A.I. in health informatics **lectures 9&10** natural language processing and biomedical texts

kevin small &
byron wallace

today

- natural language processing (NLP)
  - applications
  - techniques
- clinical and biomedical language

natural language

- humans prefer
  - scientific literature
  - technical reports
  - administrative reports
  - patient charts
  - spoken language transcriptions

structured data

- computers prefer
  - measurements in a spreadsheet
  - predefined lists (*e.g.*, diseases, genes)
  - patient data
  - billing/administrative information

natural language processing (NLP)

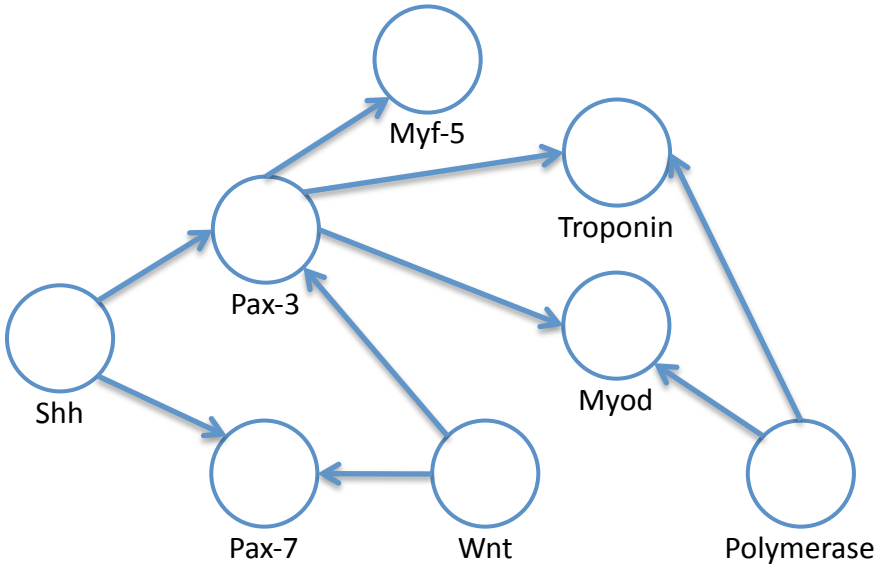
- narrative data ➔structured data
  - narrative data inefficient to (re)process
  - narrative data rife with ambiguity and variance in expression
  - structured data not always sufficient
- NLP abstracts narrative information into a structured form
- NLP is **hard**

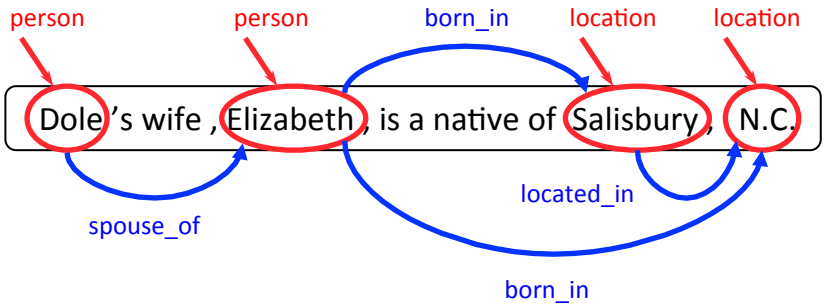
NLP goals

- time
  - PubMed has ~20M documents
  - patient documents
- consistency / objectivity
  - rules can be updated by experts
  - classifiers generalize to new data
- cost
  - labor with specialized training

information extraction (IE)

- locates important structures
  - named entities
  - relations



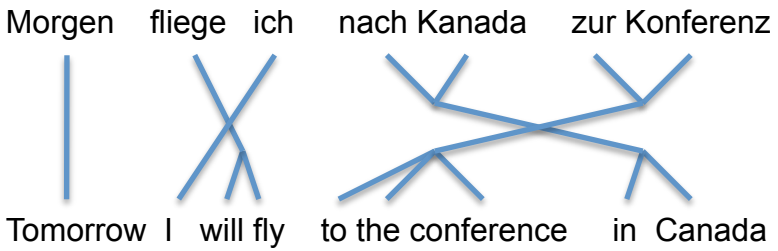


information retrieval (IR)

- access documents in large corpora
- satisfy information need of query
- term indexing
- phrase/entity/relation indexing

machine translation

- converts text in one language to another language
- study enrollment
- scientific literature



et cetera

- text generation
- summarization
- automatic editing
- user interfaces
- speech transcription
- use your imagination…

levels of knowledge

- morphology
  - morphemes generate words
- lexicography
  - global properties of words
- syntax
  - structure of phrases and sentences
- semantics
  - interpretation of linguistic structures
- pragmatics
  - understanding discourse

NLP requirements

- specifying representation
- method of acquiring knowledge to generate representation
- algorithms to support applications based on specified representation

NLP techniques

- symbolic/logical methods
  - finite state machines
  - context-free grammars (CFGs)
  - informative, brittle
- statistical methods
  - Markov models
  - probabilistic context-free grammars (PCFGs)
  - often less interpretable, robust
  - discriminative methods

NLP paradigms

- parsing
  - linguistic analysis from latently structured information to explicit structure
- generation
  - use linguistic/statistical models to generate natural language
- extraction
  - extract relevant information
  - doesn’t necessarily require full analysis

morphology

- morphemes (roots, prefixes, suffixes) are used to generate words
  - free vs. bound
  - inflectional (e.g. bigg-er)
  - derivational (e.g. judg-ment)

- biomedical data morphologically richer than general English
  - hydr-oxy-nitro-di-hydro-thym-ine
  - hepatico-cholangio-jejuno-stom-y

tokenization

- parsing string into tokens
  - sentence splitting often first step
  - includes words, numbers, symbols

q.i.d. → four times a day

M03F4.2A → gene name

(w)adh-2 → biological named entity

tokenization

5 mg. given.

- regular expressions

[a-z]+('s)?|[0-9]+|[.]

- Markov model p(5*mg.*given*.) versus p(5*mg*.*given*.)

	5	mg	mg.	given	.
5	0.1	0.8	0.9	0.4	0.6
mg	0.3	0.1	0.1	0.9	0.4
mg.	0.3	0.1	0.1	0.9	0.2
given	0.7	0.6	0.6	0.2	0.7
.	0.6	0.4	0.4	0.8	0.1

- hybrid systems

lexicography

- atomic elements of language
  - multi-word expressions (MWE)
    - foreign phrases (ad hoc)
    - prepositions (along with)
    - idioms (follow up)
    - clinical MWE (congestive heart failure)
- parts of speech (POS)
  - inflectional morphemes
    - number
    - person
    - case

rule-based POS tagging

- based on *transformation rules*

NN ➔ VB if previous tag is TO
NN ➔ JJ if following tag is NN

Before Rule Application	After Rule Application
total/NN hip/NN replacement/NN	total/JJ hip/NN replacement/NN
a/DT total/NN of/IN four/NN units/NNS	*no change*
refused/VBD to/TO stay/NN	refused/VBD to/TO stay/VB
her/PP$ hospital/NN stay/NN	*no change*
allergy/NN to/IN penicillin/NN	*no change*

- can be learned (TBL)

Markov model POS tagging

	NN	VB	VBD	VBN	TO	IN
NN	0.34	0.00	0.22	0.02	0.01	0.40
VB	0.28	0.01	0.02	0.27	0.04	0.39
VBD	0.12	0.01	0.01	0.62	0.05	0.19
VBN	0.21	0.00	0.00	0.03	0.11	0.65
TO	0.02	0.98	0.00	0.00	0.00	0.00
IN	0.85	0.00	0.02	0.05	0.00	0.08

NN NN VBD TO VB NN
NN VB VBN TO VB NN
NN VB VBN IN VB NN

- can also be lexicalized (HMM)

discriminative POS tagging

- based on structured classification
  - most common tag
  - tag distribution
  - previous tag
  - previous word
  - two words previous
  - two tags previous
  - next word
  - bigram previous…
- is the state of the art

syntax

- structure of phrases and sentences
  - lexemes form phrases
    - noun phrases (severe chest pain)
    - adjectival phrases (painful to touch)
    - verb phrases (has increased)
  - phrases form sentences
- clinical text is *telegraphic*
  - constitutes a sublanguage

symbolic parsing

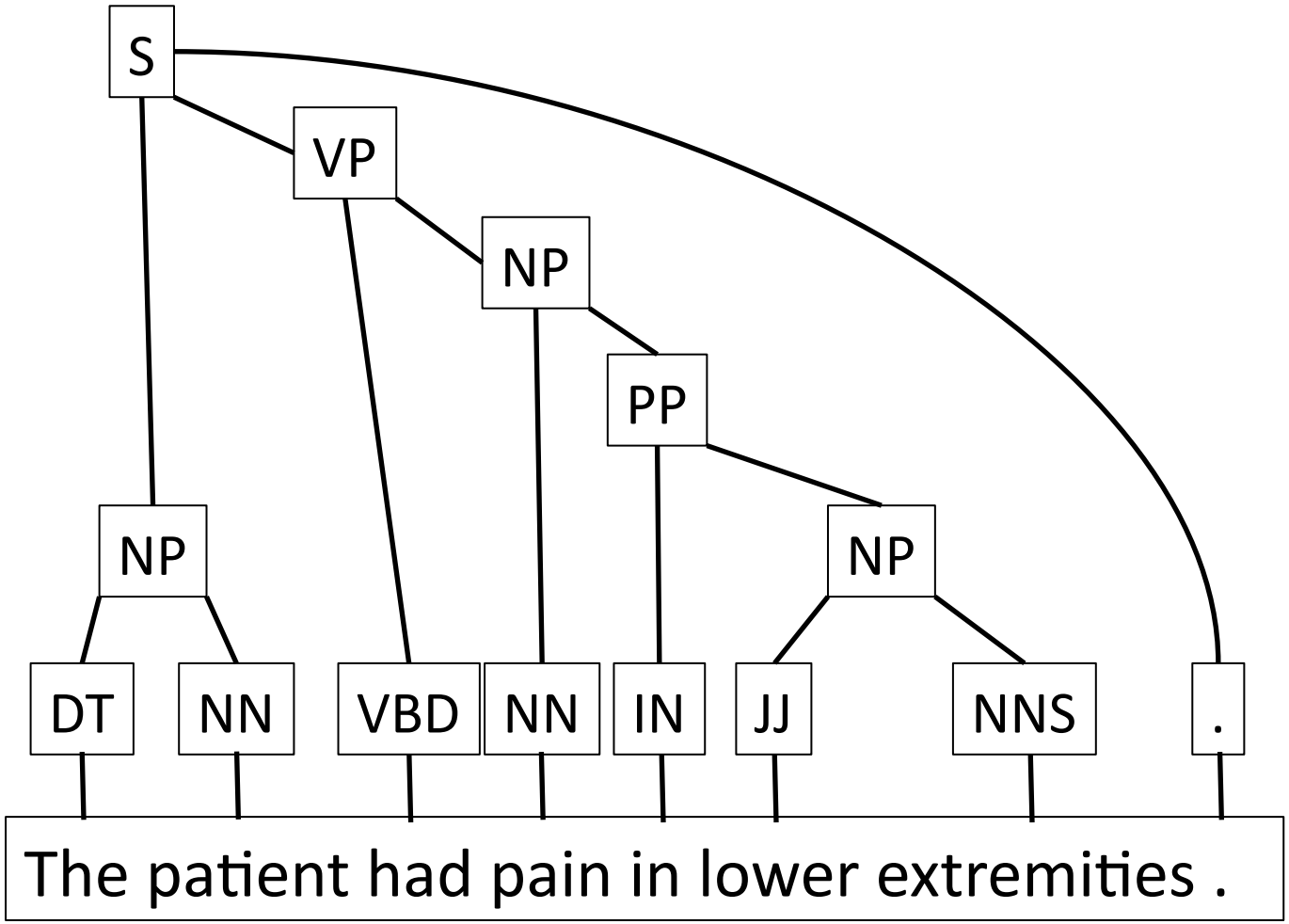
- regular expressions

DT? JJ* NN* (NN|NNS)

- context-free grammars (CFGs)

S → NP VP .
NP → DT? JJ* (NN|NNS) CONJ* PP* | NP and NP
VP → (VBZ|VBP) NP? PP*
PP → IN NP
CONJ → and (NN|NNS)

parse trees


S
VP
NP
PP
NP
NP
DT
NN
VBD
NN
IN
JJ
NNS
.
The patient had pain in lower extremities .

probabilistic CFGs (PCFGs)

S → NP VP .
NP → DT?$^{0.9}$ JJ*$^{0.8}$ (NN$^{0.6}$|NNS) PP*$^{0.8}$
VP → (VBZ$^{0.4}$|VBP) NP?$^{0.9}$ PP*$^{0.7}$
PP → IN NP

X-ray shows patches in lung.

state of the art parsing

- discriminative
  - structured prediction algorithm
- lexicalized
- active research area
- limited to newswire (Penn Treebank)

semantics

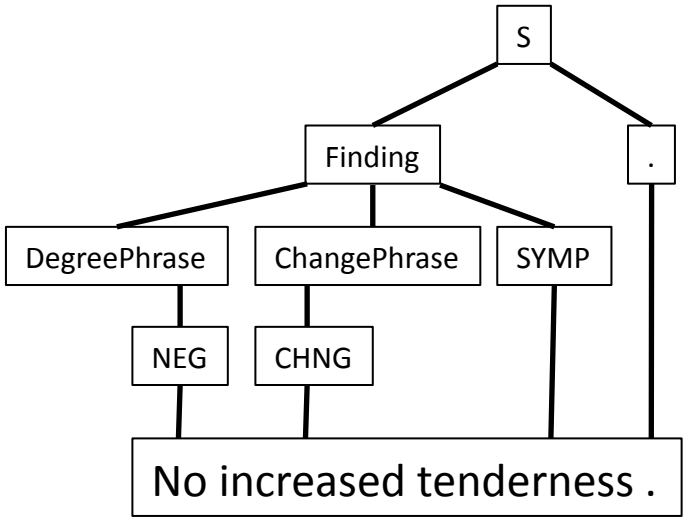
- interpretation of linguistic structures
- word sense (*e.g.*, bank, capsule)
- semantic types
  - medication, gene, disease, etc.
- semantic roles
  - medication-treats-disease, etc.
- biomedical vs. general language

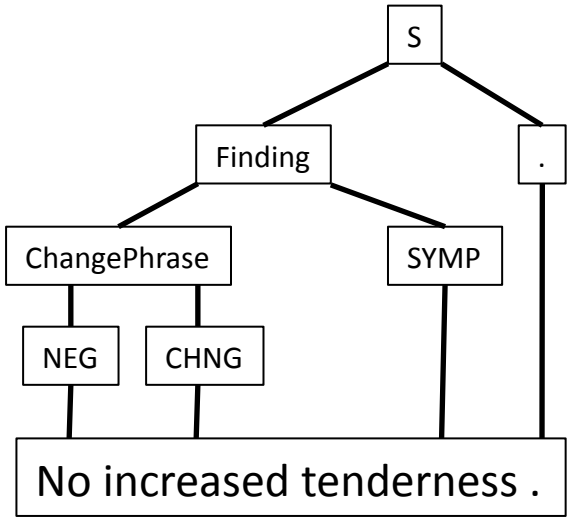
semantics

- external lexicon (*e.g.*, UMLS)
- morphological analysis
  - "-itis" and "-osis" are diseases
  - "-otomy" and "-ectomy" are procedures
- word sense disambiguation
  - same routine as syntax
- semantic role labeling (SRL)
  - regexp, semantic grammar

semantic parse

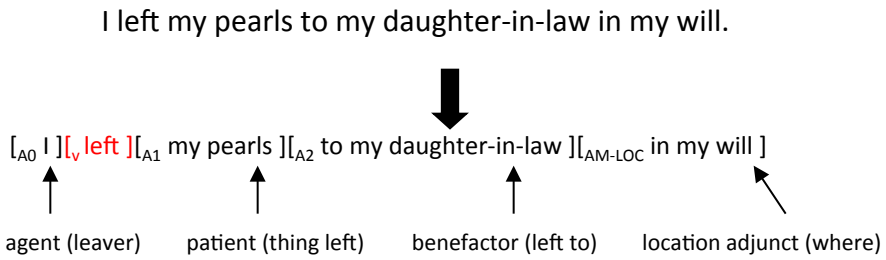
S ➔ Finding .
Finding ➔ DegreePhrase? ChangePhrase? SYMP
ChangePhrase ➔ NEG? CHNG
DegreePhrase ➔ DEGR | NEG





state of the art SRL

- discriminative
  - structured prediction algorithm
  - pipelined (attempts to collapse)
- lexicalized



- external knowledge

pragmatics

- structure of discourse
  - word and phrase senses
    - "mass" (mammography vs. radiology)
    - "drinks" (health care vs. life)
  - reference attachment
    - co-reference

> An infiltrate was noted in right upper lobe; it was patchy.

  - narrative centering

clinical language

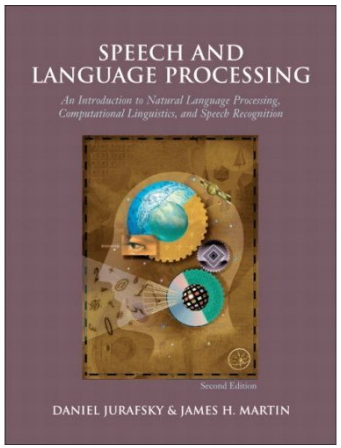
- requires high sensitivity & specificity
- lacks contextual features
- telegraphic morphological features
- global context necessary for significant ambiguity resolution
- lack of direct measurement
- standards

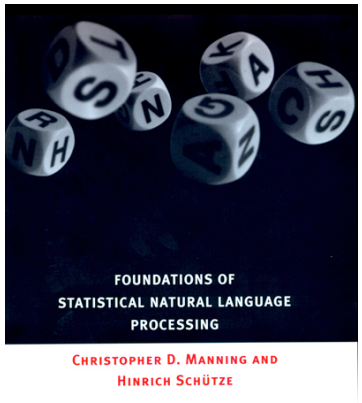
biomedical language

- time-varying
- morphological nesting
- syntactical and semantic nesting
- syntax wildly different than "standard" English

resources

- controlled vocabularies / lexicons
  - UMLS
  - SNOMED, ICD-9
  - biological databases (Flybase)
  - GENIA





take away

- representation is crucial
  - "deeper" analysis useful
  - "deeper" analysis noisy
- symbolic methods
  - relies on expert information
  - brittle, but state of the art for clinical
- statistical methods
  - relies on labeled data
  - robust, seemingly the future

next lecture

- information extraction
  - named entities
  - relation extraction