Extracting Events from Clinical Text Using Natural Language Processing

Shan Rajapakshe
Supervisor: Dr. Ramon Lawrence
Computer Science
Overview

• Background
  • Healthcare data
  • Information Extraction
  • Dataset

• My Work
  • Clinical Events
  • Features
  • Algorithms

• Results
  • Evaluation

• Moving Forward
Healthcare Has a Lot of Data

- Pharmacy
- GP visits
- Lab Reports
- Pathology
- Imaging
- Progress Notes
Nature of Healthcare Data

• Mix of both unstructured and structured data

• Structured data:
  • Easy to feed into a computer
  • I.e. Data in a spreadsheet

• Unstructured data:
  • Much messier
  • Harder to represent and ‘understand’
  • Images, sounds, video, natural language text
My Goal – Solve a Subproblem

• Extract relevant events from unstructured clinical text

Why?

• Currently Physicians need to look back on data
• Less time reading charts, more time with patients!
What is Event Extraction?

• “...Extraction of complex combinations of relations between actors (entities), performed after executing a series of Natural Language Processing steps...”

• Ie. Finding a relevant event within text data

Difficult because of ambiguity
Methods of Information Extraction

- **Expert Systems**
  - Leverage pre-existing knowledge
  - Often use patterns or rules
  - Limited by scope of knowledge

- **Data Driven**
  - Use features from text
  - Apply statistical methods and Machine learning
  - Limited by the data
Dataset

- Clinical Progress notes from a set of 262 Lung Cancer Patients from the BC Cancer Agency
- Had to be anonymized (no patient identifiable information left, but able to access the specific patient if necessary)
  - ~ 10 Charts per patient
  - ~ 2700 Charts total
  - ~ 56000 sentences
My Goal – Revisited

- Extract relevant events from unstructured clinical text
My Goal – Revisited

- Extract relevant events from unstructured clinical text using a data driven approach
My Goal – Revisited

• Extract relevant **events** from unstructured **clinical text** using a **data driven approach**
  • Determine relevant events
  • Prepare data
  • Find appropriate features
  • Choose suitable machine learning methods
  • Apply features to methods
  • Evaluate results
Tools

As well as NLTK, Numpy, Faker, and others!
Clinically Relevant Events

Had to define what were important events

• Treatments
  • Chemotherapy
  • Radiation
  • Surgery
  • Palliative

• Recurrence
Data Preparation

- Pull data from CAIS (at BCCA)
- Convert PDFs to text
- Anonymize and obfuscate patient information

<table>
<thead>
<tr>
<th>Recurrence</th>
<th>Chemotherapy</th>
<th>Radiation</th>
<th>Surgery</th>
<th>Palliative</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Features

• How to let a computer represent textual data?

• Balance between understandability for humans vs ease of use for machine
Features Used

Bag of Words representations
- 1-grams
- 2-grams
- TFDIF

Alice saw Bob
Susan called Bob

Parts of Speech Tags
- Dependency Parse Trees
- Named Entities

Alice
Bob
Saw
Susan
Called

ROOT
NSUBJ
DOBJ
Alice
saw
Bob
NOUN
VERB
NOUN
Algorithms

Logistic Regression
- Robust
- Quick
- Powerful

Multilayer Perceptron
- Slow
- Requires lots of data
- Lots of potential
Pipeline

Chart

Text

Sentences

Categories of each Sentence

Bag of Words

Words

Doc2Vec/Word2Vec

POS Tags

Dependency Parse Trees

Named Entities
Evaluation

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

\[
\text{Recall} = \frac{tp}{tp + fn}
\]

\[
F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]
Preliminary Results

- Using 3-Gram BOWs for Recurrence Events

--- LOGISTIC REGRESSION ---
Normalized Score: 0.9860
Raw Score: 2329

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>26</td>
</tr>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2308</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.54</td>
<td>0.57</td>
<td>28</td>
</tr>
<tr>
<td>Avg / Total</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2362</td>
</tr>
</tbody>
</table>

--- PERCEPTRON ---
Normalized Score: 0.9814
Raw Score: 2318

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.46</td>
<td>0.73</td>
<td>0.57</td>
<td>26</td>
</tr>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2308</td>
</tr>
<tr>
<td>1</td>
<td>0.43</td>
<td>0.11</td>
<td>0.17</td>
<td>28</td>
</tr>
<tr>
<td>Avg / Total</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2362</td>
</tr>
</tbody>
</table>
Preliminary Results

- Using 3-Gram BOWs for Recurrence Events

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2308</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.54</td>
<td>0.57</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td><strong>Avg / Total</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>2362</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.46</td>
<td>0.73</td>
<td>0.57</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2308</td>
<td></td>
</tr>
<tr>
<td>0.43</td>
<td>0.11</td>
<td>0.17</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td><strong>Avg / Total</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>2362</strong></td>
<td></td>
</tr>
</tbody>
</table>
What does this mean?

- Majority of sentences **not relevant to these events**
- Potential to **streamline physician workflow** by only showing relevant information

----- LOGISTIC REGRESSION -----

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>26</td>
</tr>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2308</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.54</td>
<td>0.57</td>
<td>28</td>
</tr>
<tr>
<td>Avg / Total</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>2362</td>
</tr>
</tbody>
</table>
Conclusions

• Open Source NLP tools can be used to extract clinical events from healthcare data

• Simple algorithms work well for small datasets

• Methods work better than a naïve classifier

• Data cleaning is hard
Going Forward

- Already sharing data with group in Victoria
- Improve labelling/data generation process
- Find a meaningful way to represent information to physicians
Thanks!

• To Dr. Ramon Lawrence for providing guidance and asking me tough questions
• To Dr. Jonn Wu for the space to work at the BCCA and the idea
• To Dr. Cheryl Ho for the patient cohort

And thank you!
Questions?