Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Natural Language Processing for EHR-based Pharmacovigilance: Current progress and future directions

xx

Abhyuday Jagannatha[1], Bhanu Pratap Singh Rawat[1], UMass BioNLP[2], and Hong Yu[1,2,3,4]

[1]UMass Amherst; [2]UMass Lowell; [3]UMass Medical School; [4]VA Bedford Healthcare System

Sentinel

U.S. Department of Veterans Affairs

Disclaimer

I have no potential conflicts of interest to report.

The views and opinions expressed in this presentation are my personal views and opinions and must not be construed as representing the views and opinions of any institute from which I have received funding, or with which I am or have been affiliated in my professional life.

Pharmacovigilance, EHR Notes, NLP

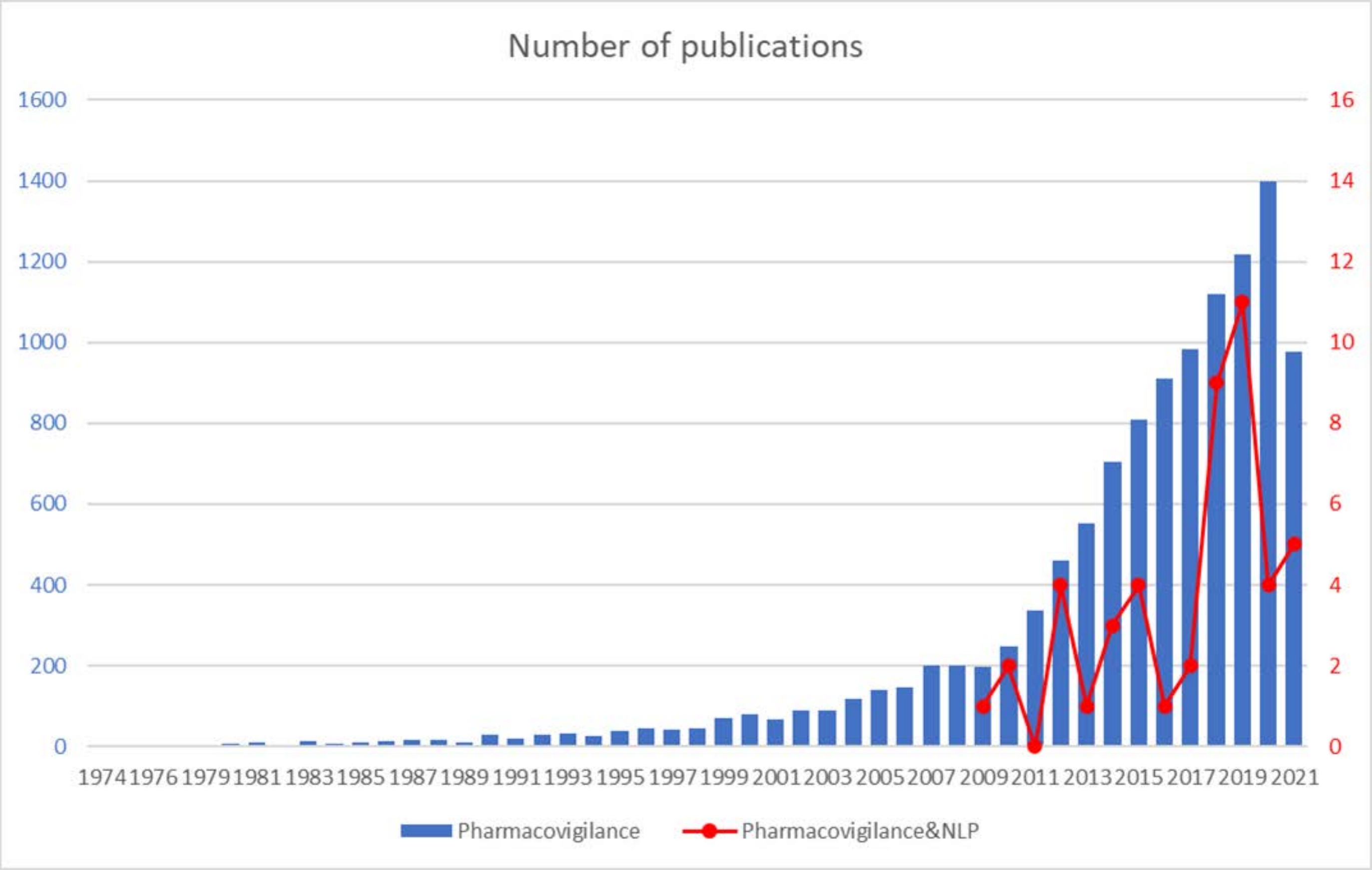
"Pharmacovigilance is the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem." -- WHO

Traditional sources of information: clinical trials, pharmaceutical industry reports, and adverse-event spontaneous reporting databases.

Electronic health records notes contain rich descriptions in adverse events frequently not available in the structured data.

Natural language processing (NLP) methods can be a powerful tool for detecting medications and adverse drug events.

Pharmacovigilance and Natural Language Processing



Search in PubMed

Talk Outline

1. Natural language processing and machine learning approaches
    - 1a. MADE 1.0 cohort
    - 1b. MADE 1.0 NLP challenges
    - 1c. Extracted ADEs
    - 1d. Naranjo question answering

2. Calibration methods

3. Membership inference attack susceptibility for NLP

Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0 Challenge)

xx

Abhyuday Jagannatha[1], Feifan Liu[2], Weisong Liu[3], Hong Yu[1,2,3,4]

[1]UMass Amherst; [2]UMass Medical School; [3]UMass Lowell; [4]VA Bedford Healthcare System

Sentinel

MADE 1.0 Challenge

An annotated cohort of 1,089 EHR notes from 21 patients with cancer, comprising 79,003 Named Entities(NE) annotated with 9 NE types 27,328 relations between Named Entities with 7 Relation types

A shared task focused on extracting fine-grained entity information related to medication and adverse drug events (ADEs)

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

The annotation team: Elaine Freund, Edgard Granillo, Heather Keating, Raelene Goodwin, and Nadya Frid

Sample Sentence and Annotations from MADE 1.0

Input: "*Hypertension* is well controlled on current dose of *atenolol* *50 mg* *daily* and *doxazosin* *4 mg* *daily*."

Output:

Named Entities

Indication: *Hypertension*

Drugname: *atenolol* , *doxazosin*

Dosage: *50 mg*, *4mg*

Frequency: *daily*, *daily*

Relations

Reason: (*Hypertension* **,** *atenolol***), (***Hypertension* **,** *doxazosin***)**

Dosage Relation : (*atenolol*, *50 mg* **), (***doxazosin*, *4 mg* **)**

Frequency Relation : (*atenolol*, *daily* **), (***doxazosin*, *daily* **)**

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

The annotation team: Elaine Freund, Edgard Granillo, Heather Keating, Raelene Goodwin, and Nadya Frid

MADE 1.0 Named Entities and Relations

Events
- ADE
- Indication
- Other SSD
- Medication

Attributes
- Severity
- Duration
- Dose
- Route
- Frequency

Relations
- Medication/Dose
- Medication/Route
- Medication/Frequency
- Medication/Duration
- SSD/Severity
- Medication/AE
- Medication/Indication

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

The annotation team: Elaine Freund, Edgard Granillo, Heather Keating, Raelene Goodwin, and Nadya Frid

MADE 1.0 Statistics

79,003 Named Entities(NE) annotated with 9 NE types 27,328

Annotation counts, and word counts for each named entity type.

NE type	Number of Annotations	Total annotated words
ADE	1940	3255
Indication	3804	8240
Other SSD	39384	82956
Severity	3908	5069
Drugname	15902	19075
Dosage	5694	11820
Duration	898	1768
Frequency	4806	11400
Route	2667	2805

Total: 79,003

Relation type	Occurrences	Relation length
ADE - Drugname	2612	82 ± 187 (3662,1)
SSD - Severity	4035	4.7 ± 34.41 (1861,0)
Drugname - Route	3006	18 ± 25(224,1)
Drugname - Dosage	6043	11 ± 22(230,0)
Drugname - Duration	1053	20 ± 27(273,1)
Drugname - Frequency	5149	25 ± 30(295,1)
Indication - Drugname	5430	96 ± 164 (2742,1)

Total: 27,328

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

The annotation team: Elaine Freund, Edgard Granillo, Heather Keating, Raelene Goodwin, and Nadya Frid

MADE 1.0 NLP Tasks

Task 1 : Named Entity Recognition

Task 2 : Relation Identification between Annotated Named Entity

Task 3 : Relation Identification

Evaluation Criterion : **Micro average F-score using exact phrase** based evaluation for Standard Track.

Test data : (213 notes)
All EHR notes from 3 patients + 4 notes from each of the remaining 18 patients(# of records > 8).
Training data : (876 notes)
Remaining notes.

Task	Teams
NER	10
RI	5
NER+RI	4

F-score : 0.8290 Worcester Polytechnic Institute (Wunnava et al)

F-score : 0.8684 University of Utah (Alec et al)

F-score : 0.6170 IBM Research (Dandala et al)

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

Methods

Team Names	LSTM	CRF	PWE	CE	Features	Relation Classifier
UCA-I3S-SPARK [65]	+	–	+	+	POS	–
UFL-gators [60]	+	–	+	–	–	–
UofUtah-Patterson [62]	–	+	+	–	POS, Surface	Random Forest
ASU-BMI [64]	+	+	+	+	Surface	Random Forest
IBMResearch-dandala [59]	+	+	+	+	POS	Attention Bi-LSTM
WPI-Wunnava [58]	+	+	+	+	–	–
UArizonaIschool-Xu [61]	+	+	+	+	Prefix, Suffix Embedding	SVM
AEHRC-HoaNGO [63]	–	+	+	–	Snomed-CT, POS, Dependency	–

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

MADE 1.0 Annotation Inconsistency

Teams discovered a few inconsistencies

- Inconsistent annotations with period.
  - E.g. “p.o.” “p.o”
- Overlapping Annotations
  - E.g. “Multivitamin” “Multivitamin (TAB-A-VITE)”
- Double Annotations
  - E.g. “pulmonary toxicity” annotated as (Other SSD, ADE)

Jagannatha et al, 2020. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf, 2019 Jan; 42 (1): 99-111.

Extracted ADEs from an EHR Cohort

We extracted ADEs from an EHR cohort: 200,129 patients, 2,449,944 notes.

We extracted a total of 2,547,445 medication and ADE pairs

The most frequent ADE is “allergies”

The remaining most frequent 20 ADEs are:

'Systemic disease'	'Aspirin'
'Myalgia'	'Hydroxymethylglutaryl-CoA Reductase Inhibitors'
'Nausea'	'Zofran'
'Constipation'	'Narcotics'
'Apnea'	'Benzodiazepine'
'Leukocytosis'	'Steroids'
'Hyperlipidemia'	'Hydroxymethylglutaryl-CoA Reductase Inhibitors'
'Diarrhea'	'Metformin'
'Dizziness'	'Adrenergic alpha-Antagonists'
'Hypoglycemia'	'Insulin'
'Acidosis, Lactic'	'Metformin'
'Reaction'	'Lisinopril'
'Myalgia'	'atorvastatin'
'Neuropathy'	'gabapentin'
'Myalgia'	'Pravastatin'
'Hyperglycemia'	'Steroids'
'Constipation'	'Miralax'
'Dizziness'	'Lisinopril'
'Angioedema'	'Lisinopril'
'Dizziness'	'Meclizine'

Jagannatha and Yu, 2016. Bidirectional Recurrent Neural Networks for Medical Event Detection in Electronic Health Records. NAACL 2016.

Jagannatha and Yu, 2016. Structured prediction models for RNN based sequence labeling in clinical text. EMNLP 2016.

Munkhdalai et al. 2018. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. JMIR Public Health Surveill. Apr 25;4(2):e29. doi: 10.2196/publichealth.9361.

Extracted ADEs from an EHR Cohort

Positive Examples:

ADE	Drug	Analysis
Angioedema	Lisinopril	ACE Inhibitor-Related Angioedema. Kaufman 2013.
Myalgia	Simvastatin	Statin induces myalgia and mytosis. O'Callaghan 2018
Constipation	Narcotics	Opiod induced Constipation and Bowel Dysfunction. Muller-Lissner et. al. 2016
Apnea	Benzodiazepine	Benzodiazepines, breathing and sleep. Guilleminault 1990

Negative Examples:

ADE	Drug	Analysis
Nausea	Zofran	Indication or ADE?
Systemic Disease	Aspirin	Generic ADE

Evaluation by Physicians: Selected 120 extracted unique ADEs based on distributions. One physician judged 115 yes and 5 no. The second physician judged 18 no, 49 yes, and 53 unevaluable.

Manuscript in preparation

Naranjo Question Answering using End-to-End Multi-task Learning Model

xx

Bhanu Pratap Singh Rawat[1], Fei Li[2], and Hong Yu[1,2,3,4]

[1]UMass Amherst; [2]UMass Lowell; [3]UMass Medical School; [4]VA Bedford Healthcare System

Sentinel

Naranjo Question Answering

Naranjo Scale was developed to standardize the assessment of causality for adverse drug reactions (ADRs).

It comprises of 10 questions which can be answered as ‘Yes’, ‘No’ and ‘Do not know’.

	#	Naranjo Questions	Yes	No	Do not know
Unable to answer from clinical notes	1.	Are there previous conclusive reports on this reaction?	1	0	0
	2.	Did the adverse event occur after the suspected drug was administered?	2	-1	0
	3.	Did the adverse reaction improve when the drug was discontinued or a specific antagonist was administered?	1	0	0
Few cases	4.	Did the adverse reaction reappear when the drug was readministered?	2	-1	0
	5.	Are there alternative causes (other than the drug) that could have on their own cause the reaction?	-1	2	0
Not always present for each patient.	6.	Did the reaction reappear when a placebo was given?	-1	1	0
	7.	Was the drug detected in the blood (or other fluids) in concentrations known to be toxic?	1	0	0
Few cases	8.	Was the reaction more severe when the dose was increased or less severe when the dose was decreased?	1	0	0
Few cases	9.	Did the patient have a similar reaction to the same or similar drugs in any previous exposure?	1	0	0
	10.	Was the adverse event confirmed by any objective evidence?	1	0	0

doubtful: <= 0
possible: 1 <= **score** <=4
probable: 5 <= **score** <=8
definite: >=9

Can we automate this QA system?

Naranjo et al., 1981. *A method for estimating the probability of adverse drug reactions. Clinical Pharmacology & Therapeutics*, *30*(2), pp.239-245.

Naranjo Cohort Selection

- We built an expert annotated EHR cohort to be used for training and evaluation for automated Naranjo QA.
- We selected clinical notes of patients who were administered one of these six anticoagulants: *Apixaban, Clopidogrel, Dabigatran, Enoxaparin, Rivaroxaban* and *Warfarin*.
- Physician annotators manually examined those notes and provided answers for each Naranjo question.
- Experts provided two levels of annotation: *relevant* sentences and *answer* for each Naranjo question.

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

The annotation team is led by Drs. Stephen Belknap and Feifan Liu. The annotators are: Stephen Belknap, William Temps, Nadya Frid, and Edgard Granillo.

Naranjo QA EHR Dataset

- 991 patients with 1,385 discharge summaries.
- Eliminated questions 1 and 6. The remaining questions 2, 3, 5, 7 and 10 were most frequently answered by the experts.

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

The annotation team is led by Drs. Stephen Belknap and Feifan Liu. The annotators are: Stephen Belknap, William Temps, Nadya Frid, and Edgard Granillo.

Naranjo QA EHR Dataset

Anitcoagulant	# Unique Patients	# Discharge Summaries
Dabigtran	38	48
Apixaban	82	121
Rivaroxaban	85	116
Enoxaparin	141	181
Clopidogrel	169	212
Warfarin	476	707

Distribution of unique patients and discharge summaries across different anticoagulants.

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

The annotation team is led by Drs. Stephen Belknap and Feifan Liu. The annotators are: Stephen Belknap, William Temps, Nadya Frid, and Edgard Granillo.

Naranjo Annotation

- Two levels of annotation:
  a. *Relevant sentence* for which the Naranjo question has been answered.
  b. *Answer* for the specific Naranjo question.

Paragraph from EHR: Upon arrival to ER, pt developed massive coffee-ground hematemesis (no BRB) x1. In ED, VS notable for 96.6, 98/58 --> 120/60s a/p 1L NS (b/I BP 130s/80s), 70-80s (on BB), 16, 100% RA. NGL notable for coffee-ground hematemesis. Recta q/ melena, no BRBPR. Hbb 7.8, INR 1.9. The pt. was then admitted to MICU for further mg'mt and was started on nexium gtt, T&S'd. 18G PIV x2 placed

Naranjo question: Did the adverse event occur after the suspected drug was administered?

Answer: Yes

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

The annotation team is led by Drs. Stephen Belknap and Feifan Liu. The annotators are: Stephen Belknap, William Temps, Nadya Frid, and Edgard Granillo.

Naranjo QA EHR Dataset

Causal Relation	Condition	# Discharge Summaries
Doubtful	$N_{score} \leq 0$	183
Possible	$1 \leq N_{score} \leq 4$	916
Probable	$5 \leq N_{score} \leq 8$	283
Definite	$9 \leq N_{score}$	3

Distribution and condition for each causal relation between the medication and its ADEs.

Distribution of answers for selected 5 questions.

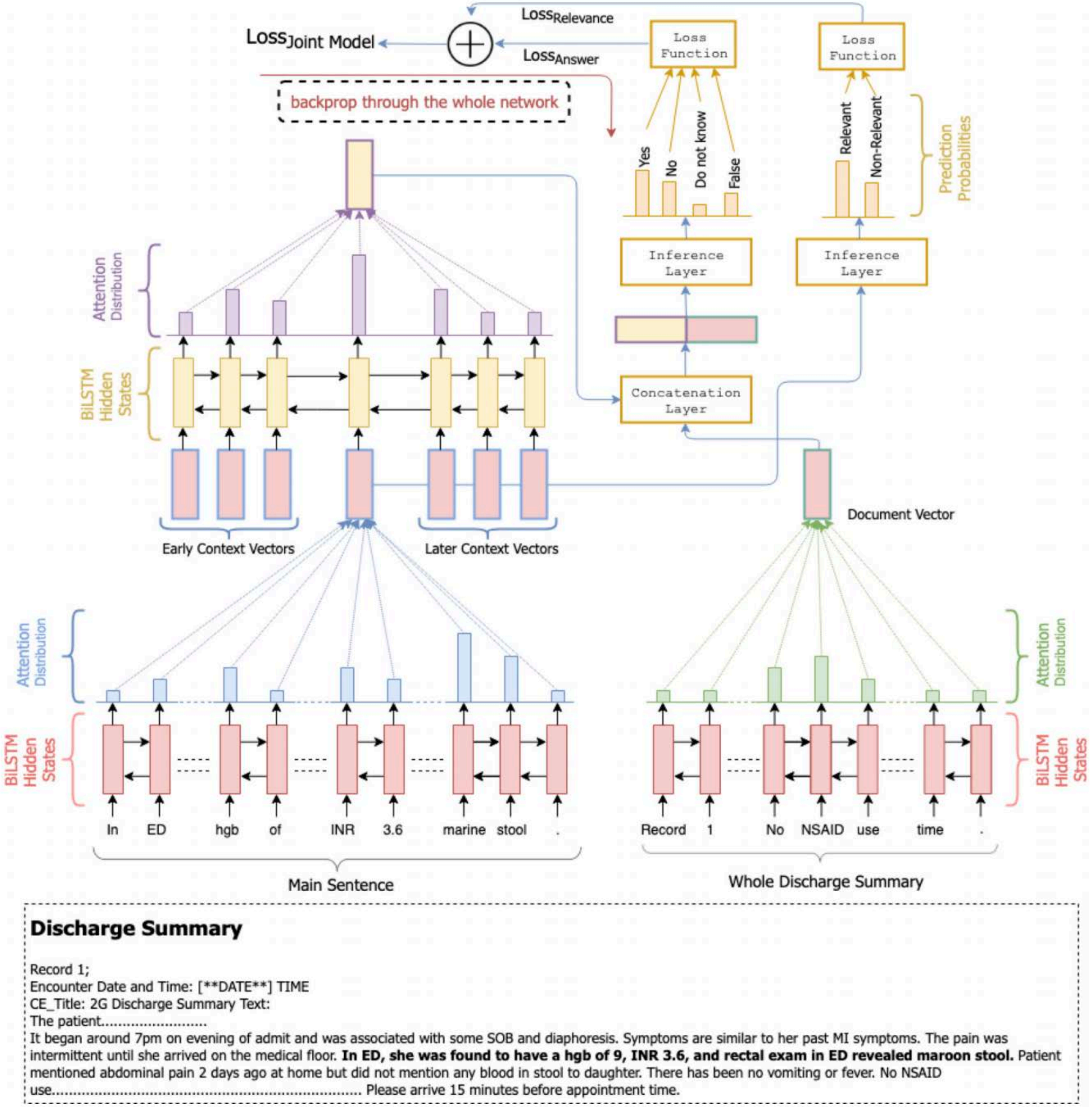
Question #	Yes	No	Do not know
2	1633	139	666
3	381	21	181
5	2186	221	316
7	619	29	76
10	1683	678	227

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

The annotation team is led by Drs. Stephen Belknap and Feifan Liu. The annotators are: Stephen Belknap, William Temps, Nadya Frid, and Edgard Granillo.

Multitask Learning for Naranjo QA

$$Loss_{JointModel} = Loss_{Relevance} + Loss_{Answer}$$



Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

Performance

Ques #	Model	MA-P	MA-R	MA-F	MI-P	MI-R	MI-F
Ques 2	Pipeline	0.3045	0.3185	0.3105	0.9313	0.9313	0.9313
	JM	0.4445	0.4472	0.4424	0.95452	00.9545	0.9545
	JM-Doc	0.4608	0.4677	0.4633	0.9592	0.9592	0.9592
	JM-Doc down	0.4874	0.4952	**0.4827**	0.9624	0.9624	**0.9624**
Ques 3	Pipeline	0.3657	0.3776	**0.3675**	0.9809	0.9809	0.9809
	JM	0.3459	0.3423	0.3434	0.9902	0.9902	0.9902
	JM-Doc	0.6640	0.3381	0.3415	0.9918	0.9918	**0.9918**
	JM-Doc down	0.3546	0.4007	0.3652	0.9780	0.9780	0.9780
Ques 5	Pipelinec	0.3137	0.3302	0.3209	0.9313	0.9313	0.9313
	JM	0.3791	0.4181	0.3884	0.9404	0.9404	0.9404
	JM-Doc	0.3722	0.3907	0.3758	0.9434	0.9434	0.9434
	JM-Doc down	0.4054	0.3859	**0.3936**	0.9523	0.9523	**0.9523**
Ques 7	Pipeline	0.2785	0.3070	0.2876	0.9728	0.9728	0.9728
	JM	0.2890	0.3523	0.3054	0.9694	0.9694	0.9694
	JM-Doc	0.3838	0.3558	**0.3678**	0.9874	0.9874	**0.9874**
	JM-Doc down	0.3587	0.3585	0.3409	0.9858	0.9858	0.9858
Ques 10	Pipeline	0.3275	0.3274	0.3260	0.9288	0.9288	0.9288
	JM	0.5017	0.4826	0.4886	0.9535	0.9535	0.9535
	JM-Doc	0.5104	0.4628	0.4779	0.9542	0.9542	**0.9542**
	JM-Doc down	0.5394	0.5365	**0.5271**	0.9494	0.9494	0.9494

Rawat et al. Naranjo question answering using end-to-end multi-task learning model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019 Jul 25 (pp. 2547-2555).
Rawat et al. Clinical Judgement Study using Question Answering from Electronic Health Records. In Machine Learning for Healthcare Conference 2019 Oct 28 (pp. 216-229). PMLR.
Rawat et al. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 1041).

Calibrating Structured Output Predictors for Natural Language Processing

xx

Abhyuday Jagannatha[1], Hong Yu[1,2,3,4]

[1]UMass Amherst; [2]UMass Medical School; [3]UMass Lowell; [4]VA Bedford Healthcare System

Sentinel

Calibration

Provide confidence scores with NLP structured output predictions.

Calibration is important for healthcare.

Widely used methods are often defined as binary or multi-class problems, not for structured outputs.

We propose a general calibration scheme for structured outputs in neural network-based NLP models.

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Main Contributions

We use the calibration framework from Kuleshov et. al. (2015) to define a general structured calibration scheme for NLP systems with the following properties:

1. It can use any binary class forecaster to calibrate the predictor confidence for a defined output entity of interest.
2. It provides better calibration than standard methods for both in-domain and out-of-domain samples.
3. The forecaster confidence can also be used to rescore output entities, and improve underlying predictor performance (in-domain and out-of-domain).

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Experiment Overview

We define calibration schemes for the following NLP tasks :

- Part-of-Speech (Penn Treebank)
- Named Entity Recognition (CoNLL-2003, MADE 1.0)
- Question Answering (SQuAD, emrQA, MADE 1.0)

Neural Network models such as BERT have calibration errors ranging from 3.5 to 30 % on these tasks.

We evaluate our calibration scheme on each of these tasks.

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Experimental Details

Forecaster Features:

- Model outputs and epistemic uncertainty.
- Entity of Interest specific features.
- Distributional Uncertainty using LM perplexity.

Evaluation Tasks:

- Part-of-Speech
- Named Entity Recognition
- Question Answering

- Forecaster Model:
  - Gradient Boosted Decision Tree.
- Uncalibrated Predictors:
  - BERT
  - DistilBERT
  - BERT-CRF
- Evaluation Criterion:
  - Expected Calibration Error (Naeini et. al. 2015).

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Results

Calibration Performance :

- **Improves** on Penn Treebank POS task
- **Improves** on CoNLL-2003 and **MADE 1.0** NER tasks.
- **Improves** on SQuAD 1.0, emrQA(medical) and MADE 1.0 (medical) QA tasks.
- **Improves** on out-of-domain evaluation (emrQA → MADE 1.0)

Classifier Performance after re-scoring :

- **Remains competitive** to baseline on Penn Treebank POS task
- **Improves** on CoNLL-2003 and **MADE 1.0** NER tasks. **0.8434 F1 score**
- **Improves** on SQuAD 1.0, emrQA (medical) and MADE (medical) QA tasks.
- **Improves** on out-of-domain evaluation (emrQA → MADE 1.0).

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Conclusion

Structured Prediction models for NLP have an estimate of the expected difference between the model confidence and accuracy as high as 30%.

We provide a general calibration scheme to :

- Calibrate only those output entities that are relevant for model prediction.
- Use any binary class calibration method as a plug-in to improve calibration for output entities of interest.
- Enrich the forecaster training data and improve calibration performance.
- Rescore the output entities to improve the predictor performance.

Abhyuday Jagannatha and Hong Yu, 2020. Calibrating structured output predictors for natural language processing. ACL: 2078-2092.

Membership Inference Susceptibility of Clinical Language Models

xx

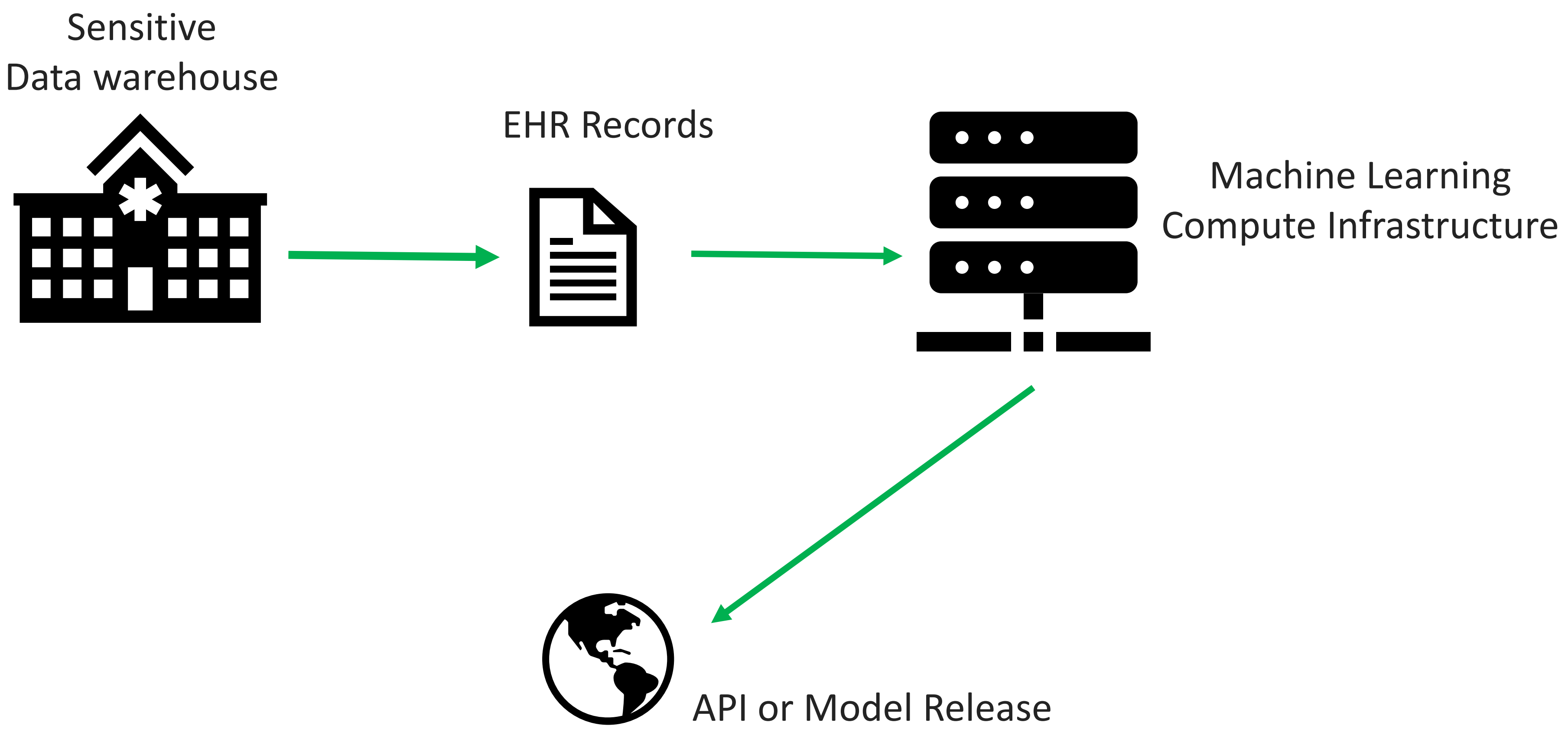
Abhyuday Jagannatha[1], Bhanu Pratap Singh Rawat[1], Hong Yu[1,2,3,4]

[1]UMass Amherst; [2]UMass Medical School; [3]UMass Lowell; [4]VA Bedford Healthcare System

Paper under review

Sentinel

How Safe are Machine Learning Models Trained on EHRs?



Jagannatha et al. Membership Inference Attack Susceptibility of Clinical Language Models.
Paper under review.

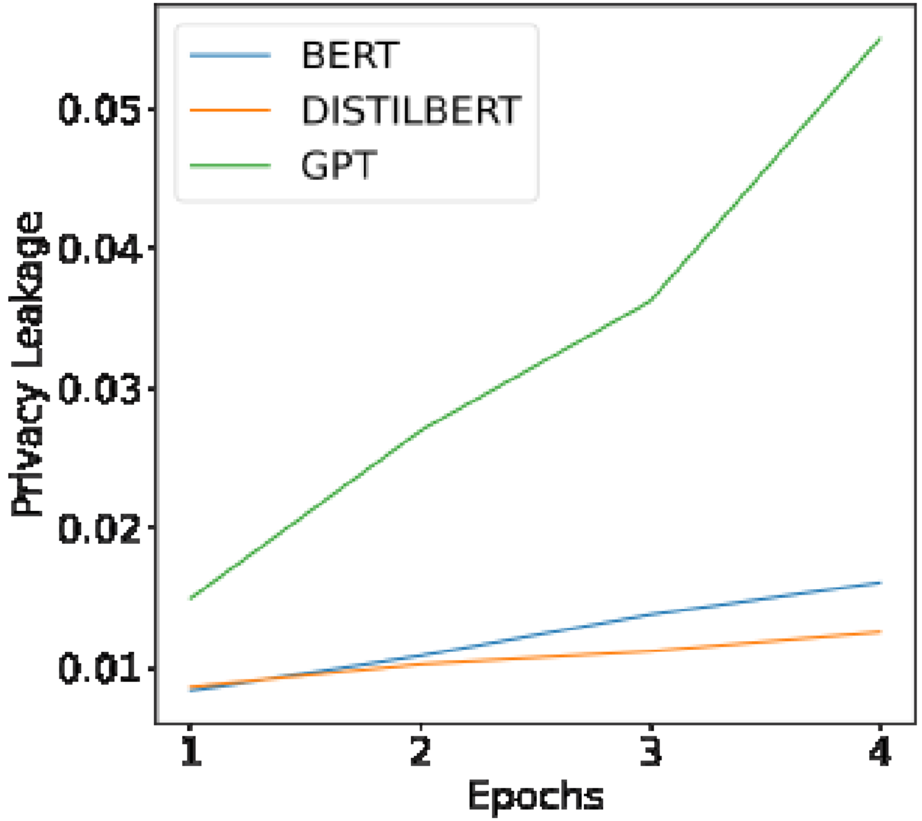
Differential Privacy

Differential privacy describes a promise, made by a data holder, or curator, to a data subject: “You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.” --Dwork and Roth, 2014

Membership Inference: Given a model and a data sample, can the attacker infer whether the data sample was a part of the training set ? --Shokri et al, 2017

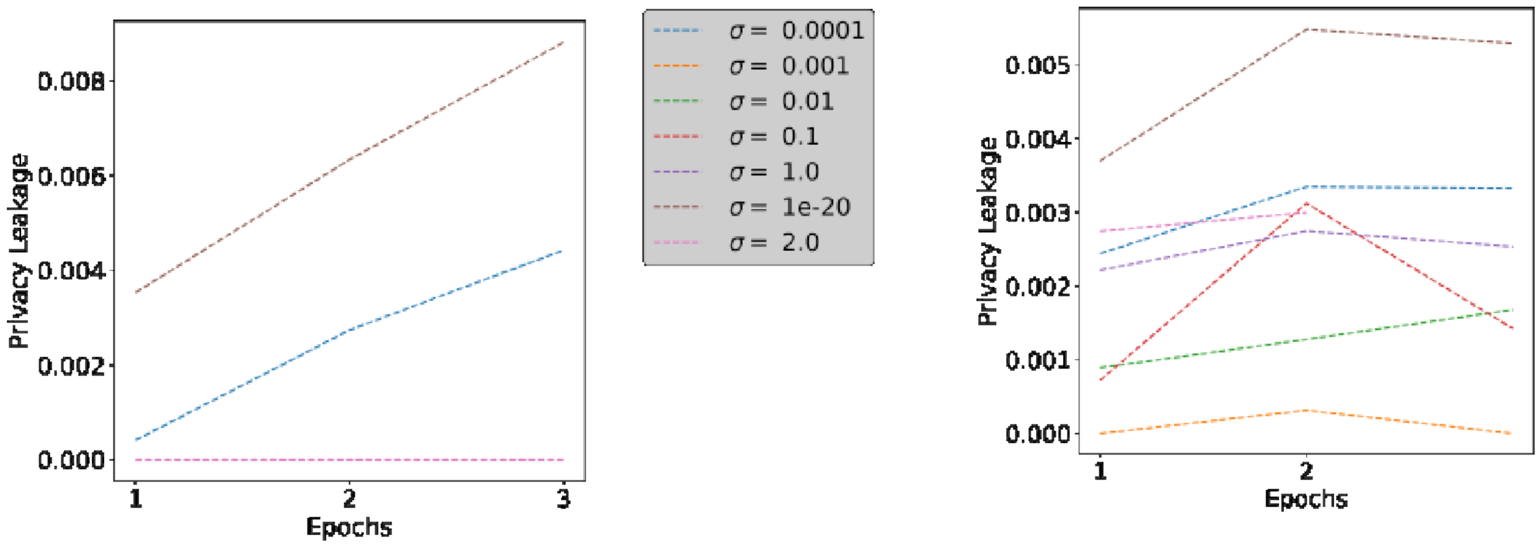
Jagannatha et al. Membership Inference Attack Susceptibility of Clinical Language Models. Paper under review.

Privacy Leakage in Language Models



Jagannatha et al. Membership Inference Attack Susceptibility of Clinical Language Models.
Paper under review.

Privacy Leakage in Private Language Models



Differentially Private GPT2

Differentially Private BERT

Privacy leakage estimates for different gradient noise σ values.

Jagannatha et al. Membership Inference Attack Susceptibility of Clinical Language Models.
Paper under review.

Conclusions

Large LMs have higher empirical privacy leakages (9%) than smaller LMs (2%).

Randomly masked LMs have lower privacy leakages than autoregressive LMs.

Training using DP-SGD (Dwork et al., 2014) can reduce empirical privacy leakages while ensuring increased model utility.

Users with rarer profiles may be more vulnerable to higher privacy leakages.

Jagannatha et al. Membership Inference Attack Susceptibility of Clinical Language Models. Paper under review.

Future Directions

XX

Future Directions

Annotation, Annotation, and Annotation!

Unsupervised learning, domain adaptation

Naranjo question answering

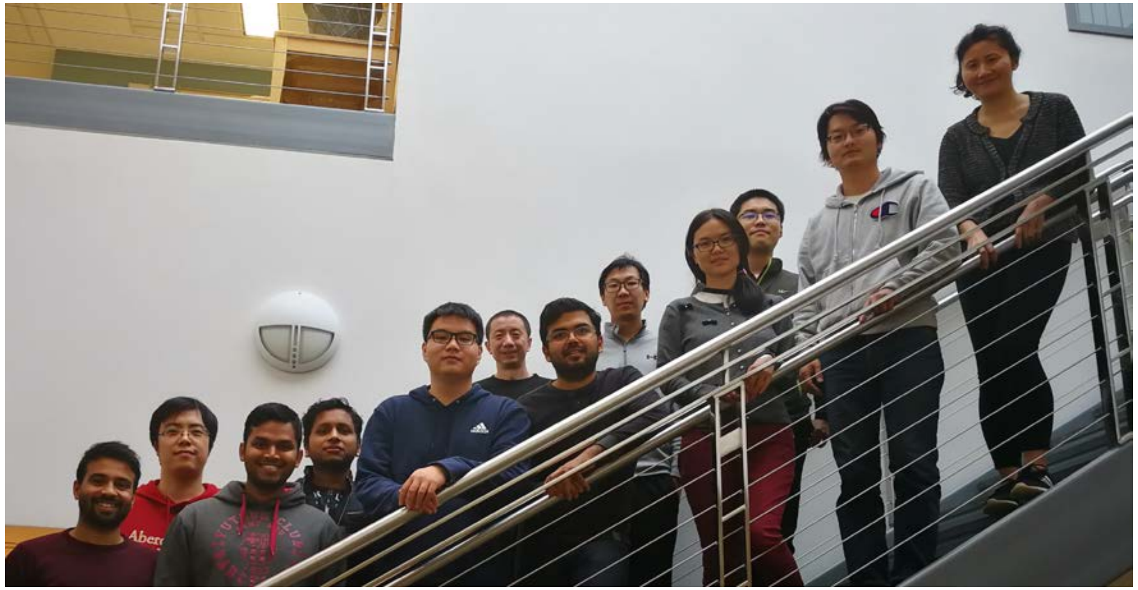
Data integration and model development based on multisource data

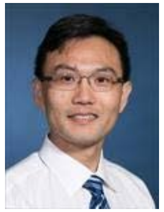
Semi-supervised tool development (e.g., ADEPT, Geva et al**)**

Acknowledgement

Abhyuday Jagannatha
PhD candidate
Umass Amherst
abhyuday@cs.umass.edu

Bhanu Pratap Singh Rawat
PhD candidate
Umass Amherst
brawat@umass.edu





Annotators: Heather Keating, Elaine Freund, Raelene Goodwin, Nadya Frid

Acknowledgement

U01 CA180975. R01 HL 125089

NIH NATIONAL CANCER INSTITUTE

NIH National Heart, Lung, and Blood Institute

Contact Information

Hong Yu, MS, PhD, FACMI
Professor, Department of Computer Science
University of Massachusetts Lowell
Hong_yu@uml.edu
https://bio-nlp.org/index.php

Research Health Scientist
VA Bedford Healthcare System

Adjunct Professor
College of Information and Computer Science, UMass Amherst
Department of Medicine, UMass Medical School