Natural Language Processing for Clinical Informatics and Translational Research Informatics

Imre Solti, M. D., Ph. D.
solti@uw.edu
K99 Fellow in Biomedical Informatics
University of Washington
Background for Clinical Natural Language Processing (NLP)

- NLP
- Comp Ling MA
- NLM Postdoctoral Fellowship in BMI
- 8 yrs Healthcare IT
- Ph. D. in Health Services Organization and Research
- Medical Doctoral Degree
Career Interest

NLP as strategic tool to achieve the six aims of the Institute of Medicine

Health Care

Patient-Centered

Effective

Timely

Efficient

Safe

Equitable
Research Interests

NLP for Clinical Informatics and Translational Research Informatics

- Individualized Medicine
- Personal Health Records
- Patient Portals

- Comparative-Effectiveness

- Patient-Centered
- Effective

- Timely
- Efficient

- Health Care

- Safe
- Patient Safety
- Adverse Drug Events
- Pharmacovigilance

- Equitable
- Discrimination based on Age or Ethnicity

- Translational Research
- Clinical Trial Eligibility Screening

- Cost of Care
- Present On Admission
Research Interests - Summary

• Information Extraction from unstructured clinical text -> Linking phenotype and genotype
• Document Classification
• Data Mining
Use Cases for Today’s Presentation

NLP Research Use Cases for the Electronic Medical Record

1. Semi-Automated Medical Problem List
2. Extraction of Medication Information
3. Automated Classification of Radiology Reports for Acute Lung Injury
4. Automated Clinical Trial Eligibility Screening
Collaborators

• University of Washington
  – Eithon Cadag, Ph. D. – Biomedical Informatics
  – John Gennari, Ph. D. - Biomedical Informatics
  – Scott Halgrim, M. A. – Computational Linguistics
  – Tom Payne, M. D. – IT Services Medical Center
  – Peter Tarczy-Hornoch, M. D. - Biomedical Informatics
  – Mark Wurfel, M. D. – Pulmonary and Critical Care Med
  – Fei Xia, Ph. D. – Computational Linguistics

• External Investigators
  – University of Pittsburgh
  – Columbia
  – Albany/MIT, i2b2 (Informatics for Integrating Biology and the Bedside)
Definitions¹

• **Natural Language Processing (NLP):**
  NLP research focuses on building computational models for understanding natural (human) language.

• **Information Extraction (IE):**
  IE involves extracting predefined types of information from text. Subfield of NLP.

• **Named Entity Recognition (NER):**
  Recognizing expressions denoting entities (i.e., Named Entities), such as diseases in free text documents. Subfield of IE.

• **Information Retrieval (IR):**
  Information retrieval (IR) is focused on finding documents.

Definitions\(^1\) – Cont.

• **Document Classification:**
  Assigning electronic documents to one or more categories.

• **Biomedical Text:**
  Text that appears in books, articles, literature abstracts.

• **Clinical Text:**
  Texts written by clinicians in the clinical setting.

• **Biomedical-NLP:**
  NLP for biomedical text.

• **Clinical-NLP:**
  NLP for the clinical text.

Agenda for Today

Past Projects:
1. Semi-Automated Medical Problem List: Clinical-NLP, IE, NER - 1 Slide
2. Extraction of Medication Information: Clinical-NLP, IE, NER - 1 Slide
3. Classification of Radiology Reports for Acute Lung Injury: Clinical Document Classification

Future Project:
4. *Automated Clinical Trial Eligibility Screening: Clinical NLP, Biomedical-NLP, IE, NER, Document Classification
   *Grant funded
Clinic Note: BPH. Congestive heart failure. Some of this is related to his tachy-brady syndrome. He has no nausea, vomiting, diarrhea.

Problem List:
1. Benign Prostatic Hypertrophy
2. Congestive Heart Failure
3. Tachy-Brady Syndrome

Automated Extraction of Medication Information

**DISCHARGE MEDICATIONS:**

Additionally, *Aspirin* 1-2 tablets p.o. q 4 prn during next ten days for his pain, ...

**IE**
- Medication: “Aspirin”
- Dose: “1-2 tablets”
- Mode of Administration: “p. o.”
- Frequency: “q 4 prn”
- Duration: “ten days”
- Medical Reason: “pain”

**MaxEnt Algorithm**

**ML Features**

**Hybrid IE System**

**Handcrafted Rules**

**RxNorm Lexicon**

**UW – i2b2 Collaboration**

**i2b2 Challenge**

**Building System**

**Challenge**

**Community Annotation**

**MA Thesis**
Classification of Radiology Reports for Acute Lung Injury (ALI)

Motivation

• 30 % Mortality
• Delayed manual chest x-ray classification

Aims

• Build NLP-based classifier
• Intuitive link: Machine Learning – Clinical Expertise

Methods

• Keywords
• Maximum Entropy: Character n-grams
Data (Corpus) and Gold Standard

953 chest x-ray reports

96: Clean
857: Noisy

MDs
Manual Classification

ALI +
ALI -
Task for Automated ALI Classification

Chest x-ray Reports

System

Automated Classification

ALI +

ALI -
Sample Report

Tubes and lines: satisfactory position and alignment

Lungs: The lung volumes are low and unchanged. There are diffuse, bilateral opacities that are worsened.

Pneumothorax: none

Effusions: none
Measurement Metrics

Recall, Precision, F-measure, Accuracy

Recall = TP / (TP + FN)
Precision = TP / (TP + FP)
Accuracy = (TP + TN) / (TP + TN + FP + FN)
F-measure = 2PR / (P + R)
Baseline

No Processing: Assign ALI+ to Every Report

System Classified ALI+

Larger Set

\[ R = \frac{TP}{TP + FN} = \frac{392}{392} = 1 \]

\[ P = \frac{TP}{TP + FP} = \frac{392}{857} = 0.46 \]

F-measure = \[ \frac{2PR}{P + R} = 0.63 \]
## Gold Standard – Smaller Corpus

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>R</th>
<th>P</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.94</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
<td>0.98</td>
<td>0.88</td>
</tr>
<tr>
<td>5</td>
<td>0.62</td>
<td>1.00</td>
<td>0.77</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>0.83</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>8</td>
<td>0.70</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>0.70</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>10</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>11</td>
<td>0.92</td>
<td>0.98</td>
<td>0.95</td>
</tr>
</tbody>
</table>
## List of Keywords (Sample)

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Weight/3</th>
<th>Weight/10</th>
</tr>
</thead>
<tbody>
<tr>
<td>edema</td>
<td>2.5</td>
<td>8</td>
</tr>
<tr>
<td>lung opacities</td>
<td>2</td>
<td>5.5</td>
</tr>
<tr>
<td>diffuse</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>bilateral</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

48 Key Phrases
## Keyword & Weight-Based Results

<table>
<thead>
<tr>
<th>Method</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>96-raw</td>
<td>0.88</td>
<td>0.83</td>
<td>0.85</td>
<td>0.844</td>
</tr>
<tr>
<td>96-w3</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.833</td>
</tr>
<tr>
<td>96-w10</td>
<td>0.72</td>
<td>0.88</td>
<td>0.80</td>
<td>0.800</td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>0.46</td>
<td>0.63</td>
<td>0.46</td>
</tr>
</tbody>
</table>
MaxEnt Character n-gram Features

- Unigram, Bigram, ... 6-gram
- “diffuse”, 6-gram, sliding window

nnnn_d
nnn_di
nn_dif
n_diff
_diffu
diffus
iffuse
ffuse_
fuse_n
etc ...
## MaxEnt Results (Smaller Corpus)

<table>
<thead>
<tr>
<th>System</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>0.83</td>
<td>0.78</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>n1</td>
<td>0.62</td>
<td>0.58</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td>n2</td>
<td>0.67</td>
<td>0.81</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>n3</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>n4</td>
<td>0.85</td>
<td>0.97</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>n5</td>
<td>0.77</td>
<td>0.73</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>n6</td>
<td>0.84</td>
<td>0.82</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>0.46</td>
<td>0.63</td>
<td>0.46</td>
</tr>
</tbody>
</table>
# MaxEnt vs Keyword

<table>
<thead>
<tr>
<th>System</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.88</td>
<td>0.83</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>W3</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>n3</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>n4</td>
<td>0.85</td>
<td>0.97</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>n6</td>
<td>0.84</td>
<td>0.82</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>0.46</td>
<td>0.63</td>
<td>0.46</td>
</tr>
</tbody>
</table>

ROC statistics - Not significant difference Keyword vs MaxEnt
<table>
<thead>
<tr>
<th>N-gram Feature</th>
<th>Clinical Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>edema_</td>
<td>edema</td>
</tr>
<tr>
<td>a_and_</td>
<td>edema and</td>
</tr>
<tr>
<td>ffuse_</td>
<td>diffuse</td>
</tr>
<tr>
<td>teral_</td>
<td>bilateral</td>
</tr>
<tr>
<td>y_opac</td>
<td>patchy opacities</td>
</tr>
<tr>
<td>al_opa</td>
<td>bilateral opacities</td>
</tr>
</tbody>
</table>

**Missing from 48-Phrase List**

| perihi         | perihilar                     |
Limitations

1. Two corpora (Selection and GS Criteria)
2. Not tested – Other ALI Research Team Corpora
3. Features limited to n-grams
4. Different performance peaks (96 vs 857-set)
Related Work ALI Classification

- Herasevich et al., - Mayo Clinic, Rochester (2009)
- Azzam et al., - UPenn (2009)
- Rule-based systems, focus -> ALI screening not on NLP component
- No details -> Not directly comparable
Conclusions for ALI Classification

1. Aims achieved:
   I. Built NLP-based classifier(s)
   II. Visualized ML features for clinicians

2. Advantages and disadvantages: Keyword and ML-based systems ->

3. What approach is better?
Use Cases for Today’s Presentation

NLP Research Use Cases for the Electronic Medical Record

1. Semi-Automated Medical Problem List
2. Extraction of Medication Information
3. Automated Classification of Radiology Reports for Acute Lung Injury
4. Automated Clinical Trial Eligibility Screening
Automated Clinical Trial Eligibility Screening - Task

200,000/year

Notes

700,000/year

Eligibility Screening

10,000 trials/year in US
$ 30 billion/year*

84,000 trial announcement

http://clinicaltrials.gov
178 OHSU studies

Automated Patient-Centered Clinical Trial Eligibility Screening

Background and Significance:

- **Low Rate:** 4% adult cancer patients
- **Physician Bias:** older age, minority status
- **Not Mentioned:** 25% b cc surg -> 0 offer, 40% -> 1-10% offer

Aims:

- Identify concept elements
- Build inf application to extract and match
- Interactive input module
- Evaluation of performance
Related Work

• Protocol Authoring Tools
• Standardized Terminology
  – Clinical Data Interchange Standards Consortium
  – Biomedical Research Integrated Domain Group
  – HL7
  – Trial Bank/Open Trial Bank – Ida Sim
  – Columbia – Patel and Weng
• Cincinnati - Embi
• Others...
Excerpts – Trial Announcement

• First-degree relative with bilateral breast cancer who developed the first breast cancer at \( \leq 50 \) years of age

• Postmenopausal, defined as at least 1 of the following:
  • Over 60 years of age
  • Bilateral oophorectomy
  • \( \leq 60 \) years of age with a uterus and amenorrhea for at least 12 months

• No cancer within the past 5 years except nonmelanoma skin cancer or carcinoma in situ of the cervix
Points of Intervention for NLP Systems

- **Patient Portal**
- **EMR1**
- **EMR2**
- **ATS**
- **IE**
- **PHR NLP-Module**
- **Eligibility Criteria**
- **Trial Announcements**
- **Automated Text Summarization**
- **IE**
- **IE**
- **IE**
- **Diseases**
- **Medications**
- **Labs, Vitals**
- **Family History**
- **Personalized Recommendations**

Patient-Centered Clinical Trial Eligibility Screening
NLP Use Cases for Clinical Informatics and Translational Informatics

Task for Automated ALI Classification

Points of Intervention for NLP Systems

Summary – Questions?
Summary – Questions? (Text Version)

Past Projects:
1. Semi-Automated Medical Problem List: Clinical-NLP, IE, NER - 1 Slide
2. Extraction of Medication Information: Clinical-NLP, IE, NER - 1 Slide
3. Classification of Radiology Reports for Acute Lung Injury: Clinical Document Classification

Future Project:
4. *Automated Clinical Trial Eligibility Screening: Clinical NLP, Biomedical-NLP, IE, NER, Document Classification

*Grant funded