Overview of Natural Language Processing (NLP) Approaches

Robert Carroll

Some slides adapted from Josh Denny

What is NLP?

- Natural Language Processing is a field around extracting information from text.
- Natural language is comprised of our writing: emails, books, clinical notes, etc.
- It's often used to describe a broad range of techniques:
  - Keyword searches
  - Regular expressions
  - Concept indexing

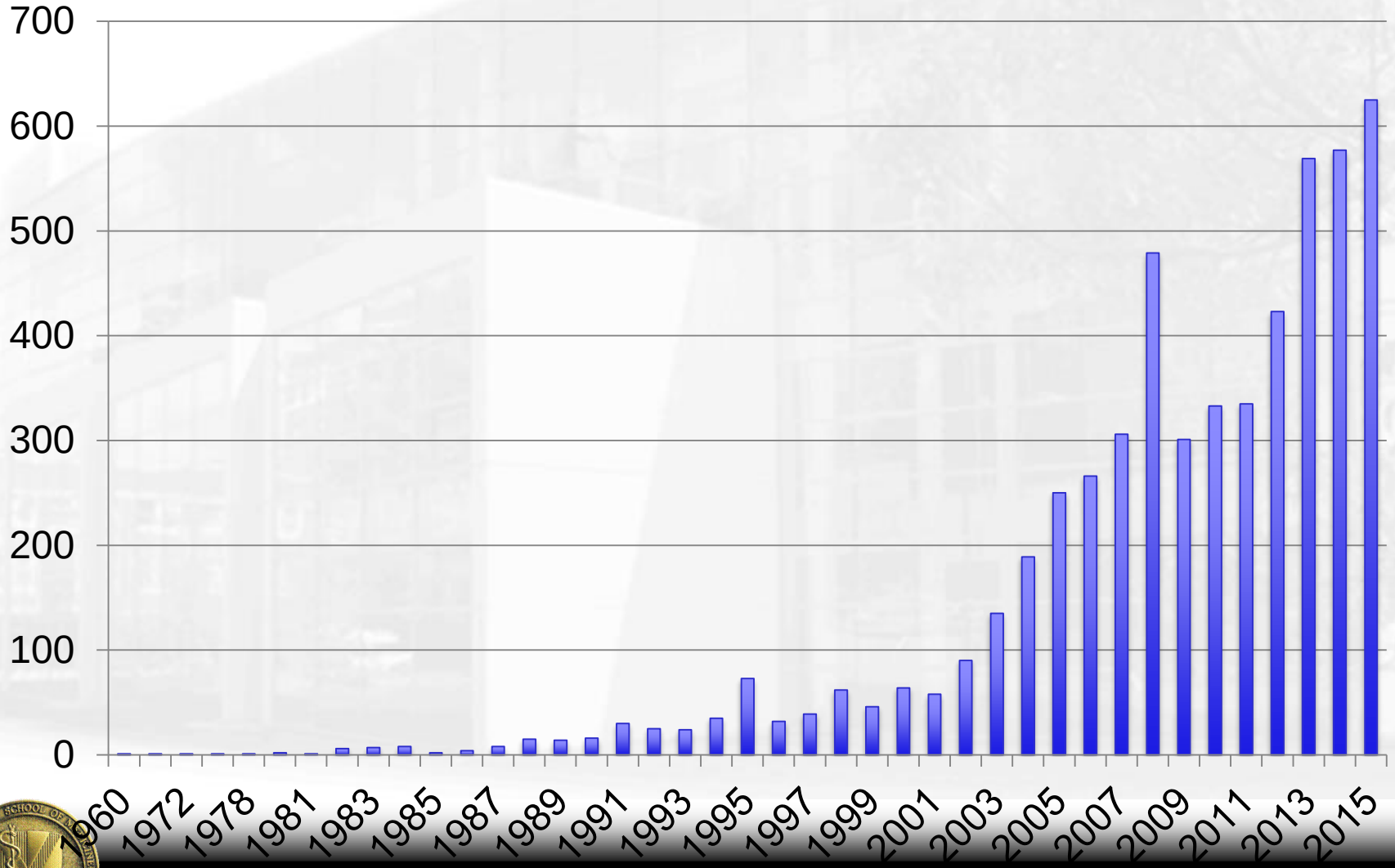
Why Biomedical NLP?

- Natural language text is the primary means of communication in health care and biological research
  - Clinical notes: VUMC 1.7M patients > 10 years data, >120M documents
    - To build intelligent processes on top of NLP structured output
  - Biomedical literature: >22M articles in MEDLINE
    - To develop knowledge bases, to keep up-to-date

Complexity of Biomedical Text

- Clinical Notes
  - Heterogeneous report structures
  - Telegraphic text formats
  - Abbreviations
- Biomedical Literature
  - Complicated sentences structures
  - Biomedical entities
    - Large & dynamic: gene, protein, cell, tissue, organism, species, diseases….
    - Cross different domains: biology, physics, chemistry …

Biomedical NLP Trends



"Major tasks" in clinical NLP

- Concept identification
- Negation detection
- Section identification
- Word sense disambiguation
- Temporal resolution

Concept vs. Text indexing

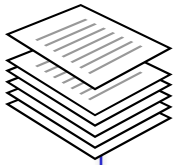
- Text indexing / keyword searches
  - Indexing by words of document
  - "Hepatolenticular degeneration" ≠ "Wilson's Disease"
- Concept indexing / Natural language processing
  - Recognizes words in document to a controlled vocabulary
  - "Hepatolenticular degeneration" = "Wilson's disease"
- Further Improvements
  - Negation detection – "**no** chest pain"
  - Section tagging – was it the chief complaint or the family medical history?
  - Temporal detection – "colonoscopy **5 years ago**"

Controlled Vocabularies

- Unified Medical Language System (UMLS)
  - Contains >100 individual vocabularies
    - Incl. SNOMED, LOINC, RxNorm… and many others
  - Sometimes "messy" and imprecise, but most "robust" coverage
- SNOMED-CT
  - current "gold standard" for medical problems and signs/symptoms
  - Well-defined, curated relationships
- LOINC – "gold standard" for labs and tests
- RxNorm – currently most robust freely available drug database

Interpreting Natural Language Text

Clinical Notes

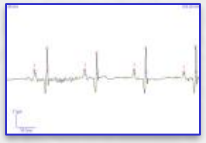


SecTag (Note Section Tagger)

Document organized by sections

```
<chief_complaint>
   SOB
</chief_complaint>
<history_present_illness>
    …no chest pain…
</history_present_illness>
```

Rad, Path Reports



Negation Tagging

KnowledgeMap Concept Identifier

Status and Date detection

```
<chief_complaint>
   C0392680: Shortness of Breath,
                Asserted
</chief_complaint>
<history_present_illness>
    C0008031: Chest Pain, Negated
</history_present_illness>
```

Structured Output

Text labeled with **Unified Medical Language System** concepts organized by section and certainty (positive, possible, negated)

Challenges to effective mapping

- Acronyms:
  - "CHF" – "Congestive Heart Failure" or "Congenital Hepatic Fibrosis"?
  - "BSE" – "Bovine spongiform encephalopathy" or "Breast self exam"?
- Pseudo-synonyms ("lungs" and "pulmonary")
- Underspecified terms ("overmatch")

Document word	Bad UMLS match	Desired UMLS match
ST	"street"	"ST *segment*"
gram-negative infection	"gram-negative" + "infection"	gram-negative *bacterial* infection

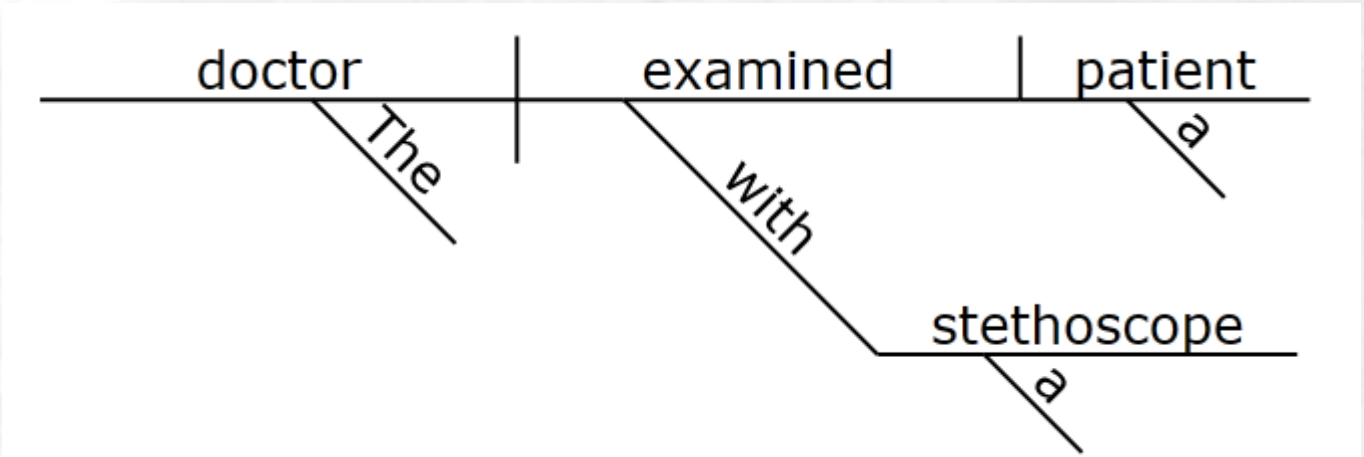
- Outline headers and poor sentence structure

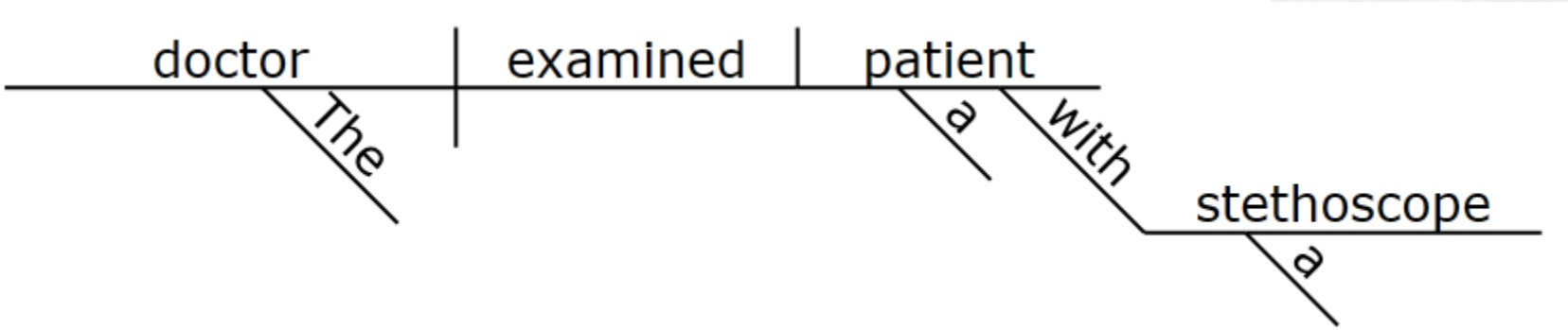
"`III. Aminoglycosides IV. Metronidazole`" **vs.** "`The patient was given IV metronidazole…`"

- Sentences defining concepts ("The aortic valve was stenosed").

What's the challenge of determining syntax?

The doctor examined a patient with a stethoscope.





Specific-focus NLP

- Smoking use
- Obesity comorbidities
- Medication information
- Diseases and treatments
- Temporal reference

If it's such a hassle…

- While there are many challenging aspects to NLP, we wouldn't try it if it wasn't necessary.
- Structured information, such as labs or billing information, can only get us so far.

Identifying cases with precision from the EMR

requires structured and unstructured information

	RA	MS	CD	T2D
Had ICD-9 codes:	3.9%	1. 8%	1.8%	17.3%
Met algorithm definition:	2.7%	1. 2%	1.6%	9.7%
Accuracy of ICD9 codes:	69%	66%	89%	56%

Demonstration Project:
Validating EMR-derived genotype-phenotype studies

Disease	Methods	Definite Cases	Controls	Case PPV	Control PPV
Atrial fibrillation	NLP of ECG impressions ICD9 codes CPT codes	168	1695	98%	100%
Crohn's Disease	ICD9 codes Medications (text)	116	2643	100%	100%
Type 2 Diabetes	ICD9 codes Medications (text) Text searches (controls)	570	764	100%	100%
Multiple Sclerosis	ICD9 codes or text diagnosis	66	1857	87%	100%
Rheumatoid Arthritis	ICD9 codes Medications (text) Text searches (exclusions)	170	701	97%	100%

When do I try NLP?

- We have talked a lot about the general topics of NLP, but what about when should you consider it?
- In particular, what approach is appropriate?
  - Keyword searching
  - Concept indexing
  - Regular expressions (pattern matching)

Text matching / Keywords

Pros

- Straight forward
- Can be done in the interface
- Casts a broad net
  - Can be used if you intend to further review who you ID
  - Can be used if you have a rare condition you want to exclude on

Cons

- Limited flexibility
- Misspellings, etc
- Casts a broad net
  - No sections* means you will likely see family history
  - No negation means you will see “risk of” and other modified mentions

Concept Indexing

Pros

- Can be done once for all individuals
- Very easy to use programmatically
- Can identify FHx, negated mentions, risk mentions, etc which may be relevant

Cons

- Requires a lot of computational time
- Cannot be done in the interface
- Disambiguation is not perfect
- Misspellings, etc, can still cause problems
- You must select your terms from the mapped terminologies

Regular Expressions

Pros

- Incredibly powerful
- Very specific
- You can extract other information, like ejection fractions estimates

Cons

- Requires a lot of development
- Can be incredibly complicated
- Updates may be required as documents change
- May not be applicable to other sites, note types, etc.

Keyword searches

- Useful when:
  - You have limited resources
  - You will further review individuals
  - You believe your text will be found reliably in certain places, eg problem lists
  - You want rough estimates of prevalence
  - You need specificity not found in other sources (though still may require further review!)

Concept indexing

- Useful when:
  - You will need to survey a lot of topics
  - You need to know negation, FHx, or other statuses of the text
  - You want to work with codified data
    - Portability
    - Scalability
    - Programmatic work

Many types of CI

- KnowledgeMap Concept Indexer
  - Developed locally by Josh and others
- cTAKES
  - Mayo and Partners
- MetaMap
  - National Library of Medicine
- Medex
  - Hua Xu and others

KMCI output format

- Cui|cui name|semantic type|sentence number|word number|score|attributes|original text|section string|note subtype
- 33860|'Psoriasis'|'Disease or Syndrome'|38|8|1|other_experiencer|psoriasis|158/history_present_illness|VMG FRANKLIN RHEUMATOLOGY CLINIC VISIT

Multiple Sclerosis

CUI	Name
270784	'Multiple sclerosis of the brainstem (disorder)'
393665	'Chronic progressive multiple sclerosis (disorder)'
751965	'Secondary progressive multiple sclerosis (disorder)'
393664	'Acute relapsing multiple sclerosis (disorder)'
2586221	'Malignant multiple sclerosis (disorder)'
581392	'Exacerbation of multiple sclerosis (disorder)'
751964	'Primary progressive multiple sclerosis (disorder)'
338475	'Multiple sclerosis of the spinal cord (disorder)'
2585570	'Benign multiple sclerosis (disorder)'
751967	'Relapsing remitting multiple sclerosis (disorder)'
393666	'Remittent-progressive multiple sclerosis (disorder)'

Regular Expressions

- Useful when:
  - You need to extract numbers or designations
  - The notes or reports follow a specified format
  - Your target word or phrase has many variations not likely to be captured by a general purpose system

What are RegEx?

- Regular expressions are a way of defining complex search patterns.
- Its simplest form is the use of wildcards:
  - From: *@Vanderbilt.edu
  - Find my_csv*.doc
- They can be incredibly complicated (and potentially powerful!).
- http://regexr.com/

Finding exercise stress tests

- What if we want to know the results of an exercise stress test?
  - Resting and stress heart rate
  - Resting and stress blood pressure
  - Exercise duration
  - Metabolic equivalents of task (METs)
- These numeric values won't easily be captured by other means.

Easy extraction

- If notes have a simple format, this isn't a tough job.
- "Stress test results" notes contain nice lines like:
  - METS: 5
  - Heart Rate (rest): 70
  - Systolic Pressure (peak): 150 mmHg

Easy extraction

- METS: 5
- METS:\s+([0-9.]+)
- Grab a number following the “METS:”
- The same approach works with the others as well!

Hard extraction

- What if the values you need are found in free text, not tables?
  - Heart rate rose from 70 to 150
  - 70 bpm at rest to maximum heart rate of 120 bpm
- What if there are wording changes?
  - Heart rate increased from 70 to 150
  - 70 at rest to 120 beats per minute
- What about typos?

Hard extraction

- For cardiac perfusion tests, I had to create different versions of the matching criteria for different examples.
- One may also need to clean up the text, making sure “beats per minute”, “beats-per-minute”, and “bpm” all look the same.
- Same goes with spacing, lines, mmHg, etc.

Hard extraction

- ([0-9]+) (?:bpm)?(?:at rest |rose)?to (?:(?:maxim(?:al|um))?heart rate of)?([0-9]+) bpm
- Heart rate (?:changed|increased|rose) from ([0-9]+) to ([0-9]+)

Cover your eyes extraction

```
(?:(?<pre>\.|normal|\d\d(?:\.\d+)?\s*(?:%|percent)
?(?:\s*(?:-
+|to)?\s*\d\d(?:\.\d+)?\s*(?:%|percent)?)?)(?<predi
st>.*?))?(?:(?:left ventricular )?ejection(?:
fraction)?|\bEF\b|\bLVEF\b)(?::)?(?: of ?)?(?:
approx(?:imately)? ?)?(?: greater than | ?&gt; ?|
?>
?)?(?:(?<postdist>.*?)(?<post>\.|normal|\d\d(?:\.\d
+)?\s*(?:%|percent)?(?:\s*(?:-+|(?:increased
)?to)?\s*\d\d(?:\.\d+)?\s*(?:%|percent)?)?\b))?
```

Which to use?

- Keywords
  - You need: something simple
  - You have: yourself
- Concept Indexing
  - You need: broad survey results
  - You have: computational time
- Regular Expressions
  - You need: specialized text extraction
  - You have: a developer (or funds)

Time

- All NLP takes time.
- Review is helpful no matter what approach you go.
- There's a tradeoff between the costs- sometimes doing by hand is easiest.