Deep Learning for Medical Event Detection in EHRs

Hong Yu, Abhyuday Jagannatha, and the UMass BioNLP group

University of Massachusetts Medical School
University of Massachusetts-Amherst
Bedford VAMC

In proceedings of NAACL-HLT 2016, *long paper*, San Diego, USA

Objective

- Extraction of Medical Events and their attributes from Electronic Health Records and identification of their relations
- Categories of medical entities:

Medical Events

1. Indication
2. Adverse Drug Events (ADE)
3. Other Sign Symptom or Diseases(Other SSSD)
4. Medication (Drugname)

Attributes

1. Severity
2. Route
3. Dosage
4. Duration
5. Frequency

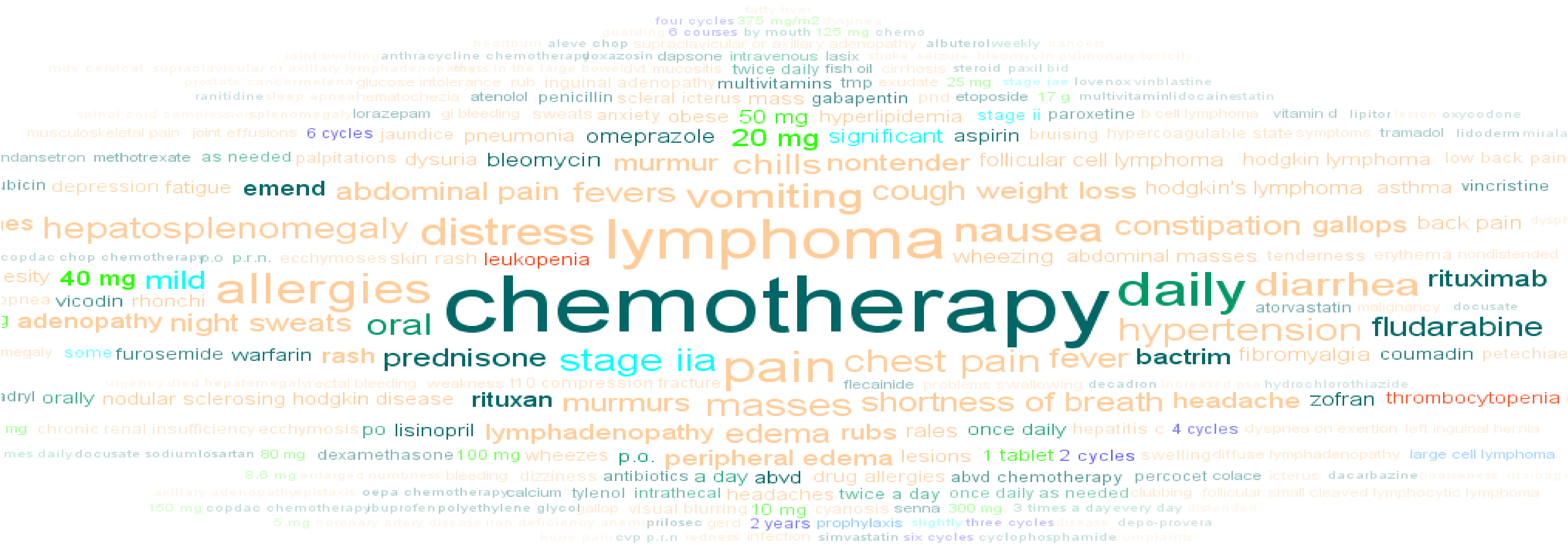
Sample Sentence from dataset

"The follow-up needle biopsy results were consistent with *bronchiolitis obliterans*, which was likely due to the Bleomycin component of his *ABVD chemo*."

Expected Output

ADE: *bronchiolitis obliterans* **Drugname:** *ABVD chemo*

Annotating Cancer EHRs



Challenges

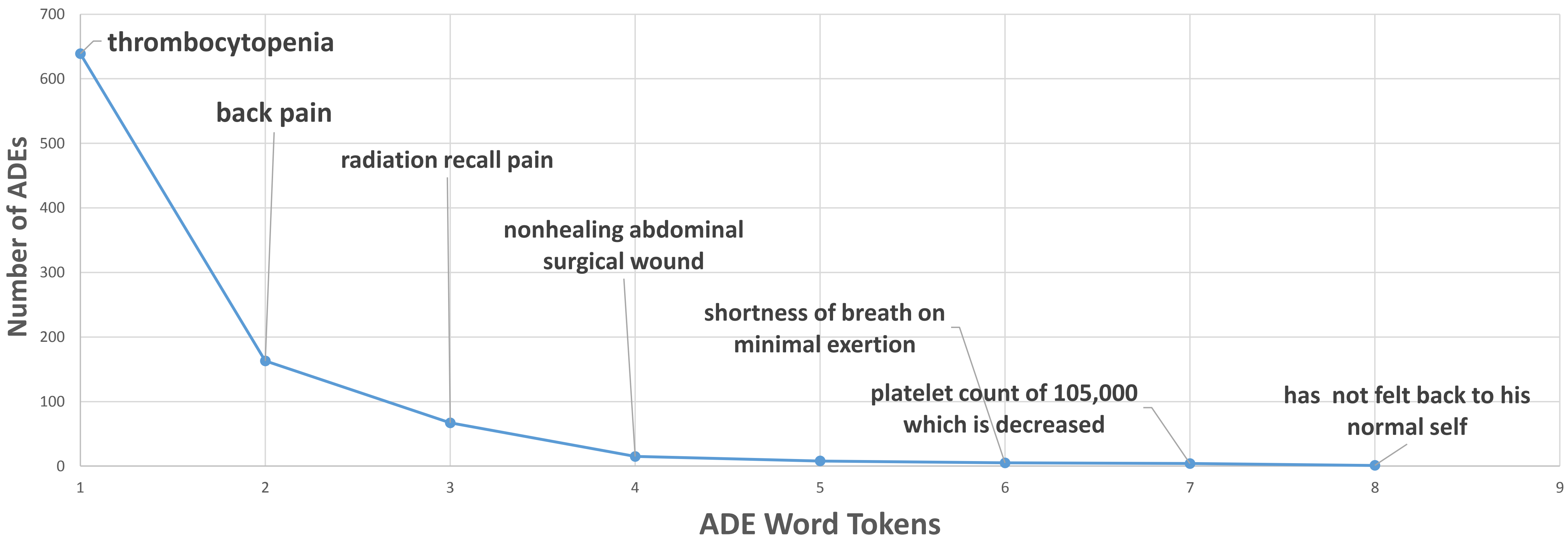
- Common vocabulary between different medical entities. E.g.

Example Sentence from Dataset

Anemia and *Thrombocytopenia*: Anemia likely multifactorial d/t CTCL, chronic disease, and romidepsin; *thrombocytopenia* likely d/t romidepsin.-Discussed with Dr. [**Last Name (STitle) 33**]. *Thrombocytopenia* will not likely worsen much further. [**Doctor First Name **] to continue coumadin for now.

  - In our corpus, an Adverse Drug Event is tagged only when there is a direct evidence in the text , linking it as a side effect of a medication.
  - The second mention of *thrombocytopenia* is an Adverse Drug Event.
  - The first and third mentions of *Thrombocytopenia* are not used to describe a side effect directly, so they are labeled as *Other SSD* .

Challenges



700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
Number of ADEs
ADE Word Tokens
thrombocytopenia
back pain
radiation recall pain
nonhealing abdominal surgical wound
shortness of breath on minimal exertion
platelet count of 105,000 which is decreased
has not felt back to his normal self

Challenges

- Unbalanced data and severe label sparsity.

Labels	Annotations	Avg. Annotation Length
ADE	905	1.51
Indication	1988	2.34
Other SSD	26013	2.14
Severity	1928	1.38
Drugname	9917	1.20
Duration	562	2.17
Dosage	3284	2.14
Route	1810	1.14
Frequency	2801	2.35

Data statistics for the version used in [Jagannatha and Yu, NAACL HLT, 2016]

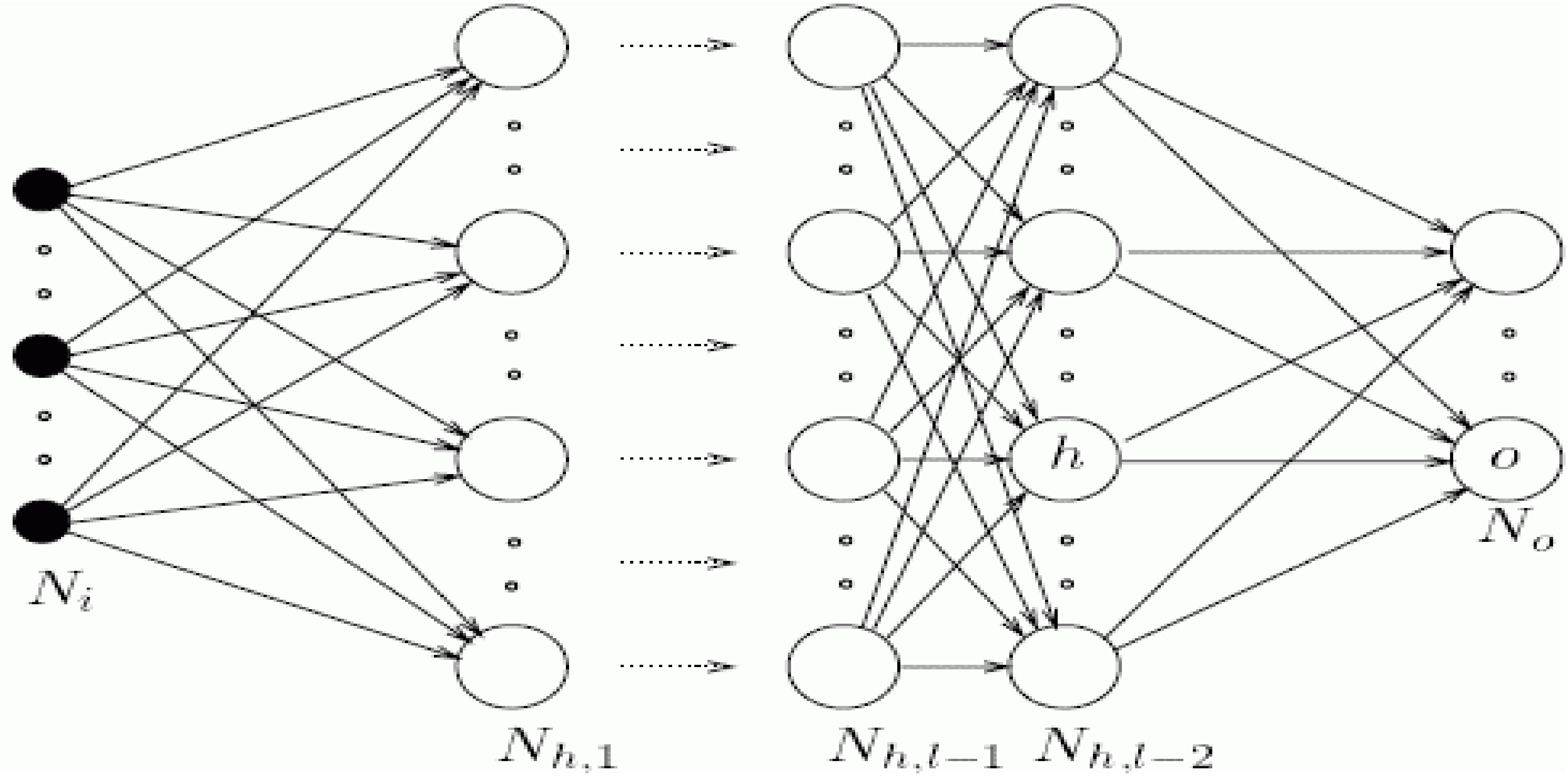
Challenges

- Noisy or incomplete sentence structure
  - Sentence structures in EHRs are unique to medical domain. Models trained using Standard NLP corpus (Treebank) do not adapt well to EHRs.
- Inconsistent syntactic cues
  - SSD mention can have one of many syntactic forms like noun, noun phrase, verb phrase, etc.
- Abbreviations and medical jargon
  - Multiple word forms exist for the same disease name.
- Multi-word entities
  - Several medical entities have variable lengths ranging from a single word to multi-word phrases (sometimes even more than ten words).

Methodology

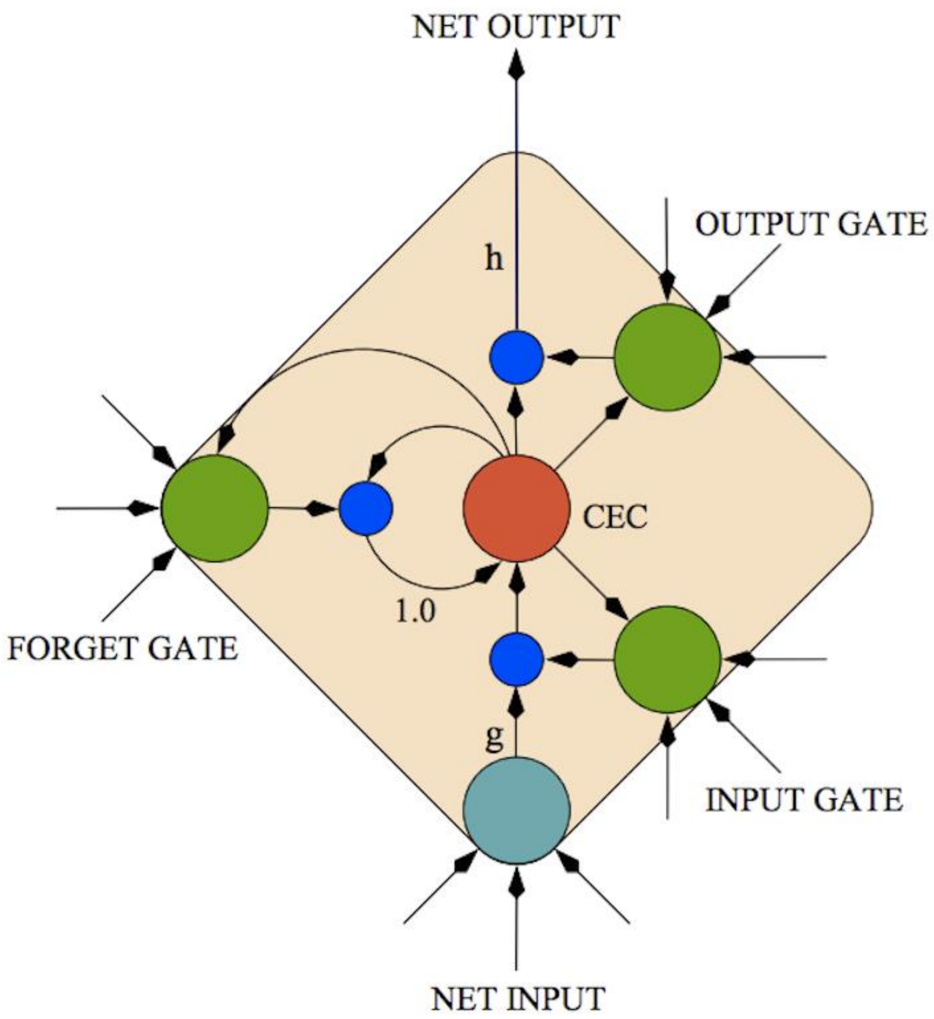
- We treat extraction of Medical Entities as a Sequence Labeling problem.
- Sequence Labeling in Natural Language Processing (NLP) treats a text document as a sequence of word and special character tokens.
- The algorithm learns to label sub-sequences which correspond to a particular category.
- <u>State of the Art</u>**: Conditional Random Fields**
- <u>Our Approach</u>**: Deep Learning**

Our approach: Deep Learning



Recurrent Neural Networks (RNN)

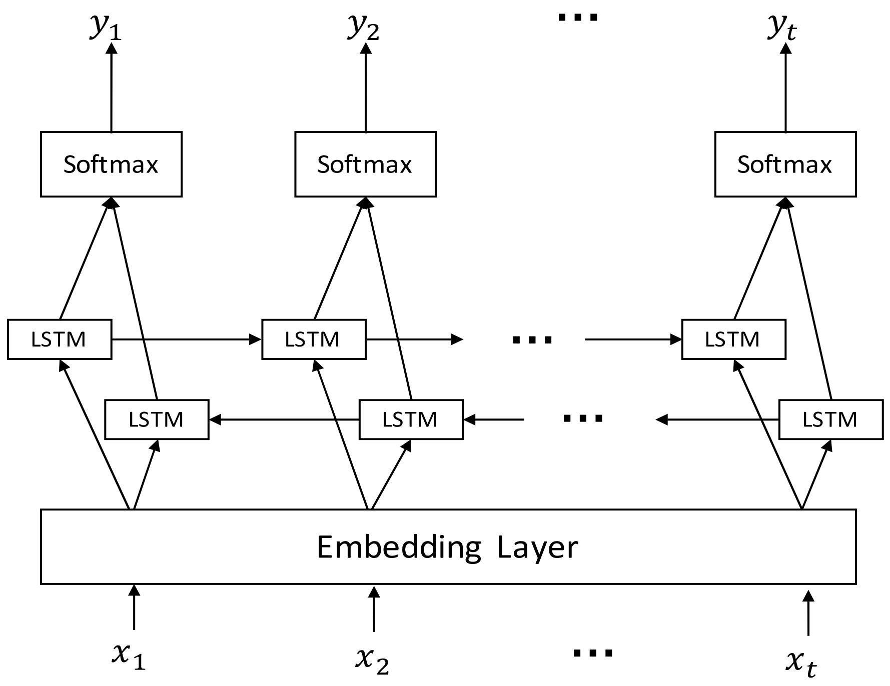
- RNNs are artificial neural networks, with recurrent connections.
- This recurrent connection helps them remember previous context history when processing a sequence of tokens.
- The figure shows the internal structure of one of the RNN models we use, Long Short Term Memory (LSTM).
- RNN models do not need handcrafted features. They are able to recognize patterns in variable context lengths using just word inputs.



Graves et. al. 2009

Bidirectional RNNs

- The figure shows a Bidirectional RNN layer using LSTM neurons.
- (x_1,x_2 … x_t) are input words.
- (y_1,y_2 … y_t) are corresponding output labels.
- The network has two recurrent chains, going in either directions.
- Bidirectional chains enables the model to use the context from both directions to predict the output label.
- The RNN models are trained at both sentence and document levels.



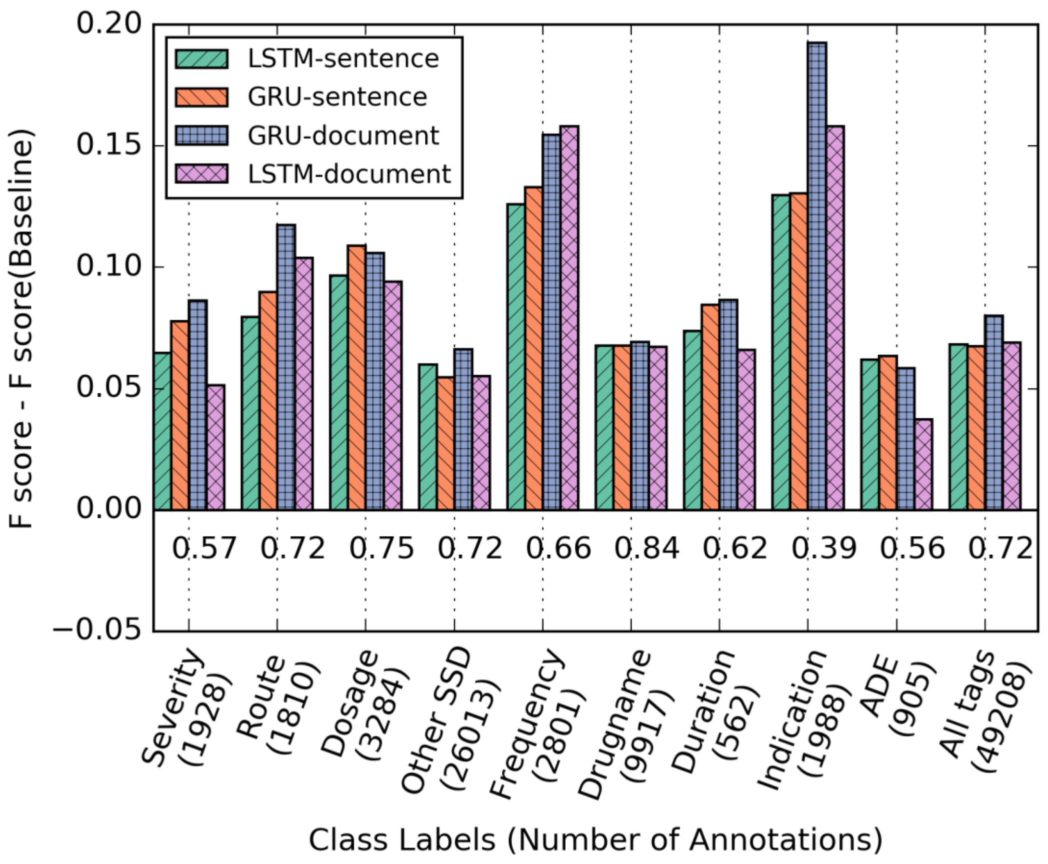
Sequence Labeling Model for LSTM Network.

Results

Models	Recall	Precision	F-score
CRF-nocontext	0.6562	0.7330	0.6925
CRF-context	0.6806	0.7711	0.7230
LSTM-sentence	0.8024	0.7803	0.7912
GRU-sentence	0.8013	0.7802	0.7906
LSTM-document	0.8050	0.7796	0.7921
GRU-document	0.8126	0.7938	0.8031

Strict Evaluation results for micro averaged Recall, Precision and F-score. All results use ten fold cross validation.

Results



- Figure shows change in F-score for RNN models with respect to CRF-context (baseline). The values below the plotted bars represent the baseline f-scores for each class label.
- Using the entire document is very helpful in extracting *Route, Frequency*, and *Indication*.
- Even though Document models improve the performance overall, using isolated sentences for prediction seems to be better for detecting important labels like *ADE* and *Other SSD*.

Result

CRF-nocontext

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	19	30	0	1.4	0	0.3	0.4	49	0.1	0
Outside	0	98	0.1	0.1	0	0.1	0	1.2	0	0.1
Dosage	0	13	85	0.5	0.1	0.9	0	0.1	0.3	0
Drugname	0.1	11	1.2	86	0.2	0.7	0.1	0.8	0.3	0
Duration	0	50	1	0.1	45	2.7	0	0.3	0.1	0
Frequency	0	31	0.8	0.8	0.3	66	0.1	0	0.3	0
Indication	0.1	17	0	1.3	0	0.3	18	62	0	0.7
Other SSD	0.3	26	0	0.1	0	0	1.3	73	0	0.2
Route	0	17	2.4	2.8	0	0.6	0	0.6	77	0
Severity	0	41	0	0	0	0	0.6	8.1	0	50

CRF-context

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	39	29	0	0.7	0	0.1	3.7	28	0	0
Outside	0	98	0.1	0.1	0	0.2	0.2	1.9	0	0.1
Dosage	0	12	80	2.8	0.3	3.5	0.8	0.2	0.6	0
Drugname	0	11	1.6	82	0.2	1.9	1.4	1.5	0.2	0
Duration	0	38	0.9	0.8	56	3.8	0.1	0.8	0	0
Frequency	0	19	1.2	0.9	0.5	77	1.1	0.1	0.3	0
Indication	0.4	18	0.1	1.4	0.1	0.6	44	34	0.1	0.6
Other SSD	0.2	21	0	0.1	0	0	1.5	77	0	0.2
Route	0	14	1.8	3.5	0.1	2.9	1	0.4	76	0
Severity	0.1	35	0	0.2	0	0	1.7	10	0	52

GRU-sentence

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	46	21	0	0.3	0	0	5.6	27	0.1	0
Outside	0	99	0	0.1	0	0.1	0	1.1	0	0.1
Dosage	0	5.5	93	0.7	0.2	0.4	0.1	0	0.3	0
Drugname	0	6.7	0.4	92	0	0.1	0.1	0.4	0.1	0
Duration	0	23	0.7	0.1	72	3.8	0.1	0	0	0.2
Frequency	0	8	0.3	0.1	0.4	91	0.1	0	0	0
Indication	0.5	9.6	0	0.5	0	0.1	58	30	0	0.8
Other SSD	0.2	13	0	0	0	0	1.3	85	0	0.2
Route	0	9.4	0.5	0.7	0	0.1	0.1	0.1	89	0
Severity	0	25	0	0	0.1	0	0.6	5.3	0	69

LSTM-sentence

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	48	19	0	0.6	0	0	5.1	26	0.1	0.1
Outside	0	99	0	0.1	0	0.1	0	1.1	0	0.1
Dosage	0	4.8	93	0.9	0.1	0.7	0.1	0.1	0.2	0
Drugname	0.1	6.2	0.5	93	0	0.1	0.2	0.4	0.1	0
Duration	0	23	0.8	0.1	72	4.8	0.1	0	0	0
Frequency	0	7.1	0.5	0.1	0.3	92	0.1	0	0	0
Indication	0.6	10	0	0.6	0	0	59	28	0.1	0.7
Other SSD	0.2	14	0	0	0	0	1.6	84	0	0.2
Route	0	8	0.8	0.7	0	0.3	0.1	0.1	90	0
Severity	0	26	0.1	0	0.1	0	0.4	5.1	0	68

LSTM-document

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	49	21	0	0.4	0	0	5	24	0.2	0.1
Outside	0	99	0	0.1	0	0.1	0	1	0	0.1
Dosage	0	5.1	93	0.8	0.1	0.5	0	0.1	0.2	0.1
Drugname	0.1	6	0.6	93	0	0.1	0.1	0.5	0.1	0
Duration	0	25	0.5	0.1	68	6.5	0.1	0	0	0.1
Frequency	0	6.1	0.3	0.1	0.3	93	0.1	0	0	0
Indication	0.9	10	0	0.6	0	0	64	24	0.1	0.9
Other SSD	0.3	13	0	0	0	0	1.1	85	0	0.2
Route	0	9.6	0.6	0.8	0	1.2	0	0	88	0
Severity	0	26	0.4	0	0	0	0.4	5.8	0	67

GRU-document

	ADE	Outside	Dosage	Drugname	Duration	Frequency	Indication	Other SSD	Route	Severity
ADE	48	22	0	0.4	0	0	5.1	25	0.1	0.1
Outside	0	99	0	0.1	0	0.1	0	1	0	0.1
Dosage	0	4.9	93	0.8	0.1	0.5	0	0	0.2	0
Drugname	0	5.8	0.7	93	0	0.1	0.1	0.3	0.1	0
Duration	0	23	0.4	0.1	71	6	0.1	0	0	0
Frequency	0	6.1	0.4	0.1	0.1	93	0.1	0	0.1	0
Indication	0.8	9.8	0	0.5	0	0.1	65	23	0	0.8
Other SSD	0.2	13	0	0	0	0	1	85	0	0.2
Route	0	8.8	0.5	0.4	0	0.3	0.2	0.1	90	0
Severity	0	24	0	0.1	0	0	0.5	4.7	0	70

Heat-maps of Confusion Matrices of each method for the different class Labels. Rows are reference and columns are predictions. The value in cell *(i, j)* denotes the percentage of words in label *i* that were predicted as label *j*.

Conclusions

- Recurrent Neural Networks perform significantly better than Conditional Random Fields.
- This is mainly because RNNs are able to learn from dependencies that occur at variable context lengths.
- Since RNNs are able to better recognize contextual cues, they have a significantly higher recall than CRF models.
- RNN models are useful even for domain specific tasks with small datasets.
- Training RNN models on the entire document, instead of on individual sentences seems promising. The performance of document models can likely improve even for labels like ADE, with a larger training dataset.

Goal: EHR-based cancer drug safety research

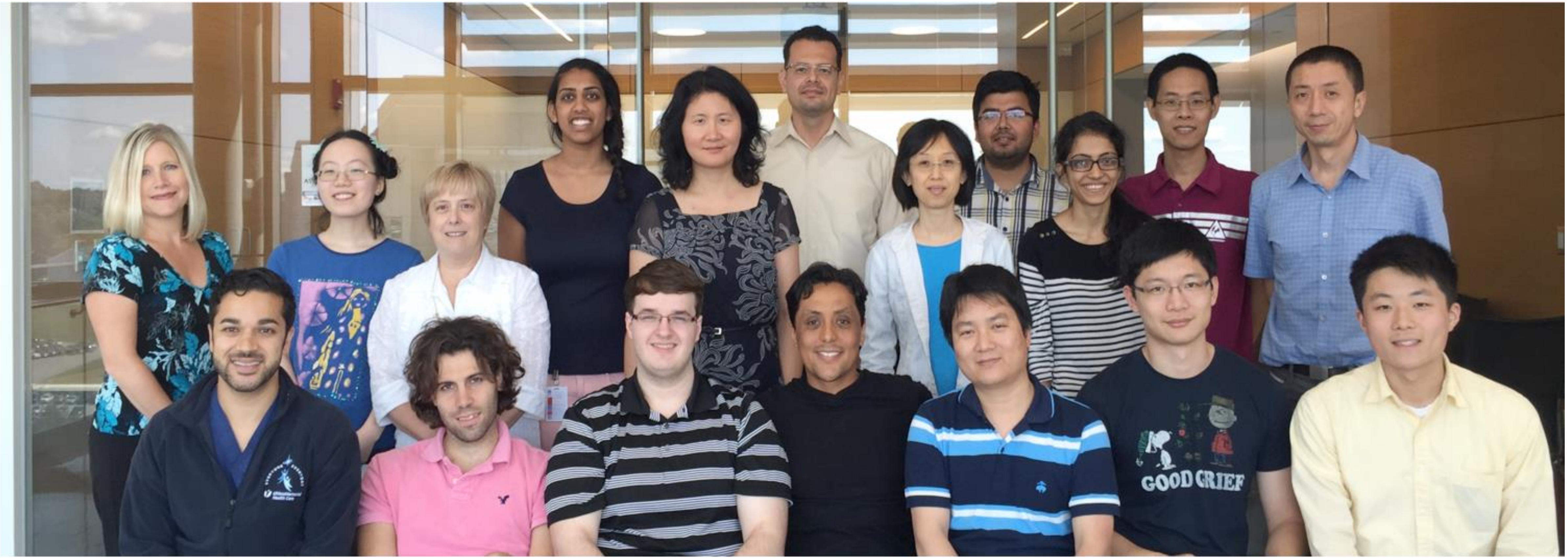
- What we have achieved
  - Developing annotation guideline and annotated 1000 EHR notes
  - Build a high-performance state-of-the-art NLP system
  - New Work:
    - Improve to 0.87 F1 score from 0.8
    - Information extraction 0.97 F1 score
    - Mapping entity to controlled terminology

Deploying the NLP system to many hospital systems

- Umass Medical School
- Tufts Medical School
- Northwestern University
- Upenn
- Yale University
- VA
- …

The UMass BioNLP Group

- My group's research involves developing **innovative biomedical informatics algorithms and tools** for gathering, analyzing and interpreting heterogeneous **data** from multiple sources for personalized medicine.



- https://youtu.be/cSXLXwWdlck