Clinical NLP schemas

Noémie Elhadad
noemie@gmail.com
Columbia University
Outline

• Types of NLP outputs
  – Unstructured text $\rightarrow$ structured output
  – Unstructured text $\rightarrow$ bag of words
  – Unstructured text $\rightarrow$ word embeddings

• The ShARE schema for structured output

• Some low level details
Primary Provider Clinic Note
Patient MRN: 0000000
Created: XXXX-XX-XX XX:XX:XX.XXXX
Pt: Bob Builder
contact info: 715-788-9999

General Medicine Clinic Note - follow up visit

HPI:
77 yo old m with h/o HTN, CAD s/p CABG 1988. Endorses intermittent dyspnea. Right eye blindness. CRI (bl 1.5-1.7). Pt has persistent gas/epigastric discomfort.

Social Hx:
lives with wife and son in the Bronx. Requires help with all ADLs. History of tobacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Father died of unknown at 80, Mother died 92.

ALL: PCN (rash)

Meds:
1) ASA 81mg po daily
3) Lisinopril 5mg po daily
4) Metformin 1000mg po bid
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable appearing NAD
PERRLA, anicteric sclerae, OP moist, no exudates
normal rate, irreg rhythm, no murmurs or gallops
+BS, soft, nt/nd EXT: WNP, no edema.

Labs:
- Na 142, k 4.8, Cl 107, CO2 23, BUN 20, Cr 1.6, Gluc 106, Ca 9.2
- hgba1c 6.9
- urinary microalbumin 2.2

A/P:
- pt 77 yo old man with HTH CAD s/p CABG 1988, Here for f/u.
-leave patient off lasix and Ace-1
- Continue Cozaar and HCTZ
-continue metformin 1000mg po bid
-will follow cr
- will refer to eye clinic
- f/u 1 month
Structured output

• Clinical NLP pipeline output

Patient should come back *if* severe *facial rash* occurs

<table>
<thead>
<tr>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>span: 35-45 (facial rash)</td>
</tr>
<tr>
<td>CUI: C0239521</td>
</tr>
<tr>
<td>body location: C0015450 (facial; 27-32)</td>
</tr>
<tr>
<td>conditional: true (if; 25-26)</td>
</tr>
<tr>
<td>negation: false (NULL)</td>
</tr>
<tr>
<td>severity: severe (severe; 28-33)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Structured output

• Useful for
  – phenotyping
  – cohort identification
  – information extraction
  – ...

Bag of words (observations)

- Vocabulary of all the words in the given notes of an institution
  - Filter out infrequent words, stop words, identifiers
  - Words are not filtered according to a terminology
- Each note is represented as a bag of words
  - Note $n = w_{43:3}, w_{118:9}, w_{210:2}, w_{534:10}, ...$
  - Lose the sequence of the words
  - Less semantic interpretation, more raw individual observations
Bag of words (observations)

- Let the burden of identifying features to further processes

<table>
<thead>
<tr>
<th>Topic 3 (heart failure)</th>
<th>Topic 32 (diabetes)</th>
<th>Topic 29 (dialysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasix</td>
<td>units</td>
<td>q15</td>
</tr>
<tr>
<td>volume</td>
<td>insulin</td>
<td>dialysis</td>
</tr>
<tr>
<td>edema</td>
<td>subcutaneous</td>
<td>fistula</td>
</tr>
<tr>
<td>heart</td>
<td>lantus</td>
<td>volume</td>
</tr>
<tr>
<td>failure</td>
<td>glucose</td>
<td>bid</td>
</tr>
<tr>
<td>worsening</td>
<td>diabetes</td>
<td>lasix</td>
</tr>
<tr>
<td>diuresis</td>
<td>times</td>
<td>placement</td>
</tr>
<tr>
<td>severe</td>
<td>70/30</td>
<td>improved</td>
</tr>
<tr>
<td>diastolic</td>
<td>diabetic</td>
<td>heparin</td>
</tr>
<tr>
<td>overload</td>
<td>days</td>
<td>examined</td>
</tr>
</tbody>
</table>

Bag of observations (words)

• Let the burden of identifying features to further processes

Word embeddings

• Observations are still words, but now a word is represented as a vector
  – Dimension reduction
  – Distributional semantics, word embeddings
• For each word, the representation is learned optimized for a particular task
  – Optimize for language model
  – Optimize for phenotype recognition
  – ...
• Often need to keep some sequential information
Outline

• Types of NLP outputs
  – Unstructured text  \rightarrow structured output
  – Unstructured text  \rightarrow bag of words
  – Unstructured text  \rightarrow word embeddings

• The ShARE schema for structured output

• Some low level details
Structured output schema

- Clinical Element Model templates
- Data modeling across several initiatives and institutions (ShARE, SHARP, THYME)

Patient should come back if **severe facial rash** occurs

Disorder

- span: 35-45 (facial rash)
- CUI: C0239521
- body location: C0015450 (facial; 27-32)
- conditional: true (if; 25-26)
- negation: false (NULL)
- severity: severe (severe; 28-33)
- ...
ShARE disorder annotations

• CUI (normalization)
  "presented with facial rash"
  Facial rash (CUI C0239521)

• Negation
  "patient denies numbness"

• Subject
  "son has schizophrenia"

• Uncertainty
  "evaluation of MI"

• Course
  "The cough got worse over the next two weeks."

• Severity
  "slight bleeding"

• Conditional
  "Pt should come back if any rash occurs"

• Generic
  "she went to the HIV clinic"

• Body Location
  "patient presented with facial rash"
  Face (CUI: C0015450)
Other semantic types

**Sign/Symptom**
- Alleviating Factor
- Associated Code
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- End Time
- Exacerbating Factor
- Generic
- Negation Indicator
- Relative Temporal Context
- Severity
- Start Time
- Subject
- Uncertainty Indicator

**Procedure**
- Associated Code
- Body Laterality
- Body Location
- Body Side
- Conditional
- Device
- End Date
- Generic
- Method
- Negation Indicator
- Relative Temporal Context
- Start Date
- Subject
- Uncertainty Indicator

**Lab**
- Abnormal Interpretation
- Associated Code
- Conditional
- Delta Flag
- Estimated flag
- Generic
- Lab Value
- Negation Indicator
- Ordinal Interpretation
- Reference Range
- Narrative
- Subject
- Uncertainty Indicator

**Disease/Disorder**
- Associated Code
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- Severity
- Start Time
- Subject
- Uncertainty Indicator
- End Time
- Exacerbating Factor
- Generic
- Negation Indicator
- Relative Temporal Context

**Medication**
- Associated Code
- Change Status
- Conditional
- Dosage
- Duration
- End Date
- Form
- Frequency
- Generic
- Negation Indicator
- Route
- Start Date
- Strength
- Subject
- Uncertainty Indicator

**Anatomical Site**
- Associated Code
- Body Laterality
- Body Site
- Conditional
- Generic
- Negation Indicator
- Subject
- Uncertainty Indicator

Slide courtesy of G. Savova
ShARe community task (disorders only)

- **Task 1** – 16 teams (concept recognition and normalization)

- **Task 2b** – 9 teams (concept + attributes normalization)

Outline

• **Types of NLP outputs**
  – Unstructured text $\rightarrow$ structured output
  – Unstructured text $\rightarrow$ bag of words
  – Unstructured text $\rightarrow$ word embeddings

• **The ShARE schema for structured output**

• **Some low level details**
ShARe dataset

- Annotated clinical notes
  - 400,000+ notes
  - 122 M words

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notes</td>
<td>298</td>
<td>133</td>
<td>100</td>
</tr>
<tr>
<td>Words</td>
<td>182K</td>
<td>153K</td>
<td>109K</td>
</tr>
</tbody>
</table>
Many concepts and attributes can be...

- Pipe delimited

```plaintext
report_name|disorder-span|cui|Norm_NI|Cue_NI|Norm_SC|Cue_SC|Norm_UI|Cue_UI|
Norm_CC|Cue_CC|Norm_SV|Cue_SV|Norm_CO|Cue_CO|Norm_GC|Cue_GC|Norm_BL|Cue_BL|
Norm_DT|Norm_TE|Cue_TE

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|*no|*NULL|*patient|*NULL|*no|*NULL|*false|
*NULL|*unmarked|*NULL|severe|*NULL|*false|*NULL|C0040132|*NULL|Before|*None|*NULL
```
Many concepts and attributes can be...

- **Pipe delimited**

| report name | disorder-span | cui   | Norm_NI | Cue_NI | Norm_SC | Cue_SC | Norm_UI | Cue_UI | Norm_CC | Cue_CC | Norm_SV | Cue_SV | Norm_CO | Cue_CO | Norm_GC | Cue_GC | Norm_BL | Cue_BL | Norm_DT | Norm_TE | Cue_TE |
|-------------|---------------|-------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|---------|--------|
| 09388-093839-DISCHARGE_SUMMARY.txt | 30-36 | C0040128 | *no    | *NULL  | *patient | *NULL | *no    | *NULL | *false | *NULL | *unmarked | *NULL | severe | *NULL | C0040132 | *NULL | Before | *None | *NULL | |

- **Composed in** [FHIR](https://fhir.org)®

Screenshot courtesy of G. Savova
Many concepts and attributes can be...

- Pipe delimited

  report_name|disorder-span|cui|Norm_NI|Cue_NI|Norm_SC|Cue_SC|Norm_UI|Cue_UI|
  |Norm_CC|Cue_CC|Norm_SV|Cue_SV|Norm_CO|Cue_CO|Norm_GC|Cue_GC|Norm_BL|Cue_BL|
  |Norm_DT|Norm_TE|Cue_TE

- Composed in FHIR®

- Lucene indexes

  Source of: file:///home/tsylthin/Work/DeepPhe/data/sample/fhir

  09388-093839-DISCHARGE_SUMMARY.txt[30-36][C0040128]|*
  |NULL|unmarked|severe|NULL|false|NULL|C0040132|

  Before|None|NULL

  Screenshot courtesy of G. Savova
Points for discussion

• NLP for what tasks and requirements on NLP output
• Tables and schema as minimum viable products given NLP technology
  – Note table vs/and NLP output table
• How to store many observations and their attributes