKI in any patient group.[231,232] | 1++

A First line single agent tyrosine kinase inhibitors should be offered to patients with advanced NSCLC who have a sensitising *EGFR* mutation. Adding combination systemic anticancer therapy to a TKI confers no benefit and should not be used.

A Patients who have advanced disease, are performance status 0-1, have predominantly non-squamous NSCLC and are *EGFR* mutation negative should be offered combination systemic anticancer therapy with cisplatin and pemetrexed.

A All other patients with NSCLC should be offered combination systemic anticancer therapy with cisplatin/carboplatin and a third generation agent (docetaxel, gemcitabine, paclitaxel or vinorelbine).

A Platinum doublet systemic anticancer therapy should be given in four cycles; it is not recommended that treatment extends beyond six cycles.

SIGN Guideline: Management of lung cancer (2014)

Processing Clinical Notes

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed. Since then, self-monitoring of blood glucose (SMBG) showed blood glucose levels of 250-270 mg/dL. She was referred to an endocrinologist for further evaluation.

On examination, she was normotensive and not acutely ill. Her body mass index (BMI) was 18.7 kg/m2 following a recent 10 lb weight loss. Her thyroid was symmetrically enlarged and ankle reflexes absent. Her blood glucose was 272 mg/dL, and her hemoglobin A1c (HbA1c) was 10.3%. A lipid profile showed a total cholesterol of 261 mg/dL, triglyceride level of 321 mg/dL, HDL level of 48 mg/dL, and an LDL of 150 mg/dL. Thyroid function was normal. Urinanalysis showed trace ketones.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years, and limited her alcohol intake to 1 drink daily. Her mother's brother was diabetic.

Clinical Element Model

Disorder CEM
text: diabetes mellitus
code: 73211009
subject: patient
relative temporal context: 3 months ago

Medication CEM
text: Glyburide
code: 315989
subject: patient
frequency: once daily
negation indicator: not negated
strength: 2.5 mg

Tobacco Use CEM
text: smoking
code: 365981007
subject: patient
relative temporal context: 25 years
negation indicator: not negated

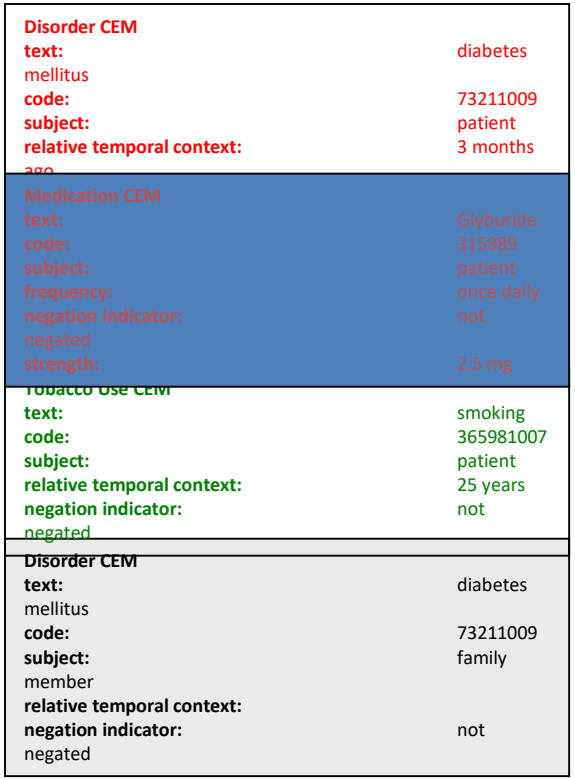
Disorder CEM
text: diabetes mellitus
code: 73211009
subject: family member
relative temporal context:
negation indicator: not negated

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years

Her mother's brother was diabetic.

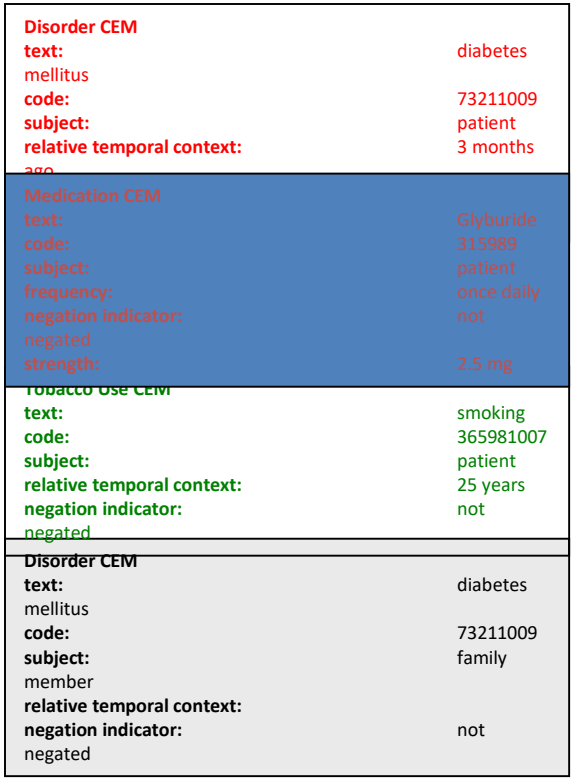
Comparative Effectiveness



Compare the effectiveness of different treatment strategies (e.g., modifying target levels for glucose, lipid, or blood pressure) in reducing cardiovascular complications in newly diagnosed adolescents and adults with type 2 diabetes.

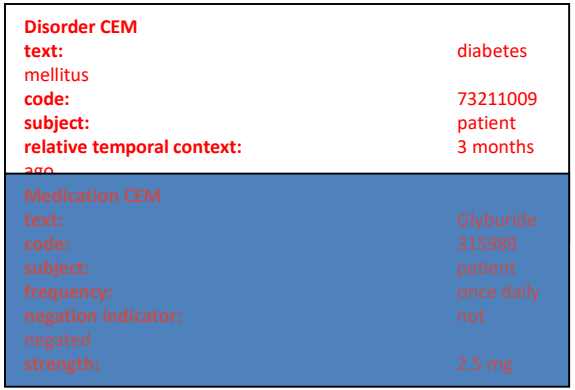
Compare the effectiveness of traditional behavioral interventions versus economic incentives in motivating behavior changes (e.g., weight loss, smoking cessation, avoiding alcohol and substance abuse) in children and adults.

Meaningful Use



- Maintain problem list
- Maintain active med list
- Record smoking status
- Provide clinical summaries for each office visit
- Generate patient lists for specific conditions
- Submit syndromic surveillance data

Clinical Practice



- Provide problem list and meds from the visit

Natural Language Processing

Some basic terms:

- Syntax: the allowable structures in the language: sentences, phrases, affixes (-ing, -ed, -ment, etc.).
- Semantics: the meaning(s) of texts in the language.
- Part-of-Speech (POS): the category of a word (noun, verb, preposition etc.).
- Bag-of-words (BoW): a featurization that uses a vector of word counts (or binary) ignoring order.
- N-gram: for a fixed, small N (2-5 is common), an n-gram is a consecutive sequence of words in a text.

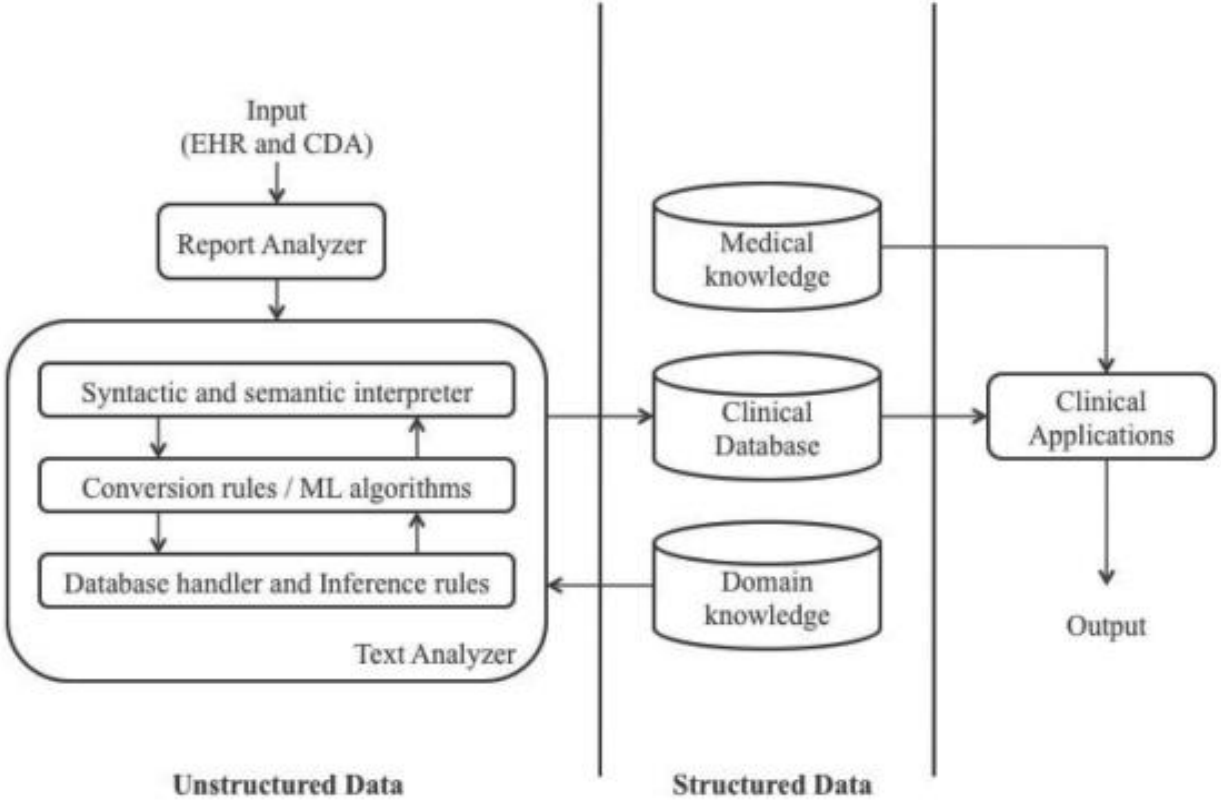


FIGURE 7.1: General workflow of a NLP system.

Core NLP Components

Morphological Analysis
Lexical Analysis
Syntactic Analysis
Semantic Analysis
Data Encoding

Input Text → Pt took aspirin 325 mg for knee pain

Tokenization →

Pt	*took*	aspirin	325	mg	for	knee	pain

Morphological analysis →

Pt	*take*	aspirin	325	mg	for	knee	pain

Lexical analysis, Syntactic analysis and Semantic analysis →

```
(ROOT
  (S
    (NP (NNP Pt))
    (VP (VBD took)
      (NP
        (NP
          (QP (CD aspirin) (CD 325))
          (NN mg))
        (PP (IN for)
          (NP (NN knee) (NN pain)))))))
```

Mining Information from Clinical Text

A. Information Extraction

1. **Preprocessing**
   - **a. Spell Checking**
   - **b. Word Sense Disambiguation**
   - **c. POS Tagging**
   - **d. Shallow and Deep Parsing**

2. **Context- Based Extraction**
   - **a. Concept Extraction**
   - **b. Association Extraction**
   - **c. Coreference Resolution**
   - **d. Negation**
   - **e. Temporality Analysis**

3. **Extracting Codes**

B.Current Methodologies

a. Rule-Based Approaches

b. Pattern-Based Algorithms

c. Machine Learning Algorithms

C. Clinical Text Corpora and Evaluation Metrics

D. Informatics for Integrating Biology and the Bedside

Mining Information from Clinical Text

- Clinical text mining is an **interdisciplinary area of research** requiring knowledge and skills in computer science, engineering, computational linguistics, and health science.

- It is a subfield of **biomedical NLP to determine classes of information** found in clinical text that are useful for basic biological scientists and clinicians for providing better health care.

- **Text mining and data mining techniques to uncover the information** on health, disease, and treatment response support the electronically stored details of patients' health records.

- A significant **chunk of information in EHR and CDA are text and extraction of such information by conventional data mining methods is not possible.**

- The **semi-structured and unstructured data** in the clinical text and even certain categories of test results such as **echocardiograms and radiology reports** can be mined for information by utilizing both data mining and text mining techniques.

Information Extraction

- Information extraction (IE) is a **specialized field of NLP for extracting predefined types of information** from the natural text.
- It is defined as the **process of discovering and extracting knowledge from the unstructured text.**
- IE differs from **information retrieval (IR) that is meant to be for identifying and retrieving relevant documents.**
- In general, **IR returns documents** and **IE returns information or facts**.
- A typical IE system for the clinical domain is a combination of components such as:
- **tokenizer, sentence boundary detector, POS tagger, morphological analyzer, shallow parser, deep parser (optional), gazetteer, named entity recognizer, discourse module, template extractor, and template combiner.**

- A **careful modeling of relevant attributes with templates** is required for the performance of high level components such as discourse module, template extractor, and template combiner.

- The high level components always depend on the performance of **the low level modules such as POS tagger, named entity recognizer, etc.**

- IE for clinical domain is meant for **the extraction of information present in the clinical text**.

- The Linguistic String Project–Medical Language Processor (LSP–MLP), and Medical Language Extraction and Encoding system (MedLEE) are the commonly adopted systems to extract UMLS concepts from clinical text.
- The Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) , Special Purpose Radiology Understanding System (SPRUS), SymText (Symbolic Text Processor), and SPECIALIST language-processing system are the major systems developed by few dedicated research groups for maintaining the extracted information in the clinical domain.

- Other important systems widely used in the clinical domain are MetaMap, IndexFinder, and KnowledgeMap.
- Among all, MetaMap is found to be useful with patients' EHR for automatically providing relevant health information.
- MetaMap and its Java version MMTx (MetaMap Transfer) were developed by the US NLM to index text or to map concepts in the analyzed text with UMLS concepts.
- Furthermore, NLP systems such as Lexical Tools and MetaMap [6] use UMLS with many other applications.

Table shows the major clinical NLP systems available along with their purpose and contents

TABLE 7.1: Major Clinical NLP Systems

Clinical NLP system	Purpose
LSP-MLP	NLP system for extraction and summarization of signs/symptoms and drug information, and identification of possible medication and side effects
MedLEE	A semantically driven system used for (1) extracting information from clinical narrative reports, (2) participating in an automated decision-support system, and (3) allowing NLP queries
cTAKES	Mayo clinical Text Analysis and Knowledge Extraction System
SPRUS	A semantically driven IE system
SymText	NLP system with syntactic and probabilistic semantic analysis driven by Bayesian Networks
SPECIALIST	A part of UMLS project with SPECIALIST lexicon, semantic network, and UMLS Metathesaurus
IndexFinder	A method for extracting key concepts from clinical text for indexing
KnowledgeMap	A full-featured content management system to enhance the delivery of medical education contents
Lexical Tools	A set of fundamental core NLP tools for retrieving inflectional variants, uninflectional forms, spelling variants, derivational variants, synonyms, fruitful variants, normalization, UTF-8 to ASCII conversion, and many more
MetaMap	A highly configurable program to map biomedical text to UMLS Metathesaurus concepts

Information Extraction

1. **Preprocessing**
   - **a. Spell Checking**
   - **b. Word Sense Disambiguation**
   - **c. POS Tagging**
   - **d. Shallow and Deep Parsing**

a. Spell Checking

- The **misspelling in clinical text is reported to be much higher than any other types of texts.**

- Traditional spell checker

- various research groups have come out with a **variety of methods for spell checking in the clinical domain:**

-**UMLS-based spell-checking error correction tool**

-morpho-syntactic disambiguation tools

1. Preprocessing

- The **primary source of information in the clinical domain is the clinical text** written in natural language.

- However, the **rich contents of the clinical text are not immediately accessible** by the clinical application systems that require input in a more structured form.

- An initial module adopted by various clinical NLP systems to **extract information is the preliminary preprocessing of the unstructured text** to make it available for further processing.

- The **most commonly used preprocessing techniques in clinical NLP** are

- spell checking
- word sense disambiguation
- POS tagging
- shallow and deep parsing.

Spell-checking literal text fields

Spelling errors are common in text describing health conditions, medical jargon, and descriptions of deaths.

Without handling errors in some way, a model will treat different spellings of a word as entirely unrelated.

Example:

Does "*gestation iabetes and placental abrupton*"

equal

"*gestational diabetes and placental abruption*"?

Spell-checking literal text fields

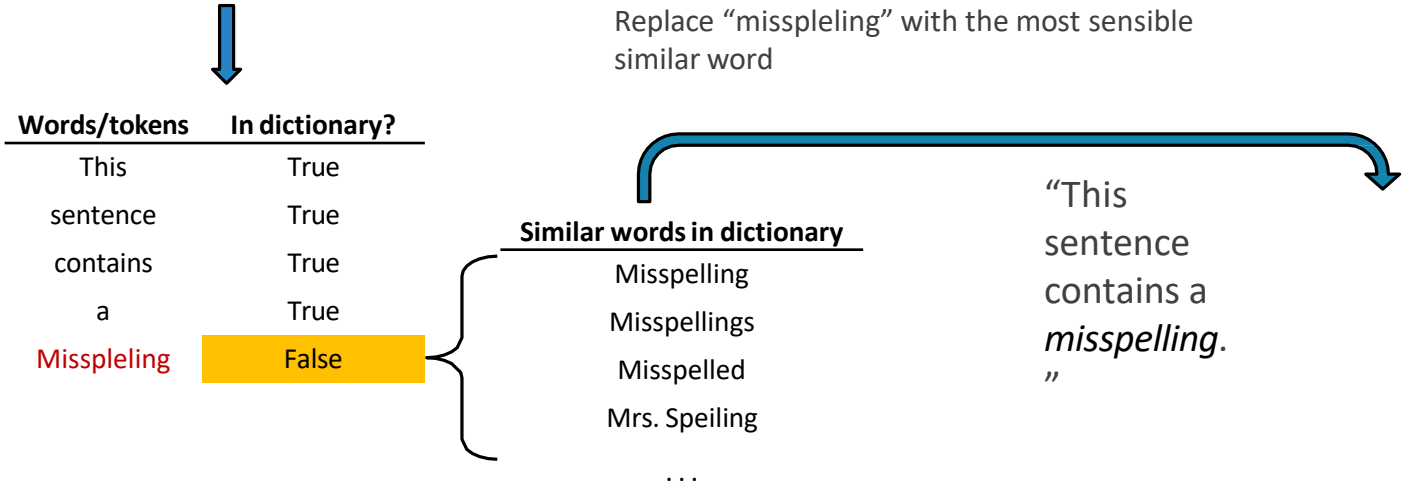
A good spell checker has three main components:

1. Dictionary
2. A method of measuring the “distance” between two strings
3. Language model or decision rules about which word from the dictionary was misspelled in the text

NOTE: The quality of all three parts corresponds to the overall quality of the spell checker. A bad dictionary, poor choice of distance metric, or an improper language model will cause poor results even if the other elements are well implemented.

Spell-checking literal text fields

"This sentence contains a *misspleling*."



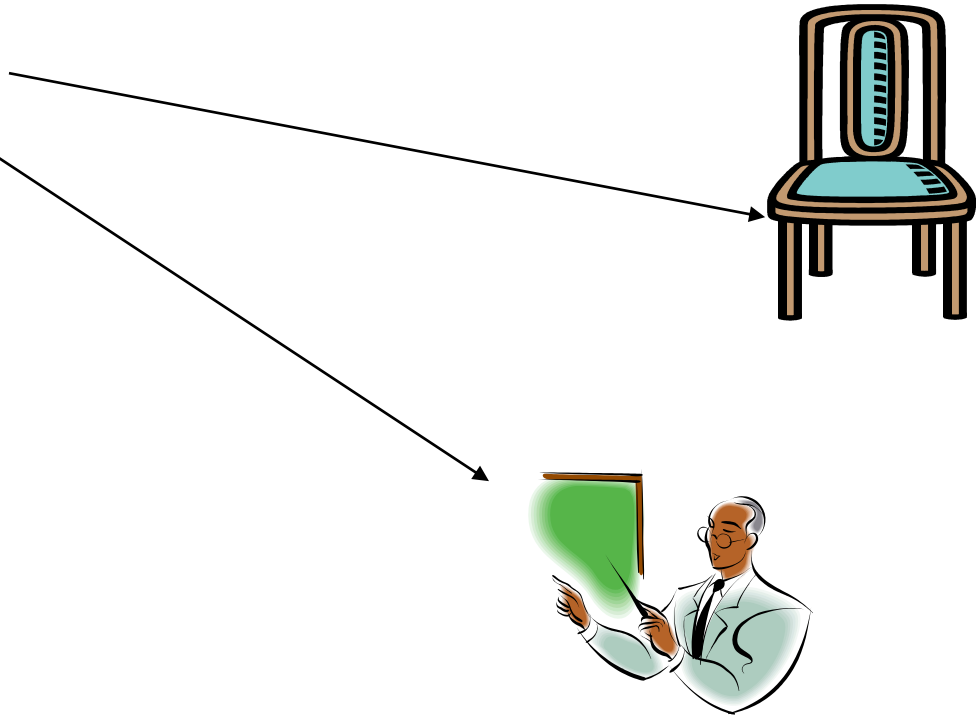
Word Sense Disambiguation

- The **process of understanding the sense of the word in a specific context** is termed as word sense disambiguation.

- The supervised ML classifiers and the unsupervised approaches automatically perform the word sense disambiguation for biomedical terms

Ambiguity

Chair



MOTIVATION

- One of the central challenges in NLP.
- Ubiquitous across all languages.
- Needed in:
  - **Machine Translation**: For correct lexical choice.
  - **Information Retrieval**: Resolving ambiguity in queries.
  - **Information Extraction**: For accurate analysis of text.
- Computationally determining which ***sense*** of a word is activated by its use in a particular ***context***.
  - E.g. I am going to withdraw money from the ***bank.***
- A classification problem:
  - Senses → Classes
  - Context → Evidence

Example to Illustrate:

- Consider the following synset:

  {heroin, diacetyl morphine, horse, junk,scag, smack}.

- It is annotated with the Medicine domain because heroin is a drug, and that is maybe best described as medical knowledge.

Example to Illustrate: Cont.

- On the other hand (on the text side), if we consider a news collection – Reuters corpus for example – the word heroin is likely to occur in the context of either:

✓ Crime news.

✓ Administrative news.

And without any strong relation with the medical field.

Why is domain information interesting?

- Due to its utility in many scenarios such as:

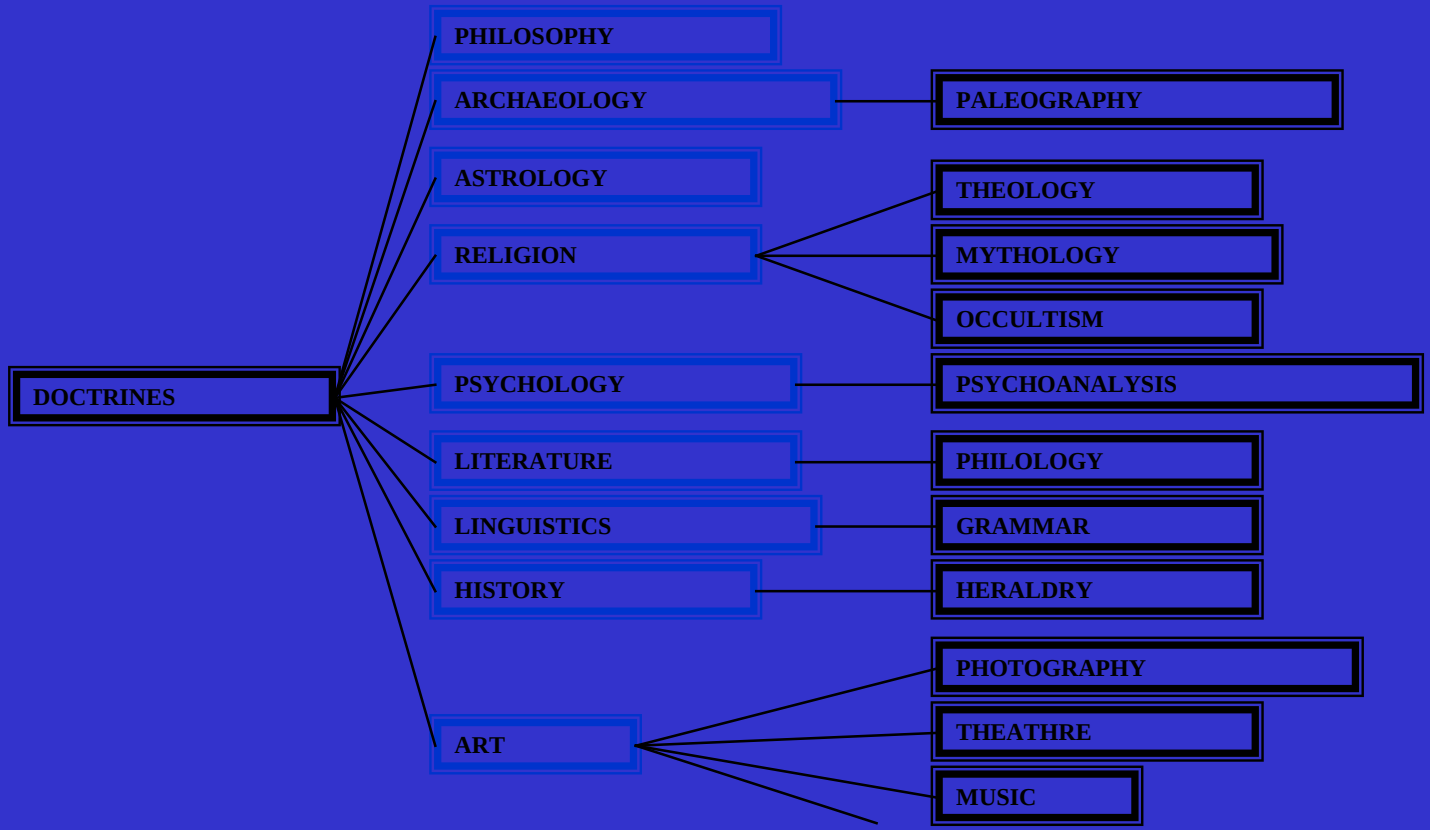
➢ Word Sense Disambiguation (WSD): where information from domain labels are used to establish semantic relations among word senses.

➢ Text Categorization (TC): Where categories are represented as symbolic labels.

WordNet Domains.

- Domains have been used to mark technical usages of words.
- In dictionaries, it is used only for a small portion of the lexicon. Therefore:
- WordNet Domains is an attempt to extend the coverage of domain labels with an already existing lexical database.
- WordNet (version 1.6) Synsets have been annotated with at least one domain label selected from a set of about 200 labels hierarchically organized.

WordNet Domains



WordNet Domains.

- The word “bank” has 10 different senses.
- Three of them (#1, #3, and #6) can be grouped under the Economy domain.
- While #2 and #7 both belong to the Geography and Geology domain.
- → Reduction of the polysemy from 10 to 7 senses.

Sense	Synset and Gloss	Domains
#1	Depository financial institution, bank, banking, banking company.	Economy
#2	bank (sloping land …)	Geography, Geology
#3	bank (a supply or stock held in a reserve)	Economy
#4	bank, bank building (a building …)	Architecture, Economy
#5	bank, (an arrangement of similar objects.	Factotum
#6	savings bank, coin bank, money box.	Economy
#7	bank, (a long ridge or pile…)	Geography, Geology
#8	Bank (the funds held by a gambling house …)	Economy, Play
#9	bank, cant camber (a slope in the the turn of a road …)	Architecture
#10	bank (a flight maneuver…)	Transport

Example.

- Consider the following synset:

  {beak, bill, neb, nib}

- It will be automatically marked with the code Zoology, starting from the synset {bird} and following “part_of” relation.

POS Tagging

- An important preprocessing step adapted by most of the NLP systems is POS tagging that reads the text and assigns the parts of speech tag to each word or token of the text.

- POS tagging is the annotation of words in the text to their appropriate POS tags by considering the related and adjacent words in a phrase, sentence, and paragraph.

- POS tagging is the first step in syntactic analysis and finds its application in IR, IE, word sense disambiguation, etc.

- POS tags are a set of word categories based on the role that words may play in the sentence in which they appear.

- The most common set contains seven different tags: Article, Noun, Verb, Adjective, Preposition, Number, and Proper Noun.

Parts of Speech

Thrax's original list (c. 100 B.C):

- Noun (boat, plane, Obama)
- Verb (goes, spun, hunted)
- Pronoun (She, Her)
- Preposition (in, on)
- Adverb (quietly, then)
- Conjunction (and, but)
- Participle (eaten, running)
- Article (the, a)

English Word Classes

- Parts-of-speech can be divided into two broad supercategories: closed class types and open class types.
- Closed classes are those that have relatively fixed membership.
- Example : Prepositions are a closed class because there is a fixed set of them in English.
- Nouns and verbs are open classes because new nouns and verbs are borrowed from other languages.

Open Classes

- There are four major open classes that occur in the languages of the world:
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs.
- It turns out that English has all four of these.

Closed Classes

- Here’s a quick overview of some of the more important closed classes in English, with a few examples of each:
  - Prepositions: on, under, over, near, by, at, from, to, with
  - Determiners: a, an, the
  - Pronouns: she, who, I, others
  - Conjunctions: and, but, or, as, if, when
  - Auxiliary verbs: can, may, should, are
  - Particles: up, down, on, off, in, out, at, by,
  - Numerals: one, two, three, first, second, third

Other Classes

- English also has many words of more or less unique function, including :
  - interjections (oh, ah, hey, man, alas, uh, um)
  - negatives (no, not)
  - politeness markers (please, thank you)
  - greetings (hello, goodbye), and
  - the existential there (there are two on the table)
- among others.

What are tagsets in NLP?

- The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging, POS-tagging, or simply tagging.

- Parts of speech are also known as word classes or lexical categories.

- The collection of tags used for a particular task is known as a tagset.

Tagsets for English

- The popular tagsets for English are evolved from the 87-tag tagset used for the Brown corpus.
- The Brown corpus is a 1 million word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.)
- This corpus was tagged with parts-of-speech by first applying the TAGGIT program and then hand-correcting the tags.

Other commonly used tagset

- The small 45-tag Penn Treebank tagset
- The medium-sized 61 tag C5 tagset used by the Lancaster UCREL project's CLAWS (the Constituent Likelihood AutomaticWord-tagging System) tagger to tag the British National Corpus (BNC)

Parts of Speech (Penn Treebank 2014)

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
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Example Tagset

- A much more elaborated set of tags is provided by complete **Brown Corpus tag-set** (www.hit.uib.no/icame/brown/bcm.html) with **eighty seven basic tags** and

- **Penn Treebank tag-set** (www.cis.upenn.edu/treebank) with **forty five tags**.

Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	' or "
POS	Possessive ending	*'s*	"	Right quote	' or "
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	[, (, {, <
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	],), }, >
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	. ! ?
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	: ; ... – -
RP	Particle	*up, off*			

Example

- The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- There/EX are/VBP 70/CD children/NNS there/RB

Grammars

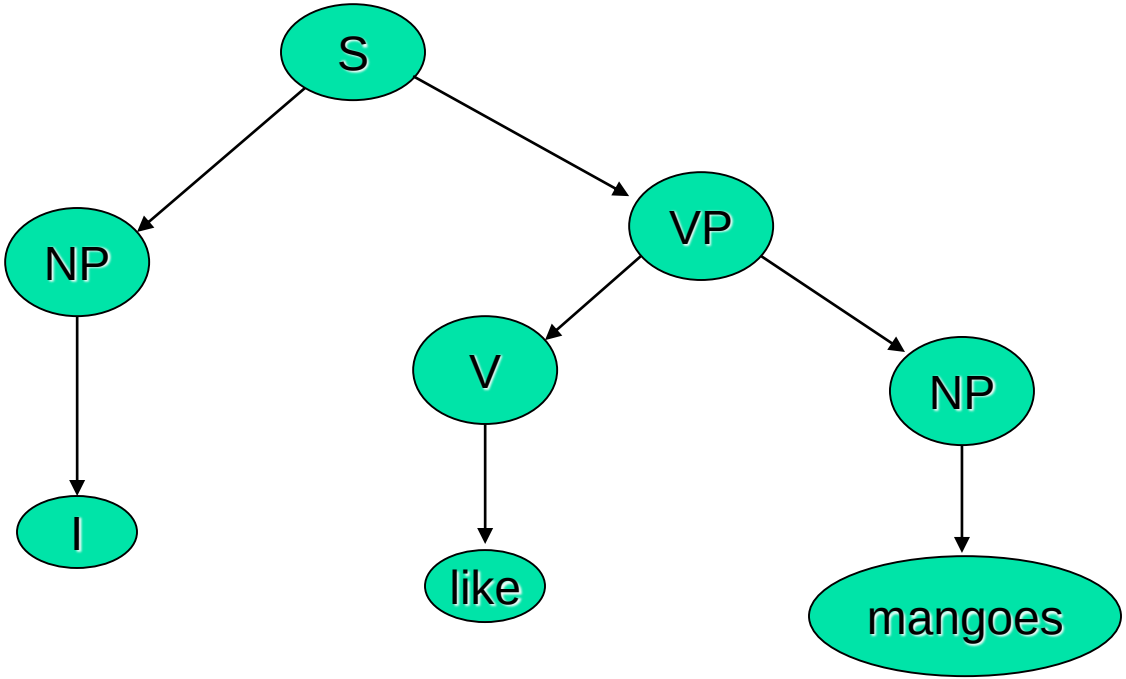
- Grammars comprise rules that specify acceptable sentences in the language: (S is the sentence or root node)
- S → NP VP
- S → NP VP PP
- NP → DT NN
- VP → VB NP
- VP → VBD
- PP → IN NP
- DT → “the”
- NN → “mat”, “cat”
- VBD → “sat”
- IN → “on”

Grammars

- Grammars comprise rules that specify acceptable sentences in the language: (S is the sentence or root node) “the cat sat on the mat”
- S → NP VP
- S → NP VP PP (the cat) (sat) (on the mat)
- NP → DT NN (the cat), (the mat)
- VP → VB NP
- VP → VBD
- PP → IN NP
- DT → “the”
- NN → “mat”, “cat”
- VBD → “sat”
- IN → “on”

Syntax Processing Stage

Structure Detection



Parsing Strategy

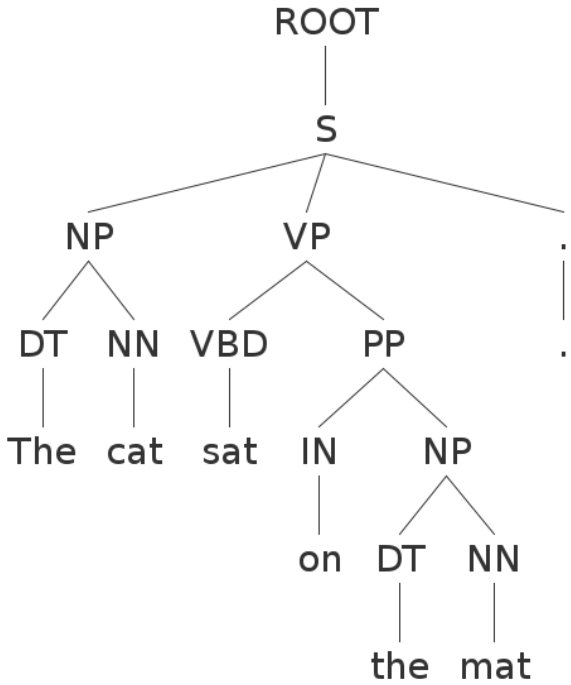
- Driven by grammar
  - S-> NP VP
  - NP-> N | PRON
  - VP-> V NP | V PP
  - N-> Mangoes
  - PRON-> I
  - V-> like

Grammars

- English Grammars are context-free: the productions do not depend on any words before or after the production.
- The reconstruction of a sequence of grammar productions from a sentence is called “parsing” the sentence.
- It is most conveniently represented as a tree:

Parse Trees

- “The cat sat on the mat”



Parse Trees

- In bracket notation:

- (ROOT
- (S
- (NP (DT the) (NN cat))
- (VP (VBD sat)
- (PP (IN on)
- (NP (DT the) (NN mat))))))

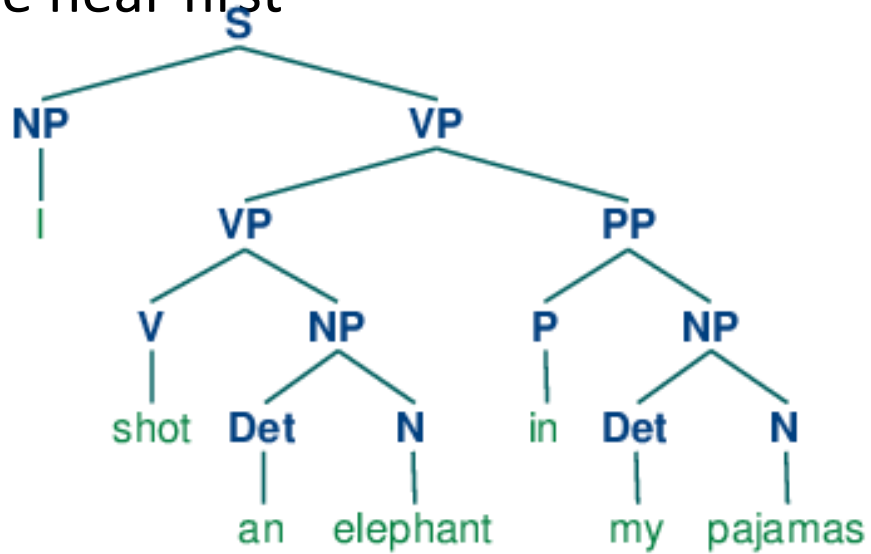
Grammars

- There are typically multiple ways to produce the same sentence. Consider the statement by Groucho Marx:

  - “While I was in Africa, I shot an elephant in my pajamas”
  - “How he got into my pajamas, I don’t know”

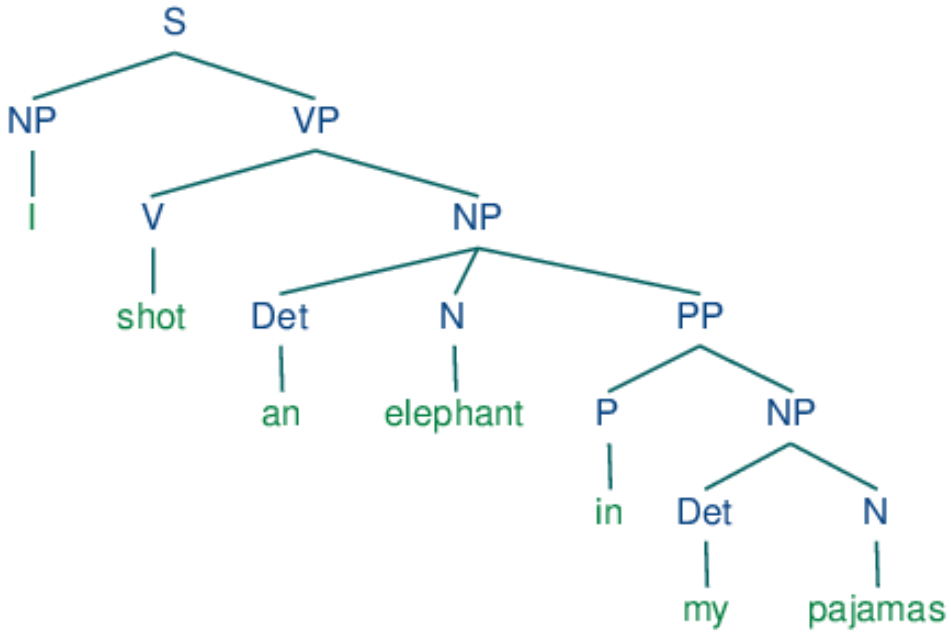
Parse Trees

- "…,I shot an elephant in my pajamas" -what people hear first



Parse Trees

- Groucho's version

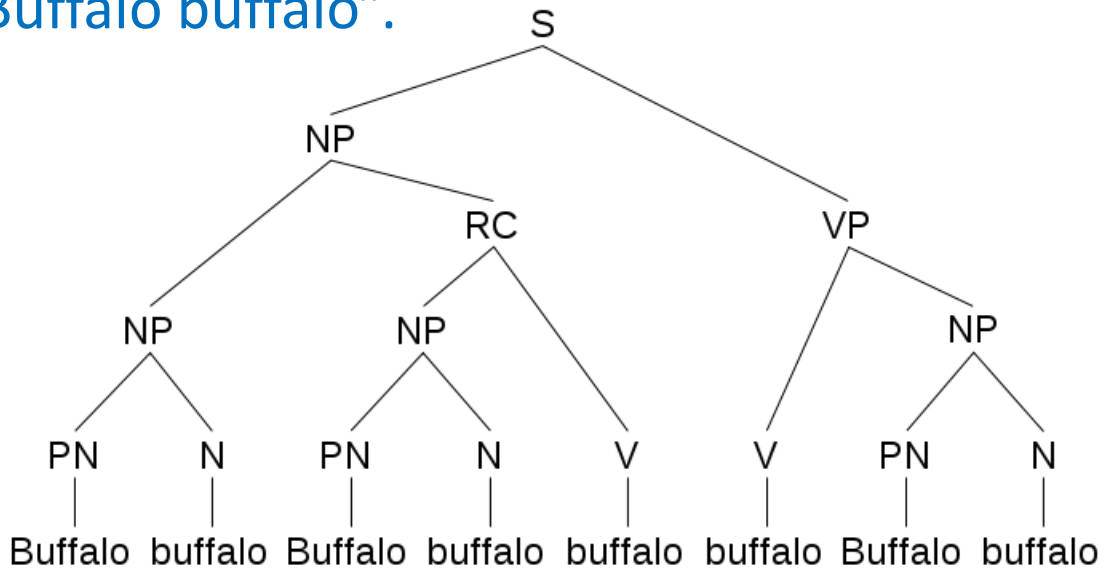


Grammars

- Recursion is common in grammar rules, e.g.
- NP → NP RC
- Because of this, sentences of arbitrary length are possible.

Recursion in Grammars

- “Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo”.



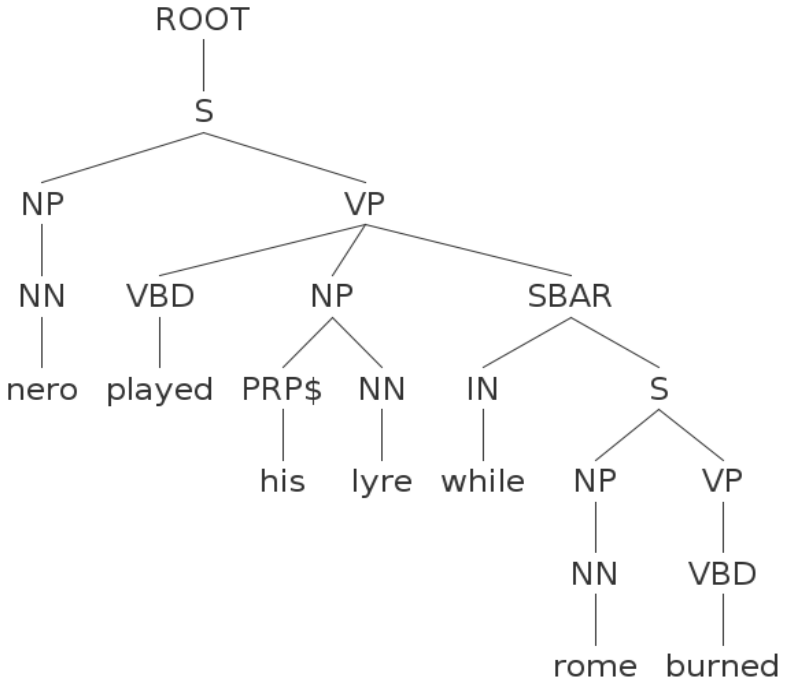
Grammars

- Its also possible to have “sentences” inside other sentences…

- S → NP VP
- VP → VB NP SBAR
- SBAR → IN S

Recursion in Grammars

- “Nero played his lyre while Rome burned”.



PCFGs

- Complex sentences can be parsed in many ways, most of which make no sense or are extremely improbable (like Groucho's example).

- Probabilistic Context-Free Grammars (PCFGs) associate and learn probabilities for each rule:
- S → NP VP 0.3
- S → NP VP PP 0.7

- The parser then tries to find the most likely sequence of productions that generate the given sentence. This adds more realistic "world knowledge" and generally gives much better results.
- Most state-of-the-art parsers these days use PCFGs.

Systems

- **NLTK:** Python-based NLP system. Many modules, good visualization tools, but not quite state-of-the-art performance.
- **Stanford Parser:** Another comprehensive suite of tools (also POS tagger), and state-of-the-art accuracy. Has the definitive dependency module.
- **Berkeley Parser:** Slightly higher parsing accuracy (than Stanford) but not as many modules.
- Note: high-quality parsing is usually very slow, but see: https://github.com/dlwh/puck

Mining Information from Clinical Text

A. Information Extraction

1. **Preprocessing**
   - **a. Spell Checking**
   - **b. Word Sense Disambiguation**
   - **c. POS Tagging**
   - **d. Shallow and Deep Parsing**

2. **Context- Based Extraction**
   - **a. Concept Extraction**
   - **b. Association Extraction**
   - **c. Coreference Resolution**
   - **d. Negation**
   - **e. Temporality Analysis**

3. **Extracting Codes**

B.Current Methodologies

a. Rule-Based Approaches

b. Pattern-Based Algorithms

c. Machine Learning Algorithms

C. Clinical Text Corpora and Evaluation Metrics

D. Informatics for Integrating Biology and the Bedside

Shallow and Deep Parsing

- Parsing is the **process of determining the complete syntactic structure of a sentence or a string of symbols in a language**.

- Parser is a tool **that converts an input sentence into an abstract syntax tree such as the constituent tree and dependency tree**,

- Dependency tree,whose leafs correspond to the words of the given sentence and the internal nodes represent the grammatical tags such as noun, verb, noun phrase, verb phrase, etc.

- Most of the parsers apply ML approaches such as PCFGs (probabilistic context-free grammars) as in the Stanford lexical parser and even maximum entropy and neural network.

- **Few parsers** even **use lexical statistics** by considering the words and their POS tags.

- Such taggers are well known for **overfitting problems that require additional smoothing.**

- An **alternative to the overfitting problem is to apply shallow parsing**, which splits the text into nonoverlapping word sequences or phrases, such that syntactically related words are grouped together.
- The word phrase represents the predefined grammatical tags such **as noun phrase(NP), verb phrase(VP), prepositional phrase(PP), adverb phrase(ADP), subordinated clause (SC), adjective phrase (AP), conjunction phrase(CP), and list marker(LM).**
- The benefits of **shallow parsing are the speed and robustness of processing.** Parsing is generally useful as a preprocessing step in extracting information from the natural text

Context-Based Extraction

- The fundamental step for a clinical NLP system is the recognition of medical words and phrases because these terms represent the concepts specific to the domain of study and make it possible to understand the relations between the identified concepts.

- Even highly sophisticated systems of clinical NLP include the initial processing of recognizing medical words and phrases prior to the extraction of information of interest.

- While IE from the medical and clinical text can be carried out in many ways, this section explains the five main modules of IE.

Concept Extraction

- Extracting concepts (such as drugs, symptoms, and diagnoses) from clinical narratives constitutes a basic enabling technology to unlock the knowledge within and support more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment.

- The first and foremost module in clinical NLP following the initial text preprocessing phase is the identification of the boundaries of the medical terms/phrases and understanding the meaning by mapping the identified term/phrase to a unique concept identifier in an appropriate ontology.

- The recognition of clinical entities can be achieved by a dictionary-based method using the UMLS Metathesaurus, rule-based approaches, statistical method, and hybrid approaches.

- The identification and extraction of entities present in the clinical text largely depends on the understanding of the context.

- For example, the recognition of diagnosis and treatment procedures in the clinical text requires the recognition and understanding of the clinical condition as well as the determination of its presence or absence.

- The contextual features related to clinical NLP are negation (absence of a clinical condition), historicity (the condition had occurred in the recent past and might occur in the future), and experiencer (the condition related to the patient).

- While many algorithms are available for context identification and extraction, it is recommended to detect the degree of certainty in the context.

- A baseline approach to concept extraction typically relies on a dictionary or lexicon of the concepts to be extracted, using string comparison to identify concepts of interest.
- Clinical narratives contain drug names, anatomical nomenclature, and other specialized names and phrases that are not standard in everyday English such as “benign positional vertigo,” “l shoulder inj,” “po pain medications,” “a c5-6 acdf,” “st changes,” “resp status,” and others.
- There is also a high incidence of abbreviation usage, and many of the abbreviations have a different meaning in other genres of English.
- Descriptive expressions (such as coil embolization of bleeding vessel, a large bloody bowel movement, a tagged RBC scan and R intracerebral hemorrhage drainage) are commonly used to refer to concepts rather than using canonical terms.
- The specialized knowledge requirement and the labor-intensive nature of the task make it difficult to create a lexicon that would include all such expressions, particularly given that their use is often non-standard and varies across institutions and medical specialties, or even from one department to another in the same hospital, rendering dictionary-based approaches less adaptable in this domain.
- An alternative to the dictionary based approach is the use of ML methods such as conditional random fields (CRF) and SVM that have achieved excellent performance in concept extraction. Torii et al. studied the portability of ML taggers for concept extraction using the 2010 i2b2/VA Challenge.
- Furthermore, the authors examined the performance of taggers with the increase in size of the dataset .
- While supervised ML approaches offer a promising alternative, a reliable system usually needs a large annotated corpus with as many relevant examples as possible.
- Therefore, clinical corpora generation is emerging as a specific branch of research in clinical NLP for the development of ML approaches

Association Extraction

- Clinical text is the rich source of information on patients' conditions and their treatments with additional information on potential medication allergies, side effects, and even adverse effects.

- Information contained in clinical records is of value for both clinical practice and research; however, text mining from clinical records, particularly from narrative-style fields (such as discharge summaries and progress reports), has proven to be an elusive target for clinical Natural Language Processing (clinical NLP), due in part to the lack of availability of annotated corpora specific to the task.

- Yet, the extraction of concepts (such as mentions of problems, treatments, and tests) and the association between them from clinical narratives constitutes the basic enabling technology that will unlock the knowledge contained in them and drive more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment.

- The clinical concepts appearing in the clinical text are related to one another in a number of ways.

- A better understanding of clinical text is possible through the identification and extraction of meaningful association or relationships between the concepts present in the text.

TABLE 7.3: Resources for Association Extraction

Resource	Purpose
UMLS Semantic Network	It defines the binary relations between the UMLS semantic types.
MedLEE	System to extract, structure, and encode clinical information in textual patient reports so that the data can be used by subsequent automated processes.
BioMedLEE	System for extracting phenotypic information underlying molecular mechanisms and their relationships.
SemRep	It maps syntactic elements (such as verbs) to predicates in the Semantic Network, such as TREATS and DIAGNOSIS.

- However, the clinical text is not always written in a way that encodes the nature of the semantic relations.
- The two concepts are not likely to occur together in the same sentence or even in the same section of the clinical text.
- In other words, the association between the concepts is generally annotated explicitly to match the clinical narratives appearing in the clinical text.
- One possible approach to annotate the concepts and their association is by using the clinical text with strongly related concepts.
- However, it may not be possible to determine the exact nature of the association from clinical text.
- In such cases, the biomedical literature provides a rich source of associated concepts that are confirmed through various research groups.
- Thus, the association between the concepts can be annotated by verifying with the association information available in the biomedical literature .

- When the explicitly stated associations are not available, the association between the concepts is identified through co-occurrence between the two concepts in the same clinical text as the twoconcepts are not likely to occur together in the same sentence or even section of the note.

- The association between any pair of UMLS concepts can be calculated with similarity measurements.

- Researchers used:
- the similarity measure to calculate the similarity between gene and disease.
- the word-level similarity measures offered by UMLS-similarity to provide contextual recommendations relevant to the health information conversation system

Coreference Resolution

- Coreferential expressions are common in clinical narratives and therefore understanding coreference relations plays a critical role in the discourse-level analysis of clinical documents, such as compiling a patient profile.
- Since the language and description style in clinical documents differ from common English, it is necessary to understand the characteristics of clinical text to properly perform coreference resolution.
- A comprehensive methodological review of coreference resolution developed for general English can be applied for coreference resolution in the clinical domain.

The existing methodologies for coreference resolution are:
1. Heuristics-based approaches based on linguistic theories and rules
2. Supervised machine learning approaches with binary classification of markable mention/ entity pairs or classification by ranking markables
3. Unsupervised machine learning approaches, such as nonparametric Bayesian models or expectation maximization clustering

- The heuristics-based approaches are the early attempts for the coreference resolution task to incorporate a knowledge source to prune unlikely antecedent candidates to get the best candidate by employing a multitude of features such syntactic, semantic and pragmatic constraints and preferences
- The supervised ML approaches replaced the interest of researchers to use complete heuristics-based systems.
- The binary classification, ranking, anaphoricity and specialized models are the major methods available for supervised ML approaches.
- On the other hand, the unsupervised approaches for coreference resolution adopt a fully generative, nonparametric Bayesian model based on hierarchical Dirichlet processes.
- A multi-pass system that applies tiers of resolution models is applied for coreference analysis.
- Researchers developed a system that applies tiers of resolution models one at a time.
- Each tier (sieve) consists of similar deterministic rules and builds on outputs of previously applied sieves.
- On the other hand, researchers employed a multi-pass sieve framework to exploit a heuristic-based approach along with a supervised ML method, specifically factorial hidden Markov models (FHMMs).
- They provide a review of the approaches in the general English and biomedical literature domains and discuss challenges in applying those techniques in the clinical narrative.
- Furthermore, the 2011 i2b2/VA/Cincinnati challenge focuses on coreferential relations between common, clinically relevant classes in medical text.
- These classes include problem, treatment, test, person, and pronoun. Coreferring mentions are to be paired together, and the pairs are to be linked to form a chain that represents the entity being referenced.
- The aim of the challenge is to produce coreferential chains of these mentions at document level (i.e., coreference relations are made across paragraphs or sections within the same document, but not across documents).

Negation

- "Negation" is an important context that plays a critical role in extracting information from the clinical text.

- Many NLP systems incorporate a separate module for negation analysis in text preprocessing.

- However, the importance of negation identification has gained much of its interest among the NLP research community in recent years.

- As a result, explicit negation detection systems such as NegExpander, Negfinder , and a specific system for extracting SNOMED-CT concepts as well as negation identification algorithms such as NegEx that uses regular expression for identifying negation and a hybrid approach based on regular expressions and grammatical parsing are developed by a few of the dedicated research community.

- While the NegExpander program identifies the negation terms and then expands to the related concepts, Negfinder is a more complex system that uses indexed concepts from UMLS and regular expressions along with a parser using LALR (look-ahead left-recursive) grammar to identify the negations

Temporality Analysis

- Temporal resolution for events and time expressions in clinical notes is crucial for an accurate summary of patient history, better medical treatment, and further clinical study.

- Discovery of a temporal relation starts with extracting medical events and time information and aims at building a temporal link (TLINK) between events or between events and time expressions.

- Clinical practice and research would benefit greatly from temporal expression and relation detection.

- Therefore, temporal Natural Language Processing and Data Mining for Clinical Text expression and relation discovery in clinical NLP are timely and inevitable to improve clinical text mining.

- A comprehensive temporal information discovery in clinical text requires medical event extraction, time information, and temporal relation identification.

- Temporal expression extraction is the first step to resolve temporal relations for any advanced natural language applications, such as text summarization, machine translation, and question answering.
- Several systems are available for extracting temporal expression.
- GUTime developed by Georgetown University is an extension of the TempEx tagger, which is a temporal tagger based on Perl regular expression.
- GUTime is now available as part of the TARSQI toolkit.
- HeidelTime is a rule-based system that is built in a UIMA framework and performed best for SemEval-2 (http://semeval2.fbk.eu/).
- SUTime is also a rule-based system using regular expression and is implemented as one of the annotators in the Stanford CoreNLP pipeline.

- **Extracting Codes**
- Extracting codes is a popular approach that uses NLP techniques to extract the codes mapped to controlled sources from clinical text.
- The most common codes dealing with diagnoses are the International Classification of Diseases (ICD) versions 9 and 10 codes.
- The ICD is designed to promote international comparability in the collection, processing, classification and presentation of mortality statistics.
- ICD-10 is the latest revised codes available with coding for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury (http://apps.who.int/classifications/icd10/browse/2010/en).
- Recently, a clinically modified form of ICD-10 version called ICD-10-CMis developed by the National Center for Health Statistics (NCHS).
- The entire draft of the tabular list of ICE-10-CMis available on the NCSH website for public comment.
- The specific clinical modifications incorporated in ICD-10-CM include many resources such as

(1) the additional information related to ambulatory and managed care encounters,
(2) expanded injury codes,
(3) combined diagnosis/symptom codes to reduce the number of codes needed to describe a disease condition,
(4) incorporation of common 4_{th} and 5_{th} digit sub-classifications, and
(5) laterality and greater specificity in code assignment.

- The medical NLP challenge in the year 2007 came out with a shared task exercise with a moderately large test/training corpus of radiology reports and their ICD-9-CMcodes.
- Most of the teams utilized multi-component coding systems for extracting codes from the text.
- One of the participated group utilized NLM's Medical Text Indexer, a SVM classifier and a k-NN classifier for extracting and arranging the codes in a stack-like architecture.
- Another team used ML, rule-based system and an automatic coding system based in human coding policies for extracting codes.

- A comprehensive clinical terminology similar to ICD is the Systematized Nomenclature of Medicine Clinical Terms (SNOMED–CT) that was originally created by the College of Amercian Pathologists (CAP) and distributed by the International Health Terminology Standards Development Organisation (IHTSDO) located in Denmark.

- It is one of the suites for use in US federal governmentsystems for the electronic exchange of clinical health information. SNOMED CT is a required standard in interoperability specifications of the US Healthcare Information Technology Standards Panel.

- The SNOMED CT is also implemented as a standard code by many clinical researchers in the area of NLP for clinical text (http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html).

- In addition to the standard codes such as ICD and SNOMED, a widely adapted clinical NLP system called MedLEE is used as a code extractor in many clinical contexts.

- Many NLP systems implement MedLEE for extracting the codes: an automated pneumonia severity score coding system, an NLP system for neuroradiology standard concept extraction, and an approach to code a standard for health and health-related states.

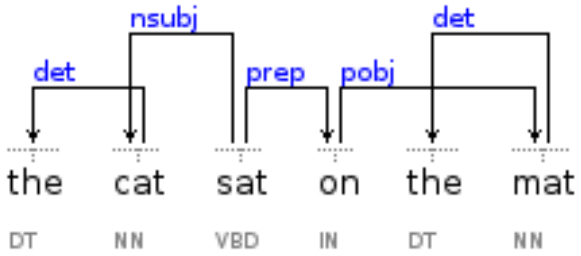
Dependencies

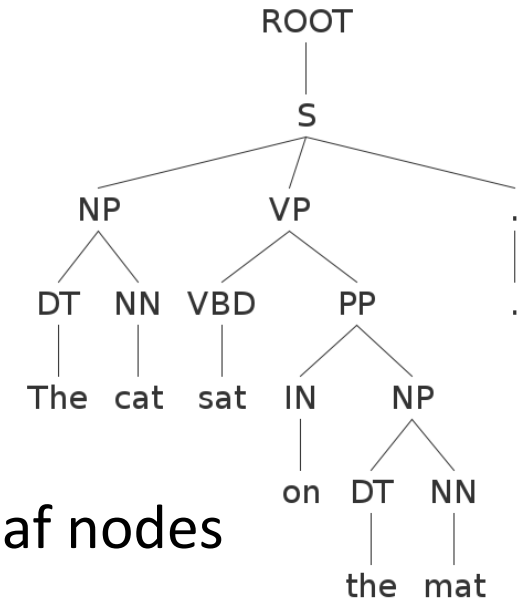
- In a constituency parse, there is no direct relation between the constituents and words from the sentence (except for leaf nodes which produce a single word).
- In dependency parsing, the idea is to decompose the sentence into relations directly between words.
- This is an older, and some argue more natural, decomposition of the sentence. It also often makes semantic interpretation (based on the meanings of the words) easier.
- Lets look at a simple example:

Dependencies

- “The cat sat on the mat”

- dependency tree





- constituency labels of leaf nodes

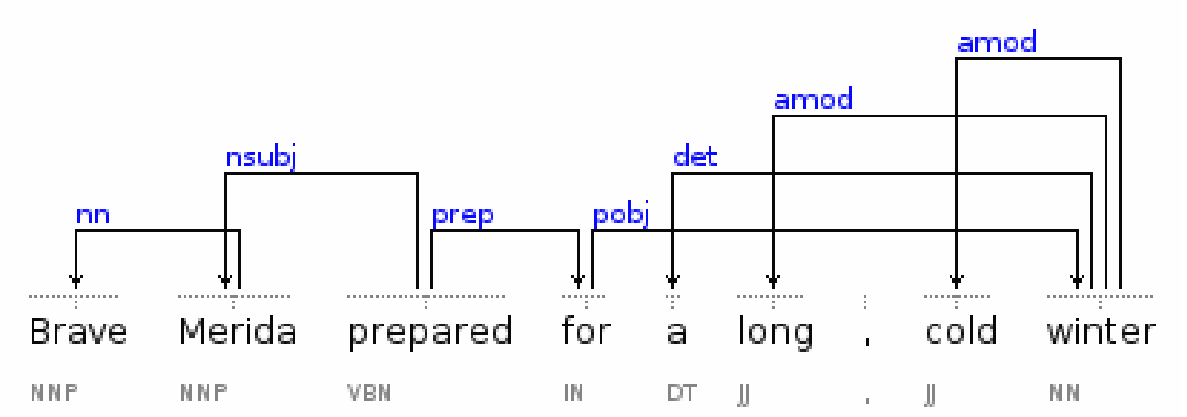
Dependencies

- From the dependency tree, we can obtain a “sketch” of the sentence. i.e. by starting at the root we can look down one level to get:
  - “cat sat on”
- And then by looking for the object of the prepositional child, we get:
  - “cat sat on mat”
- We can easily ignore determiners “a, the”.

- And importantly, adjectival and adverbial modifiers generally connect to their targets:

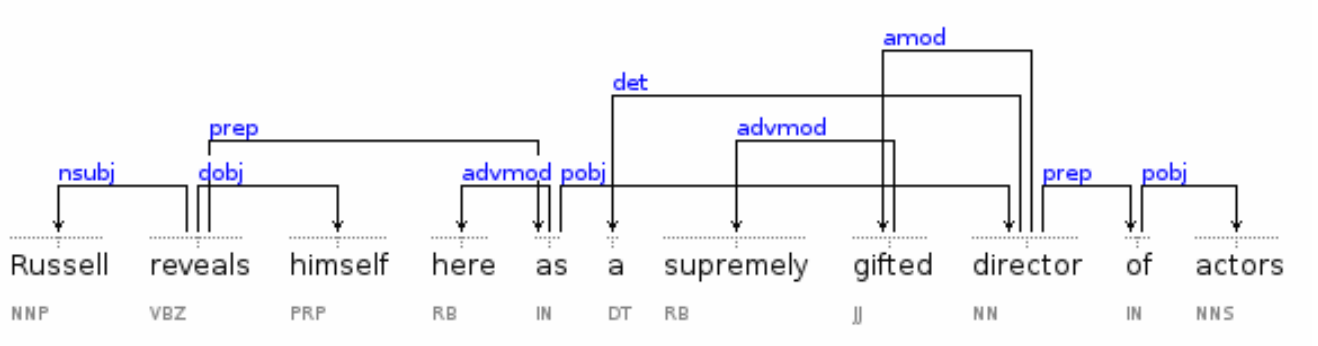
Dependencies

- “Brave Merida prepared for a long, cold winter”



Dependencies

- “Russell reveals himself here as a supremely gifted director of actors”



Dependencies

- Stanford dependencies are constructed from the output of a constituency parser (so you can in principle use other parsers).
- The mapping is based on hand-written regular expressions.
- Dependency grammars have been widely used for sentiment analysis and for semantic embedings of sentences.

Current Methodologies

- NLP as an intersection of artificial intelligence and linguistics was initially distinct from IR.
- However, NLP and IR have converged to a greater extent in recent years and applied together for indexing and searching large volumes of text.
- Currently, NLP adopts techniques and methodologies from several, very diverse fields to broaden its applications in various subtasks related to clinical text.
- The sub-tasks of NLP in two groups namely low-level and high-level tasks.
- The low-level tasks are related to sentence boundary detection, tokenization, part-of-speech tagging, morphological decomposition, shallow parsing, and problem specific segmentations.
- The high-level tasks are usually problem specific and use the low-level subtasks for building the models.
- Some of the high-level tasks with wide range of application in NLP are spelling/grammatical error identification and recovery, NER, word sense disambiguation, negation and uncertainty identification, relationship extraction, temporal inferences/relationship extraction, and IE.
- The various approaches applied for processing the clinical text range from simple rule-based methods to more sophisticated statistical, symbolic, or grammatical and hybrid approaches.
- Additionally, many researchers came out with more specific methods for specific NLP problems.

- Some researchers describe a system that leverages cloud-based approaches, i.e., virtual machines and representational state transfer (REST) to extract, process, synthesize, mine, compare/contrast, explore, and manage medical text data in a flexibly secure and scalable architecture.
- The Researchers use a simple, efficient ontology-based approach to extract medical terms present in the clinical text.

- Other researchers present a novel hybrid approach for negation detection by classifying the negations based on the syntactical categories of negation signals and patterns, using regular expression.

Rule-Based Approaches

- Rule-based approaches rely on a set of rules for possible textual relationships, called patterns, which encode similar structures in expressing relationships.
- The set of rules are expressed in the form of regular expressions over words or POS tags.
- In such systems, the rules extend as patterns by adding more constraints to resolve few issues including checking negation of relations and determining direction of relations.
- The rules are generated in two ways: manually constructed and automatically generated from the training dataset.
- Extension with additional rules can improve the performance of the rule-based system to a certain extent, but tend to produce much FP information.
- Thus, the rule-based systems tend to give high precision but low recall because the rules generated for a specific dataset cannot be generalized to other datasets.
- However, the recall of such systems can be improved by relaxing the constraints or by learning rules automatically from training data.

Pattern-Based Algorithms

- The second popular approach for extracting information from the clinical text is the pattern based algorithm.
- A set of word patterns are coded based on the biomedical entities and their relation keywords to extract special kinds of interactions.
- These approaches can vary from simple sentence based extraction to more advanced extraction methods using POS tagging with additional linguistic information.
- Similar to the rule-based approaches, the patterns defined in any pattern-based system are either manually constructed or automatically generated using suitable algorithms such as bootstrapping.
- NLP systems based on manually constructed patterns require domain experts to define the related patterns.
- The concepts knowledge, the list of tokens between these concepts and their POS tags are mandatory to generate more sophisticated patterns.
- Few systems attempt to define patterns using syntactic analysis of a sentence, such as POS tags and phrasal structure (e.g., noun/verb/preposition phrases).
- On the whole, manually generated patterns always tend to produce high precision but low recall. Such patterns do not give good performance when applied for a new domain as well as in text with information not matching with any of the defined patterns.

Machine Learning Algorithms

- Tom Dietterich says that the goal of ML is to build computer systems that can adapt and learn from their experience.
- It is a type of AI that provides computers with the ability to learn without being explicitly programmed.
- The subfields of ML include supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, learning to learn, and developmental learning.
- In general, any ML system takes a set of input variables {x1, x2, ... xm} and gives a set of output variables {y1, y2, ... yn} after learning the hidden patterns {h1, h2, ... hk} from the input variables (Figure).
- ML approaches are not needed when the relationships between all system variables (input, output, and hidden) is completely understood.
- However, this is not the case in almost all real systems.
- There is a wide range of ML approaches and successful applications in clinical text mining.
- Among these, CRF is commonly accepted to perform well for NER and SVM has been proven to be the best classifier by many researchers

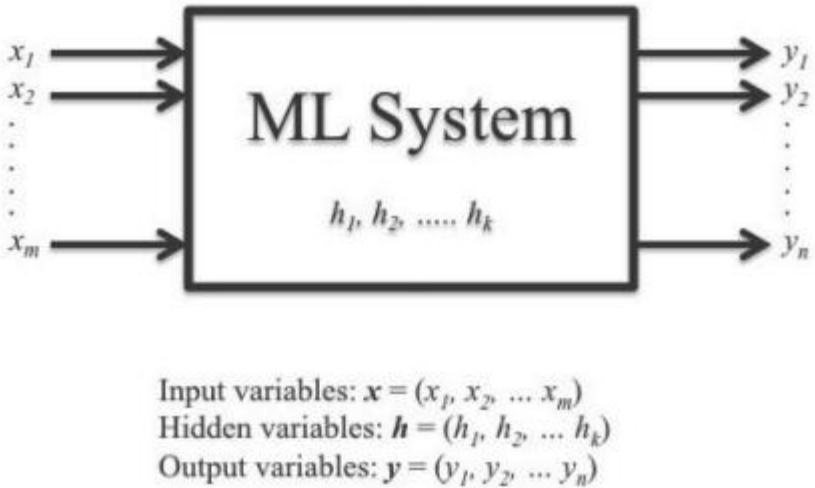


Fig. Block diagram of a generic ML System

Clinical Text Corpora and evaluation metrics

- The widespread usage of EHR in hospitals around the world promotes NLP community to create clinical text corpora for the evaluation of automatic language processing.
- The corpora of clinical 236 Healthcare Data Analytics text are the high quality gold standards annotated manually with the instances relevant to the specific NLP tasks.
- The performance of any NLP system depends heavily on the annotated corpus used for training and testing.
- The corpus for clinical text is developed based on the guidelines for annotation, identifying relevant features to annotate, and on the characterization of usability of the corpus.
- Researchers discuss the task of creating a corpus of layered annotations and developing NLP components for the clinical domain.
- Their annotation layers include Treebank annotations consisting of POS, phrasal and function tags, and empty categories in the tree structure, PropBank annotations to mark the predicate-argument structure of sentences, and UMLS entities for semantic annotations.
- The annotation schema includes 40 clinical reports for training and 20 for testing, by focusing on the semantic categories of the words that are important.
- The corpus was later examined on the agreements among annotators after they were trained with the annotation schema.

- The five qualitative dimensions deciding the usefulness of the corpus in data mining applications are focus, polarity, certainty, evidence, and directionality and developed guidelines on how to annotate sentence fragments.
- However, the difficulty of the annotation varies considerably depending on the dimension being annotated.
- The gold standard is used to test the performance and accuracy of the NLP system developed for a specific clinical task, i.e., retrieving a set of relevant documents automatically.
- The mismatches between the gold standard and system retrieved document set can be due to system error and semantic disagreement between the original text and annotation.
- Thus, it is obvious that even the gold standard annotations available are difficult to interpret against the system-generated results due to the complexity of the language.
- The quality of evaluation is given by three parameters namely, linguistic realism, accuracy, and consistency.
- Realism refers to the set of well-designed tags to bring the same category of words together, based on their similarity in syntactic distribution, morphological form, and/or semantic interpretation.
- Accuracy refers to the percentage of correctly tagged words or tokens in the corpus and calculated as precision and recall.

- Here, precision is the extent to which incorrect annotations are rejected from the output and recall is the extent to which all correct annotations are found in the output of the tagger. In corpus annotation, “correctness” is related to allows and disallows of the annotation scheme that corresponds closely with the linguistic realities recognized.
- The interannotator agreement on manual tagging is defined in terms of a consistency measure to determine the percentage of allows and disallows agreed by the annotators.
- A more sophisticated measure of inter-annotator consistency is given by kappa coefficient (K) to measure the proportion of assigning tags totally by chance. $K = P(A)-P(E)\ 1-P(E)$, where P(A) is the proportion of time that the annotators agree and P(E) is the proportion of times that we would expect them to agree by chance .
- The accuracy of clinical NLP system can be measured with eight standard measures: precision, recall, F-measure, overgeneration, undergeneration, error, accuracy, and fallout.
- The first three are the most common measures widely adopted in reporting the accuracy of a NLP system and defined as follows: In the accuracy measurement of a NLP system for document retrieval, precision quantifies the fraction of retrieved documents that are in fact relevant, i.e., belong to the target class , recall indicates which fraction of the relevant documents is retrieved, and F-measure is the harmonic mean of both for measuring the overall performance of classifiers . Likewise, the other measures are calculated as shown below: overgeneration as 1- precision, undergeneration as 1-recall, error, accuracy, and fallout.

- The accuracy measurements are calculated by categorizing the retrieved information as
- true positive (TP) when the concept is present in the document and found by the system,
- false positive (FP) when the system finds a concept that is not present in the document,
- false negative (FN) when the concept present in the document is not found by the system and
- true negative (TN) when the concept absent in the document is not found by the system.
- Precision = T P (T P+FP)
- Recall = T P (T P+FN)
- F −measure = 2PR (P+R)
- Overgeneration = FP (TP+FP)
- Undergeneration = FN (T P+FN)
- Error = (FN +FP) (T P+FN +FP)
- Accuracy = (T P+TN) (TP+FN +FP+TN)
- Fallout = FP (FP+TN)

Informatics for Integrating Biology and the Bedside (i2b2)

- The Informatics for Integrating Biology and the Bedside (i2b2) is one among the National Institutes of Health (NIH) funded National Center for Biomedical Computing (NCBC) (http: //www.bisti.nih.gov/ncbc/) for developing a scalable informatics framework to bridge clinical research data and basic science research data.
- The integration is helpful for better understanding of the genetic bases of complex diseases and the knowledge facilitates the design of targeted therapies for individual patients with diseases having genetic origins.
- The i2b2 is intended to serve various groups of clinical users including

(1) clinical investigators who are interested to use the software available within i2b2,

(2) bioinformatics scientists who wants the ability to customize the flow of data and interactions, and

(3) biocomputational software developers who involve in the development of new software capabilities that can be integrated easily into the computing environment.

Challenges in processing clinical reports

- The progress in NLP research in the clinical domain is slow and lagging behind when compared to the progress in general NLP due to multiple challenges involved in processing the clinical reports.
- The challenges to NLP development in the clinical domain are mainly due to the lack of access to shared data, lack of annotated datasets for training and benchmarking, insufficient common conventions and standards for annotations, the formidability of reproducibility, limited collaborations and lack of user-centered development and scalability.
- Shared tasks such as the i2b2/VA Challenge address such barriers by providing annotated datasets to participants for finding potential solutions.
- **Domain Knowledge :**
- The most important criteria for an NLP researcher who is involved in the development of systems and methodologies for processing clinical reports is to have adequate knowledge in the domain.
- Many NLP systems are the sound knowledge representation of models for the domain of interest and use the model to achieve a semantic analysis.
- The primary importance of the domain knowledge arises from the fact that the system output is made available for the healthcare application.

- Thus, the system is always expected to have adequate recall, precision, and F-measure for the intended clinical application, with the possible adjustment of the performance according to the needs of the application.
- Interestingly, NLP techniques can be applied to capture the domain knowledge available in the free text. For example, the NLP approach for automated capturing of ontology-related domain knowledge applies a two-phase methodology to extract terms of linguistic representations of concepts in the initial phase followed by the semantic relations extraction.
- **Confidentiality of Clinical Text:** A sample of training dataset is required for the development and testing of an NLP system.
- In a clinical domain, the training dataset is a huge collection of online patient records in textual forms. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) protects the confidentiality of patient data.
- De-identification of personal information is necessary in order to make the records accessible for research purposes.
- However, automatic detection of identifying information such as names, addresses, phone numbers, etc. is a highly challenging task that often requires a manual review.

- There are eighteen personal information identifiers, i.e., protected health information (PHI), in the clinical text required to be de-identified and found to be both time consuming and difficult to exclude as required by HIPAA.
- (Refer the table of 18 personal identifiers from TextBook)

- For example, the abbreviation PT in the clinical text could mean a patient, prothrombin, physical therapy, etc.
- The correct interpretation of clinical abbreviations is often challenging and involves two major tasks: detecting abbreviations and choosing the correct expanded forms.
- The most commonly employed methods for detecting abbreviations in the clinical domain are dictionary lookup and morphology-based matching and for choosing the correct expanded form is with machine-learning approaches.
- Researchers have contributed several methods to identify abbreviations present in the clinical texts, construct a clinical abbreviation knowledge base, and disambiguate ambiguous abbreviations.
- Furthermore, the clinical NLP systems such as MedLEE, MetaMap, etc., are developed to extract medical concepts and related abbreviations from the clinical texts.

Diverse Formats: There is no standardized format for the clinical text, especially with the medical reports of patients:

(1) The clinical text often contains the information in free-text format and as a pseudo table, i.e., text intentionally made to appear as a table. Though the contents of a pseudo table are easy to interpret by a human, it is very problematic for a general NLP program to recognize the formatting characteristics.

(2) While the sections and subsections of the reports are important for many applications, in many occasions the section headers are either omitted or merged to similar headers.

(3) Another problem commonly observed in the clinical text is the missing or inappropriate punctuation, i.e., a new line may be used instead of a period to signify the end of a sentence.

The clinical document architecture (CDA) that aims to establish standards for the structure of clinical reports addresses the problem of diverse formats related to clinical text effectively .

- **Expressiveness:**
- The language in the clinical domain is extremely expressive. The same medical concept can be described in a numerous ways, i.e., cancer can be expressed as tumor, lesion, mass, carcinoma, metastasis, neoplasm, etc.
- Likewise, the modifiers of the concept can also be described with many different terms, i.e., the modifier for certainty information would match with more than 800 MedLEE lexicons, thus making the retrieval process more complicated.
- **Intra- and Interoperability:** A clinical NLP system is expected to function well in different health care as well as clinical applications and is easy to integrate into a clinical information system. In other words, the system needs to handle clinical text in different formats.
- For example, the formats of discharge summaries, diagnostic reports, and radiology reports are different. Furthermore, the NLP system is required to generate output that can be stored in an existing clinical repository. However, due to the complexity and nested relations of the output, it is almost unlikely to map the same to the clinical database schema.

- Additionally, the output from the NLP system is required to be available for comparison through widespread deployment across institutions for a variety of automated applications.
- To achieve this, the output needs to be mapped to a controlled vocabulary such as UMLS, ICD-10, SNOMED-CT, and to a standard representation for the domain.
- Finally, the construction of a representational model is considered to be essential to interpret the clinical information and relations between the concepts. For example, one of the relations between a drug and disease is “treats.”

Interpreting Information Interpretation of clinical information: available in a report requires the knowledge of the report structure and additional medical knowledge to associate the findings with possible diagnoses.

- The complexity involved in interpreting information depends on the type of the report and section, i.e., retrieving information on the vaccination administered is more straightforward than retrieving information from a radiological report that contains patterns of lights (patchy opacity).
- An NLP system to interpret the patterns of lights to a specific disease should contain medical knowledge associated with findings.

Clinical Applications

- NLP and data mining for clinical text mining is applied to discover and extract new knowledge from unstructured data.
- Mining information from clinical text includes finding association patterns such as disease-drug information and discharge summaries by applying techniques from NLP, data mining, text mining, and even statistical methodologies.
- **General Applications:** NLP together with IR and IE approaches has been widely employed in a variety of clinical applications such as summarizing patient information in clinical reports , extracting cancer-related findings from radiology reports , ICD-10 encoding in a discharge summary , SNOMED encoding in a discharge summary , and many more.
- **EHR and Decision Support:** The use of EHR in a hospital to store information about patients' health along with the details of drug usage, adverse effects, and so on requires the implementation of NLP to process large volumes of data such as discharge summaries. The clinical information in EHR is available as free-text as a result of transcription of dictations, direct entry by the providers, and use of speech recognition applications. While the information in free-text is convenient to express concepts and events, it is more complicated for searching, summarization, decision support, and statistical analysis.

- NLP techniques are found to be highly successful to process EHR in terms of reducing errors and improving quality control and coded data.
- Many CDS systems have been developed in recent years to process and extract information from EHR.
- The goal of CDS is to help the health professionals in making clinical decisions through the interpretation of information available in EHR, i.e., to know the best possible treatment for a specific disease.
- In general, CDS is defined as any software that provides clinical decision making by matching the characteristics of an individual patient's information in EHR such as laboratory results, pharmacy orders, discharge diagnoses, radiology reports, operative notes, etc., with the computerized knowledge base, to provide patient specific assessments or recommendations.
- Patients' medical history includes laboratory results, pharmacy orders, discharge diagnoses, etc., in EHR, which can be entered manually into a CDS system by clinicians.
- However, NLP is required to retrieve the required information from the data.
- Besides, NLP is also able to represent clinical knowledge and CDS interventions in standardized formats.
- In other words, if CDS systems depend upon NLP, it would require reliable, high-quality NLP performance and modular, flexible, and fast systems.

- Such systems either are active NLP CDS applications that push the patient-specific information to users or passive NLP CDS applications that require input from user to generate the output.
- While active NLP CDS includes alerting, monitoring, coding, and reminding, passive NLP CDS focuses on providing knowledge and finding patient populations.
- Though the NLP CDS is meant for retrieving clinicians' information needs, the other active users of the system are researchers, patients, administrators, students, and coders.
- **Surveillance:**
- The process of collecting, integrating, and interpreting information related to a specific disease is called surveillance.
- The activities of surveillance for public health professionals vary from standard epidemiological practices to advanced technological systems with more complicated algorithms.
- The health care officials are expected to have awareness of surveillance programs at the federal, state, and local levels.
- The National Strategy for Biosurveillance (NSB) brings together the government, private sectors, non-government organizations, and international partners to identify and understand the health-related threats at an early stage to provide accurate and timely information.
- On the other hand, the National Association of County and City Health Officials (NACCHO) supports the local surveillance by sharing critical information systems and resources to identify and prevent the spread of a disease in an effective and timely manner.