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

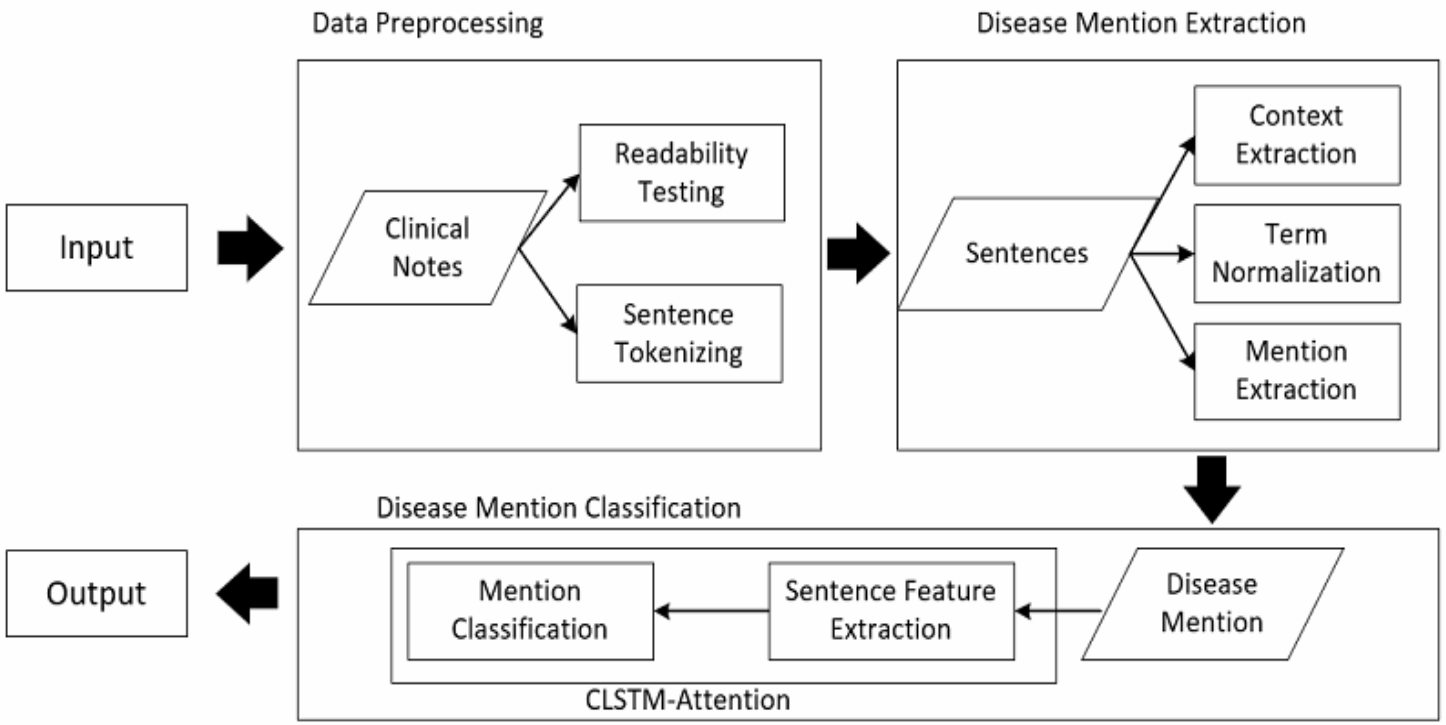
Inovalon Inc.

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.

DI++ System Architecture

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.

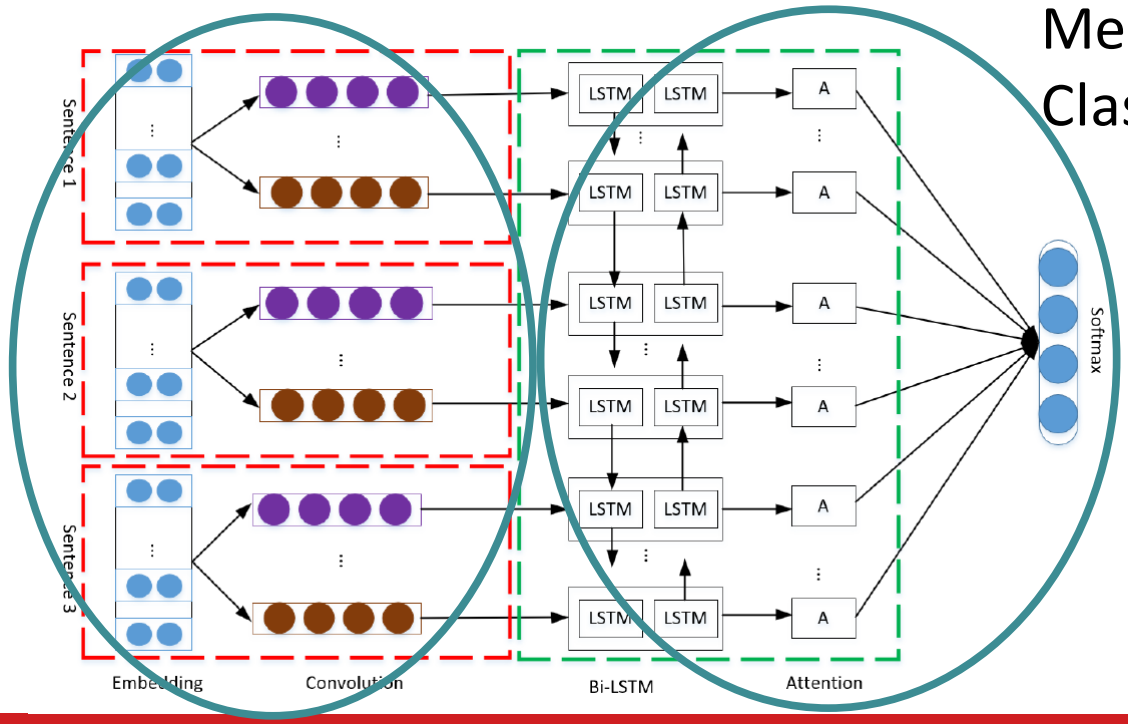
Can we get the best of both worlds?

CLSTM-Attention Architecture For Disease Mention Classification

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

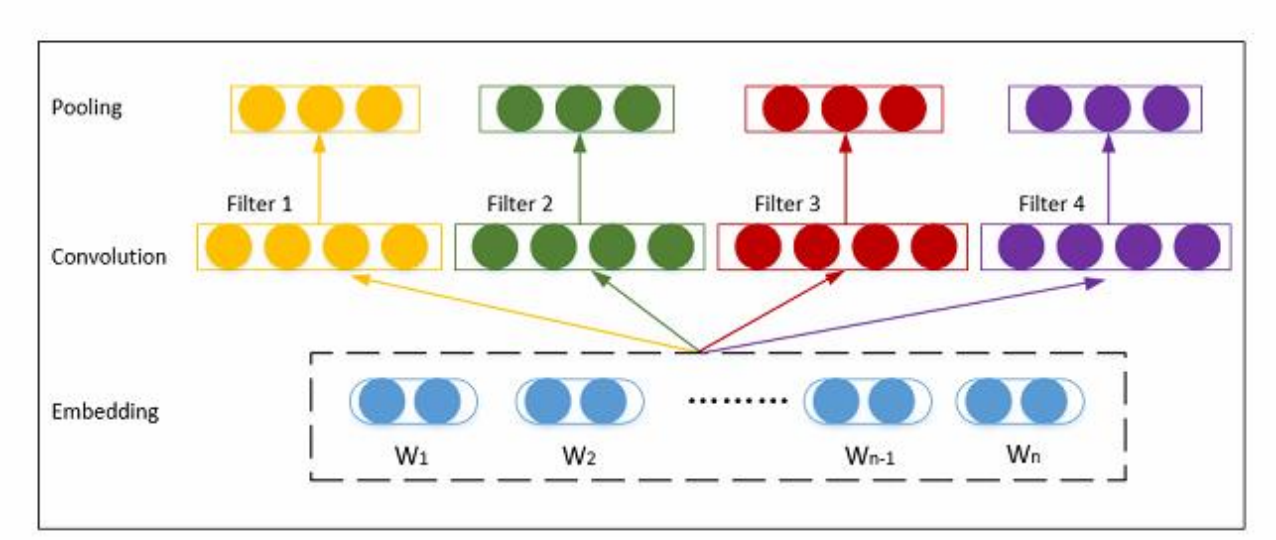
Sentence Feature Representation

Mention Classification



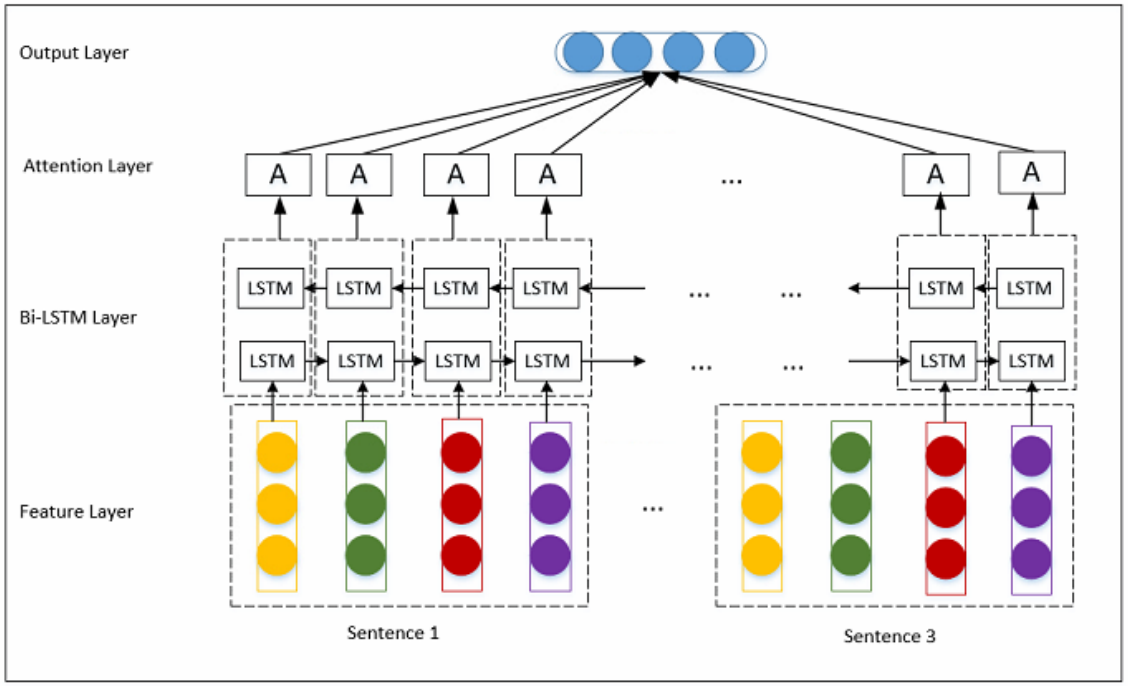
Sentence Feature Representation

- 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

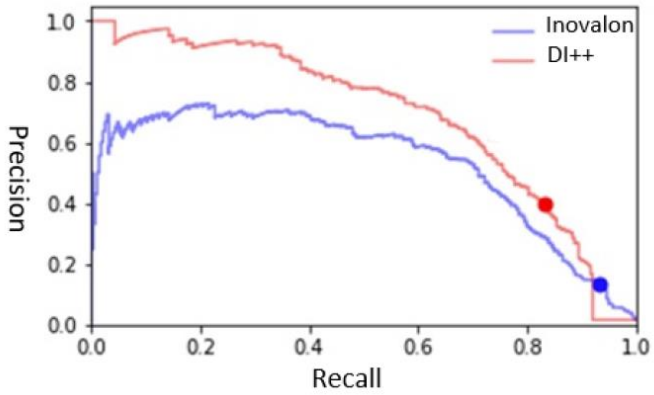
Evaluation Methods

- 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, include disease mention, from clinic notes.
  - Inovalon System: Support Vector Machine approach

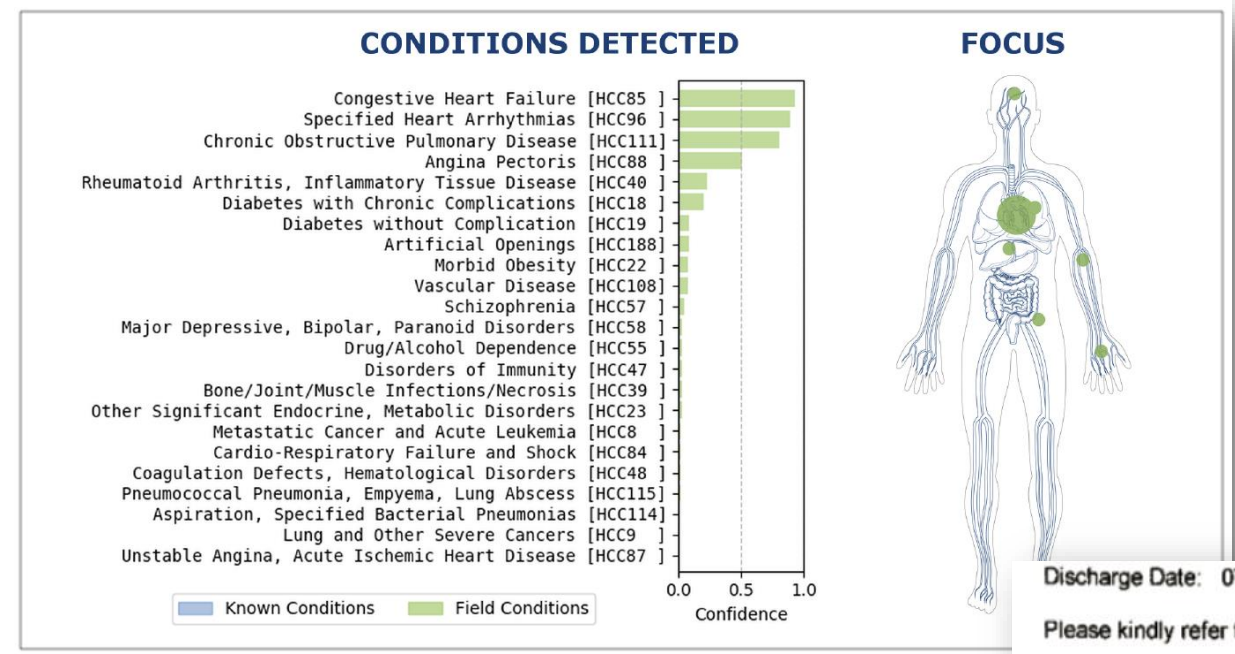
Systems	Precision	Recall	F1 Score
cTAKES	0.07	**0.89**	0.13
Inovalon System	0.54	0.65	0.60
DI++	**0.78**	0.59	**0.67**



Mention Classification

Approaches	Accuracy	AUC
CNN	83.01%	0.81
LSTM	82.36%	0.79
Hierarchical Attention	84.38	0.82
Transformer	85.76%	0.82
CLSTM-Attention	**86.52%**	**0.84**

Case Study



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.

Targeted HCC – HCC 85

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

Field HCC – HCC 18