Natural Language Processing: Fundamentals and Pitfalls

Alexander Turchin, MD MS
Quality Performance Management
Partners HealthCare
Outline

• What is NLP?
• Why NLP?
• How to NLP
• Design and validation of NLP tools
• Medical NLP tools in public domain
What is Natural Language Processing?
Mr. Smith comes today with chief complaint of back pain. Denies history of trauma, urinary retention or weakness.
Why Natural Language Processing?
Why NLP?

<table>
<thead>
<tr>
<th>SubjectID</th>
<th>DiseaseCode</th>
<th>DrugCode</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345</td>
<td>ABCD</td>
<td>1a2b</td>
</tr>
<tr>
<td>23456</td>
<td>BCDE</td>
<td>2b3c</td>
</tr>
<tr>
<td>34567</td>
<td>EFGH</td>
<td>3c4d</td>
</tr>
<tr>
<td>45678</td>
<td>FGHI</td>
<td>4d5e</td>
</tr>
</tbody>
</table>

P < 0.0001

c/o cough x 2 mon no fever, chills, SOB no hemoptysis Patient is concerned about pneumonia.

HISTORY:
Cough x 2 months

FINDINGS:
No acute cardiopulmonary process
Structured and Narrative Data are Complementary

13,993 Medication intensifications

9,819 Medication intensifications

5,627 (30.9% of the total) Medication intensifications

How to NLP
NLP: Methodology

NLP Language Models

Deterministic
- Based on a set of rules
- Easier to model complex structures
- May be more effective in identifying sparse concepts
- Labor-intensive approach; does not scale up

Probabilistic
- Based on a statistical model
- “Black Box”
- At least as effective in identifying common concepts
- Easier to scale up
NLP: Common Tasks

Sentence boundary detector

Tokenization

Stemming

Disambiguation

Mr. Smith complains of chest pains.

{Mr.} {Smith} {complains} {of} {chest} {pains}

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Mitral regurgitation

Mental retardation

Mister
NLP: Common Tasks

Sentence boundary detector

Mr. Smith complains of chest pains.

Tokenization

{Mr.} {Smith} {complains} {of} {chest} {pains}

Normalization

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Disambiguation

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Part-of-speech tagger

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

noun

verb

preposition

noun

noun

noun
NLP: Common Tasks

Sentence boundary detector

Mr. Smith complains of chest pains.

Tokenization

{Mr.} {Smith} {complains} {of} {chest} {pains}

Normalization

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Disambiguation

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Part-of-speech tagger

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

Concept mapper

{Mr.} {Smith} {complain}s {of} {chest} {pain}s

chest pain (finding)
SNOMED 29857009
NLP: Common Tasks

Sentence boundary detector

Mr. Smith complains of chest pains.

Tokenization

\{Mr.\} \{Smith\} \{complains\} \{of\} \{chest\} \{pains\}

Normalization

\{Mr.\} \{Smith\} \{complain\} s \{of\} \{chest\} \{pain\} s

Disambiguation

\{Mr.\} \{Smith\} \{complain\} s \{of\} \{chest\} \{pain\} s

Part-of-speech tagger

\{Mr.\} \{Smith\} \{complain\} s \{of\} \{chest\} \{pain\} s

Concept mapper

\{Mr.\} \{Smith\} \{complain\} s \{of\} \{chest\} \{pain\} s

Negation detection

Mr. Smith complains of chest pains but \textbf{no} dyspnea.

Reference resolution

\{Mr.\} \{Smith\} \{complain\} s \{of\} \{chest\} \{pain\} s
Design and Validation of Natural Language Processing Tools
Design of NLP Tools

- Improving speed
- Customized concept classes
- Enrichment of the data sources
Speed: Hierarchical Processing

Full analysis

Document Collection

Note-level scan

Sentence-level scan

Mandatory sentence-level concepts found

Mandatory note-level concepts found
Customized Concept Classes

<table>
<thead>
<tr>
<th>Literal meaning</th>
<th>Start, increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical meaning</td>
<td>Prescribe</td>
</tr>
<tr>
<td>Situational meaning</td>
<td>Call in</td>
</tr>
<tr>
<td>Misspellings, abbreviations</td>
<td>Increase, incr.</td>
</tr>
</tbody>
</table>
Enrichment of the Data Sources
Enrichment of the Data Sources

- Non-adherence to medications
  - Significantly elevated BP ($\geq 150/100$)
  - No intensification of anti-hypertensive medications

- Blood pressures measured at home
  - Notes with blood pressure ranges
    (e.g. 120-130/70-80)
• Stratified validation of important concept classes
• Unbiased validation
• Concurrent vs. Predictive (clinical) validation
Unbiased Validation

CORRECT

INCORRECT

Compared to manual review

Correction of reviewer rating errors concordant with software (e.g. both missed)

Correction of reviewer rating errors discrepant with software (e.g. reviewer missed, software didn’t)
Stratified Validation

Active Medications

Inactive Medications

NLP

DETECTED

MISSED
Concurrent Validation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Change at 3 Months</th>
<th>Atorvastatin Only</th>
<th>Torcetrapib plus Atorvastatin</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipids (absolute change) — mg/dl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>+1.6±20.5</td>
<td>+5.1±23.9</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>High-density lipoprotein</td>
<td>+0.5±6.2</td>
<td>+29.0±14.4</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Low-density lipoprotein</td>
<td>+0.6±15.8</td>
<td>-20.5±20.8</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

A Death from Any Cause

No. at Risk
Atorvastatin only 7534 7530 7521 7509 7487 5833 4043 2078 956 109
Torcetrapib plus atorvastatin 7533 7526 7511 7494 7464 5827 4049 2069 943 114

## Concurrent Validation

<table>
<thead>
<tr>
<th>Measure</th>
<th>Agreement</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx intensification</td>
<td>92.0 ±6.2%</td>
<td>83.8 ±4.2%</td>
<td>95.0 ±3.7%</td>
</tr>
</tbody>
</table>

+/- 95% CI

Predictive Validation

Medication intensifications / month

Average monthly change in SBP, mm Hg

Medical NLP Tools in Public Domain
Automated Retrieval Console

• Input: manually rated documents
• [Near] Instantaneous Output:
  – Individual concepts
  – Document-level classification
• Off-the shelf tool pipeline
• Developed by Leonard D’Avolio at Boston VA
• More information at:
  http://research.maveric.org/mig/arc.html
cTakes

- clinical Text Analysis and Knowledge Extraction System
- Includes NLP modules for sentence boundary detection, tokenization, normalization, named entity recognition, etc.
- Developed by Georgana Savova at Children’s Hospital, Boston (previously at Mayo Clinic)
- More information at: https://community.i2b2.org/wiki/display/NLPCTAKES/Home
MetaMap

• Maps text to UMLS concepts
• Developed by Alan Aronson at National Library of Medicine
• More information at:
  http://metamap.nlm.nih.gov/
QUESTIONS?

Alexander Turchin
aturchin@partners.org