A Novel Deep Learning System (DI++) for Patient Disease Extraction in Clinical Notes

Jinhe Shi, Yi Chen, Chenyu Ha, William C. Kinsman
New Jersey Institute of Technology
University of Maryland College Park
Inovalon Inc.
Motivation

- Patient disease information is not fully captured in structured EHR
- Accurate recording patient conditions is critical for
  - Risk prediction
  - Supporting clinic decision making
  - Ensuring correct billing

**Problem Definition:** Given clinical notes, extract the conditions of a patient.
I did talk to the patient about surveillance of the meningioma.
Disease Mention Classification: Challenges

It’s hard to judge whether or not a disease mention is a patient condition. It’s context sensitive!

- “I did talk to the family about surveillance of the meningioma”
- “his family history includes cancer in his father”
- “Flag Reference GFR estimates are unreliable in patients with severe malnutrition or obesity, rapidly changing kidney function, loss of limbs or abnormal muscle mass....”
Leveraging AI For NLP

- Breakthrough in AI: Deep Learning
- Two successful deep learning models
  - Convolutional Neural Network (CNN): Good at feature extraction, but ignore the word orders.
  - Long short-term memory (LSTM): Can learn sequence dependency, but lack the ability of feature extraction.

Can we get the best of both worlds?
CLSTM-Attention Architecture For Disease Mention Classification

- CNN: Sentence feature extraction
- LSTM: Mention feature dependency learning
- Attention: Focus on important words
Sentence Feature Representation

- Input: Mention sentence (3-sentences)
- Output: A set of feature vectors
- Model: CNN
Mention Classification

– Input: Features Vectors of Mention Context
– Output: whether the mention is Positive/Negative
– Model: Bi-LSTM, Attention
Evaluation

- Data Set: Inovalon’s MORE2 Registry Dataset
  - 11,943 clinical charts, 1.06 million pages

- Ground Truth:
  - Coder Team in Inovalon reads each page of charts and highlight the positive diseases

- Diseases:
  - We focus on identifying diseases categorized by Hierarchical Condition Category (HCC) coding
Evaluation Methods

- **Mention Extraction:**
  - To test if DI++ can extract all positive disease mentions

- **Mention Classification:**
  - Given an identified disease mention, test if CLSTM-Attention model can correctly classify it as positive or negative
Mention Extraction

Comparison Systems

- **cTAKES[5]:** Apache cTAKES is an open-source natural language processing system that extracts clinical information, including disease mention, from clinic notes.
- **Inovalon System:** Support Vector Machine approach

<table>
<thead>
<tr>
<th>Systems</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>cTAKES</td>
<td>0.07</td>
<td>0.89</td>
<td>0.13</td>
</tr>
<tr>
<td>Inovalon System</td>
<td>0.54</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td>DI++</td>
<td>0.78</td>
<td>0.59</td>
<td>0.67</td>
</tr>
</tbody>
</table>
### Mention Classification

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>83.01%</td>
<td>0.81</td>
</tr>
<tr>
<td>LSTM</td>
<td>82.36%</td>
<td>0.79</td>
</tr>
<tr>
<td>Hierarchical Attention</td>
<td>84.38</td>
<td>0.82</td>
</tr>
<tr>
<td>Transformer</td>
<td>85.76%</td>
<td>0.82</td>
</tr>
<tr>
<td>CLSTM-Attention</td>
<td>86.52%</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Case Study

CONDITIONS DETECTED

- Congestive Heart Failure [HCC85]
- Specified Heart Arrhythmias [HCC96]
- Chronic Obstructive Pulmonary Disease [HCC111]
- Angina Pectoris [HCC88]
- Rheumatoid Arthritis, Inflammatory Tissue Disease [HCC40]
- Diabetes with Chronic Complications [HCC18]
- Diabetes without Complication [HCC19]
- Artificial Openings [HCC188]
- Morbid Obesity [HCC22]
- Vascular Disease [HCC108]
- Schizophrenia [HCC57]
- Major Depressive, Bipolar, Paranoid Disorders [HCC58]
- Drug/Alcohol Dependence [HCC55]
- Disorders of Immunity [HCC47]
- Bone/Joint/Muscle Infections/Necrosis [HCC39]
- Other Significant Endocrine, Metabolic Disorders [HCC23]
- Metastatic Cancer and Acute Leukemia [HCC8]
- Cardio-Respiratory Failure and Shock [HCC84]
- Coagulation Defects, Hematological Disorders [HCC48]
- Pneumococcal Pneumonia, Empyema, Lung Abscess [HCC115]
- Aspiration, Specified Bacterial Pneumonias [HCC114]
- Lung and Other Severe Cancers [HCC9]
- Unstable Angina, Acute Ischemic Heart Disease [HCC87]

FOCUS

- Targeted HCC – HCC 85
- Field HCC – HCC 18

Discharge Date: 07/21/2014

Please kindly refer to the detailed discharge summary dictated on 07/18/2014.

PRIMARY DISCHARGE DIAGNOSIS:
1. Second-degree heart block. This has resolved and the patient has been restarted on Coreg 3.125 mg b.i.d. It should be noted that the patient is no longer on 25 mg b.i.d. of Coreg.
2. Status post left hemiarthroplasty.

SECONDARY DISCHARGE DIAGNOSIS:
1. Chronic diastolic congestive heart failure.
2. Morbid obesity.
3. Obstructive sleep apnea.
4. Chronic anemia.
5. Chronic kidney disease stage 5.