Clinical Natural Language Processing

Monica Agrawal

6.871/HST.956
February 17, 2022

Outline

- Uses of clinical text
- Idiosyncrasies of clinical text
- Clinical information extraction
- Summarization and generation of text

Outline

- **Uses of clinical text**
- Idiosyncrasies of clinical text
- Clinical information extraction
- Summarization and generation of text

Clinical text is at every step of the pipeline

Location	Example
Triage Assessment	Gross hematuria x1 episode this AM. Appears anxious.
Past History	• HPI: GH x 1 day, no LUTS, no pain, no trauma. • PMHx: none • Current smoker: 1ppd x 10 yrs
Radiology Notes	Impression: slight increase in tumor size
Cystoscopy Note	A well lubricated flexible cystoscope was inserted
Pathology Note	Poorly differentiated carcinoma, arising in an invasive papillary…
Urology Clinic Note	X is a y year old male found to have muscle-invasive bladder cancer
Patient/Provider communications	Patient called to report…

There's rich information in clinical text

Mr. Blind is a 79-year-old white white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center.

The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient's hematocrit was stable and he was given no further intervention.

The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.

On the morning of the 22nd the patient developed tachypnea with a chest X-ray showing a question of congestive heart failure. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given intravenous Lasix.

orange=demographics
blue=patient condition, diseases, etc.
brown=procedures, tests
magenta=results of measurements
purple=time

Slide credit: Pete Szolovits

The information often doesn't exist elsewhere

Example 1: Procedures that *didn't* happen

"The patient attempted a gastrofin swallow on the 21st, but was unable to cooperate with probable aspiration"

The information often doesn't exist elsewhere

Example 2: *Deviation* from medication plan

Pharmacy records say:

"The doctor ordered 6 weeks of medication starting February 1"

Notes say:

"The patient became nauseous and stopped taking the drug on February 5th"

Okay, so now how can I leverage clinical notes to improve patient care?

Point-of-care: speed up lookup

Both original diagnosis and continual care requires frequent reviews of the medical record

Narrow the Differential Diagnoses

- Obtain **data** to narrow differential diagnoses
- Malignant
  - Anywhere along the urinary tract
    - Kidney – CT scan
    - Ureter – CT scan
    - Bladder - Cystoscopy
    - Prostate – Digital rectal exam (DRE), PSA, Prostate biopsy
    - Urethra - Cystoscopy
- Non-malignant
  - Infection – History, urine culture
  - Stone – CT scan
  - Trauma - History
  - Benign prostatic hyperplasia (BPH) - History, DRE
  - Etc.

A Provider's Perspective

- Night before clinic, review medical record
  - Diagnostic process
    - Re-interpretation of tests (e.g. imaging)

Point-of-care: speed up lookup

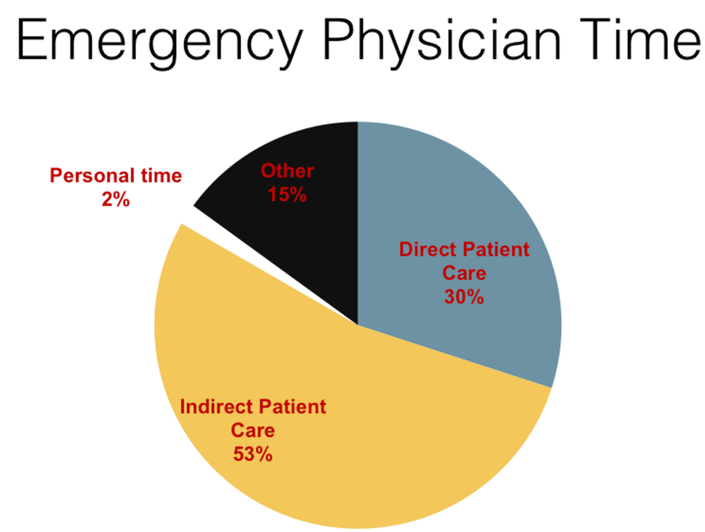
NLP Needs:

- Automated information extraction
  - E.g. smoking status
  - Other variables that feed into mental differential diagnosis algorithm
- Summarization of notes: long-tail of conditions

Point-of-care: speed up writing

NLP Needs:

- Auto-generation of text
    - Conditioned either on:
        - Imaging
        - Conversations



Chrisholm et. al. A Task Analysis of Emergency Physician Activities in Academic and Community Settings. Ann of Emerg Med. 2011.

Point-of-care: non-clinician workflows

- Clinical trial matching
  - Currently slow, often manual process
  - NLP needs: Information extraction

- Billing
  - Extra administrative cost / bloat
  - NLP needs: Information extraction

And many others…

Retrospective Research: new workflows

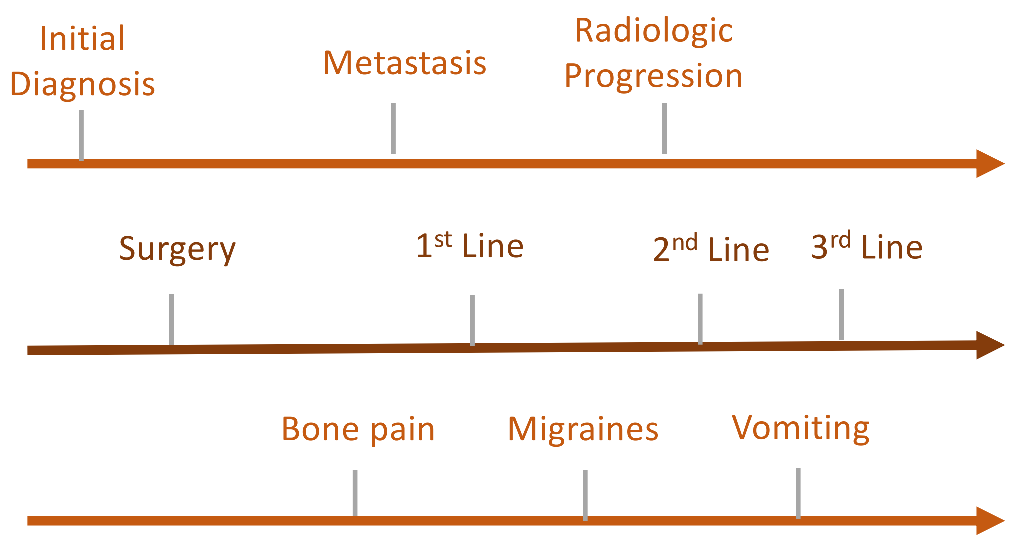
Real-world evidence could help retrospectively respond to needs unmet by trials:

- What treatment would lead to the best outcome for this patient?
  (Heterogeneous treatment effect estimation,
  reinforcement learning)

- What is the patient's expected disease trajectory?
  (Disease progression modeling)

Retrospective Research: Variables

- Disease Status
- Interventions
- Symptoms
- Confounders



Cirrhosis
CHF

NLP needs:

- Information extraction

Patient perspective

Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes

Sigall K. Bell, MD[1,2]; Tom Delbanco, MD[1,2]; Joann G. Elmore, MD, MPH[3]; et al

Author Affiliations | Article Information

JAMA Netw Open. 2020;3(6):e205867. doi:10.1001/jamanetworkopen.2020.5867

- Sent survey to 150k patients across 80 centers, 30k responded
- 1/5 of patients found errors
- 40% described them as serious errors

New Prescriptions: How Well Do Patients Remember Important Information?

Dr. Derjung M. Tarn, MD, PhD and Dr. Susan A. Flocke, PhD

- Approximately 1/3 forgot important information
- Other estimates 40%+ of medical information is immediately forgotten

Midwest Surgical Association

Readability of discharge summaries: with what level of information are we dismissing our patients?

- Only about a quarter could adequately understand their surgical summary
- 65% didn't have reading level

NLP needs:

- Information extraction
- Text generation of cleaner summary

Outline

- Uses of clinical text
- **Idiosyncrasies of clinical text**
- Extraction of relevant variables
- Summarization and generation of text
- Future of clinical NLP

Clinical NLP has unique challenges

- Smaller annotated datasets
  - E.g. can’t rely on Wikipedia
  - Requires domain expertise
  - Difficulty of data sharing across institutions
  - Not an automatic byproduct of clinical practice
- Different language with less rigid syntactic structure
- Long tail of conditions
- Lengthy notes
- High-stakes

How bad can it be to interpret?

How bad can it be to interpret?

3/11/98 IPN	
SOB & DOE ↓	
VSS, AF	
CXR ⊕ LLL ASD no Δ	
WBC 11K	
S/B Cx ⊕ GPC c/w PC, no GNR	
D/C Cef →PCN IV	

Slide credit: Pete Szolovits

How bad can it be to interpret?

3/11/98 IPN	(date of) Intern Progress Note,
SOB & DOE ↓	the patient's shortness of breath and dyspnea on exertion are decreased,
VSS, AF	the patient's vital signs are stable and the patient is afebrile,
CXR ⊕ LLL ASD no Δ	a recent new chest xray shows a left lower lobe air space density that is unchanged from the previous radiograph,
WBC 11K	a recent new white blood cell count is 11,000 cells per cubic milliliter,
S/B Cx ⊕ GPC c/w PC, no GNR	the patient's sputum and blood cultures are positive for gram positive cocci consistent with pneumococcus, no gram negative rods have grown,
D/C Cef →PCN IV	so the plan is to discontinue the cefazolin and then begin penicillin treatment intravenously.

Slide credit: Pete Szolovits

How bad can it be to interpret?

“ <u>Pt</u> given <u>carbo</u> <u>ia</u> for her <u>TNBC</u>. Will <u>dc</u>. ”

How bad can it be to interpret?

Patient? Prothrombin time? Carbodome? Intra-arterial? discontinue?

“ Pt given carbo ia for her TNBC. Will dc. ”

Physical therapist? Carboplatin? Intra-articular? D/C current? discharge? Doctor of Chiropractic?

How bad can it be to interpret?

"Pt given carbo ia for her TNBC. Will dc."

Pt	given	carbo	ia	for	her	TNBC.	Will	dc.
Patient		Carboplatin	Intra-arterial			Triple-neg. breast cancer		Discontinue

Why are notes this way?

Review > Acad Emerg Med. 2004 Nov;11(11):1127-34. doi: 10.1197/j.aem.2004.08.004.

Where's the beef? The promise and the reality of clinical documentation

Steven J Davidson [1], Frank L Zwemer Jr, Larry A Nathanson, Kenneth N Sable, Abu N G A Khan

1. Recording of medical care and communication among providers
2. Payment for hospital and physician
3. Legal defense from medical negligence allegations
4. Symptom/disease surveillance, public health, and research functions

Why are notes this way?

A Provider's Perspective

Cystoscopy Note Continued…

- Assessment: 1 papillary tumor along the right lateral wall. We discussed the role of TURBT given the appearance of a bladder tumor. Specifically, we discussed the diagnostic and therapeutic role. We discussed the potential risks and complications of the procedure, including, … The patient had an opportunity to ask questions and then signed the consent.
- Plan:
  1. TURBT with paralysis, post-operative gemcitabine
  2. PSA to assess risk of prostate cancer

Why are notes this way?

1. Lack of time
2. Note bloat / copy-forwarding

A Provider's Perspective

- MRN
- Name
- DOB

- Urologic Oncology Clinic Note
- ID: Mr. Jones is a 65 year old male who underwent a TURBT on 02/05/2022. He returns today for follow-up.
- HPI: Patient reports that he is doing well since surgery. He denies any pain or lower urinary tract symptoms.
- Pathology: (copy/paste report)

This isn't new

Linguistic Characteristics of Medical Notes

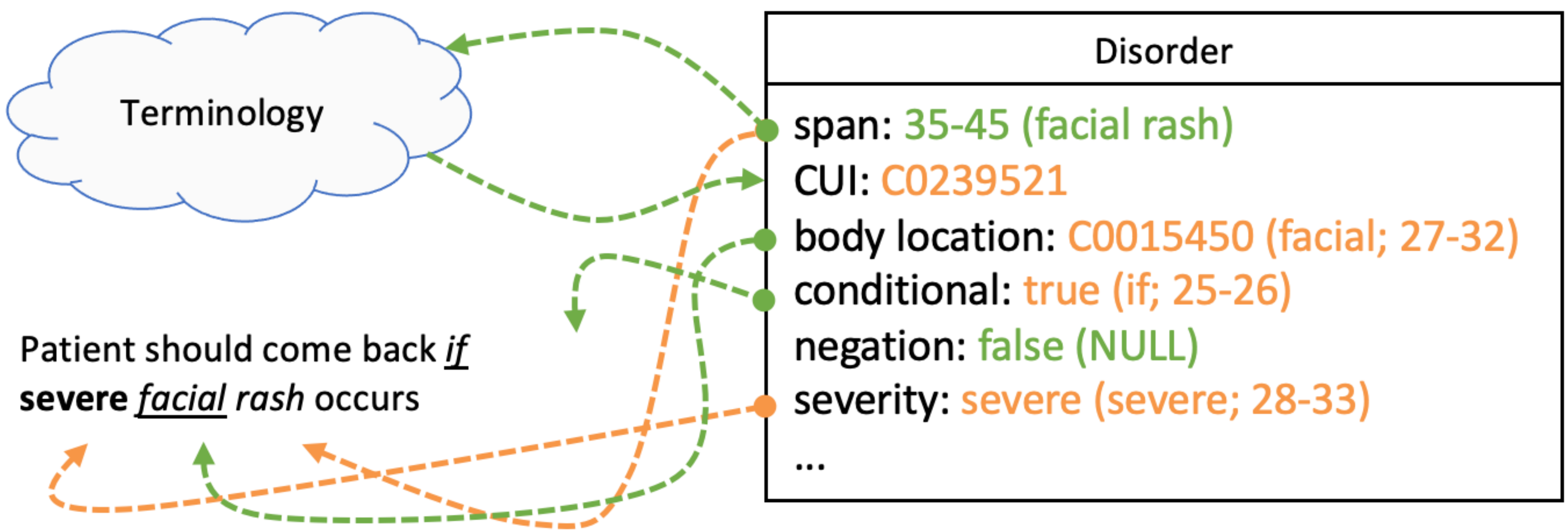
Many of the entries on the medical records are in the form of notes which are neither complete sentences nor single word entries, but linguistic strings of an intermediate type, which we will hereafter call fragments. Fragments are a compressed type of linguistic material resulting from various transformations which have the effect of making linguistic strings shorter by reducing or deleting material. The writer of these stretches of material must make his entries brief, in order to save time and effort, but also make them informative and unambiguous. For this reason the deleted material has to be easily recover-

Anderson et al , *Grammatical Compression in Notes and Records, ACL 1975*

Outline

- Uses of clinical text
- Idiosyncrasies of clinical text
- **Clinical information extraction**
- Summarization and generation of text
- Future of clinical NLP

Information extraction



Slide credit: Noemie Elhadad

Longstanding interest in clinical NLP:

Dataset	Term Types	Entity Recognition?	Entity Normalization?	# Notes
2010 i2b2[1]	Conditions, tests, treatments	✓		871
2012 i2b2[2]	Events	✓		310
2013 ShARe/Clef[3]	Conditions	✓	✓	300
2014 SemEval Task 7[4]	Conditions	✓	✓	+131
2015 SemEval Task 14[5]	Conditions	✓	✓	+100
MCN / 2019 n2c2[6]	Conditions, tests, treatments		✓	100

[1]Uzuner et al 2011, [2]Sun et al 2013, [3]Suominen et al 2013, [4]Pradhan et al 2014, [5]Elhadad et al 2015, [6]Luo et al 2019

Very longstanding interest in clinical NLP:

MEDICAL CLASSES	DESCRIPTION	EXAMPLES IN ENGLISH AND FRENCH
	*** PATIENT AREA ***	
H-PT	words referring to patient	*she, le patient, elle, Mme XXX*
H-PTAREA	anatomical area	*edge, left, surface, rebord, gauche*
H-PTFUNC	physiological function	*BP, TA, appetite, tonalité, digestif*
H-PTLOC	location relation	*radiating, localisé, irradiant*
H-PTMEAS	anatomical measure	*height, size, corpulence, taille*
H-PTPART	body part	*arm, liver, bras, foie*
H-PTPALP	palpated body part	*abdomen, liver, foie*
H-PTSPEC	specimen from patient	*blood, sang, urine*
H-PTVERB	verb with patient subject	*complains of, se plaint de, subi*
	*** TEST / EXAM ***	
H-TXCLIN	clinical exam, action	*auscultation*
H-TXPROC	examination procedure	*ultrason, gastroscopie*
H-TXSPEC	test of specimen	*urine analysis*
H-TXVAR	test variable	*glucose, GB, sédiment*
	*** TREATMENT AREA ***	
H-TTGEN	general medical management	*follow-up, soins, consultation*
H-TTMED	treatment by medication	*aspirine, clamoxyl*
H-TTFREQ	frequency of medication	*bid*
H-TTMODE	mode of administration	*IM, IV*
H-TTCHIR	surgical procedures	*hysterectomy, cholécystectomie*
H-TTCOMP	complementary treatments	*bedrest, repos, physiothérapie*
	*** TIME AREA ***	
H-TMBEG	beginning	*onset, dévelope, apparition*
H-TMEND	termination	*discontinue, arrêt, stopper*
H-TMPER	duration	*persistant, constant*
H-TMREP	repetition	*habituelle, intermittent*
H-TMPREP	time preposition	*during, après, avant, depuis*
H-TMLOC	location in time	*recently, actuelle, déjà, post-op*
	*** RESULT AREA ***	
H-AMT	amount or degree	*much, totale, sévère, tout à fait*
H-BEH	behavior	*works, studies, travaille*
H-DIAG	diagnosis	*diabetes mellitus*
H-INDIC	disease indicator word	*fever, swelling, pain, thrombose*
H-NORMAL	non-problematical	*within normal limits, bon état, simple*
H-ORG	organism	*staph*
H-TXRES	test/exam result word	*positif*
H-RESP	patient response	*relief*
H-CHANGE	indication of change	*augmenté, diminution*
	*** EVIDENTIAL AREA ***	
H-NEG	negation of finding	*no, not, ne pas, jamais*
H-MODAL	uncertainty of finding	*evocatrice, probable, suspicion, semble*
	*** CONNECTIVE AREA ***	
H-BECONN	classifier verb	*is (a), est (un)*
H-CONN	P/V/ADJ/N connects two I-F's	*due to, secondaire à*
H-SHOW	V connects test and result	*shows, confirme, montre*

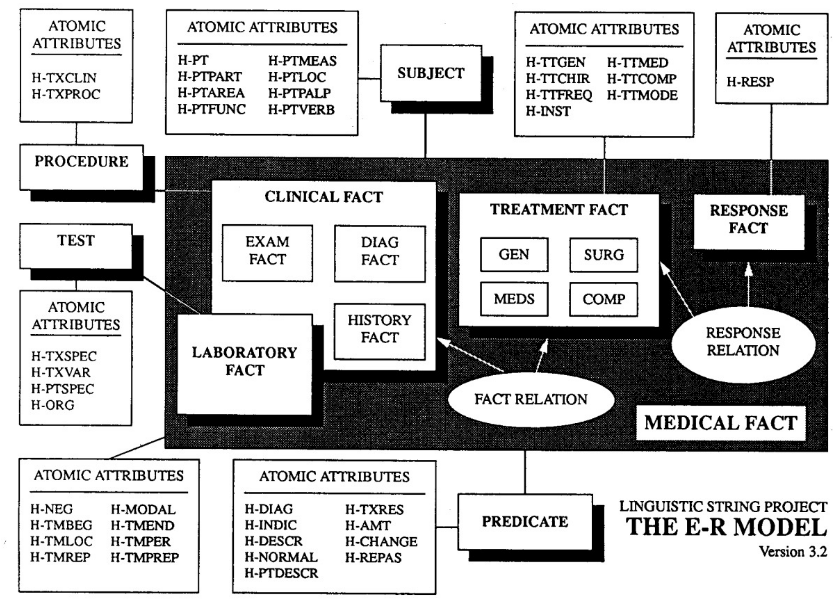


Figure 4 Schematic overview of the types of medical facts seen in patient documents and their associated lexical ("atomic") attributes. The CLINICAL FACT subtypes are distinguished by the paragraph they occur in: EXAM, DIAG, LAB, HISTORY. The TREATMENT FACT type is subdivided into general medical management (GEN), surgery (SURG), medications (MEDS), and all other therapies (COMPlementary). An instance of a TREATMENT FACT is often coupled to a RESPONSE FACT via a RESPONSE RELATION, e.g. *Much improved on penicillamine 750 mg daily*.

Sager et al, *Natural Language Processing and the Representation of Clinical Data,* 1993

Very very longstanding interest in clinical NLP:

(1) "the" and "anterior" modify incision
(2) "the" modifies "crease"
(3) "the" and "chest" modify "wall"
(4) "the", "previous", and "radical" modify "mastectomy"
(5) "made" is dependent upon the main verb "was"
(6) "along the crease" modifies "made" adverbially
(7) "against the chest wall" modifies "made" adverbially
(8) "following the previous radical mastectomy" modifies "made" adverbially

JAMA, March 17, 1969 • Vol 207, No 11

Information in Natural Languages: A New Approach, JAMA 1969

Industry interest

Google Cloud

Healthcare Natural Language API

What is Text Analytics for health in Azure Cognitive Service for Language?

Components

- Clinical entities
  - PHI, conditions, symptoms, interventions, dates
  - Often want to map to an ontology, e.g. UMLS
- Temporality
- Relation Extraction
  - Given two phrases, determine relationship between them, if any:
  - E.g., precedes, causes, treats, prevents, indicates, …
- Negation

Formalizing which components can be a whole project in itself

Concept Normalization

Words → multiple concepts
Concepts → many words

" Pt given carbo ia for her TNBC. Will dc. "

Pt	given	carbo	ia	for	her	TNBC.	Will	dc.
Patient		Carboplatin	Intra-arterial			Triple-neg. breast cancer		Discontinue

Concept Normalization

Words → multiple concepts
Concepts → many words

“ Pt given carbo ia for her TNBC. Will dc. ”

Pt	given	carbo	ia	for	her	TNBC.	Will	dc.
Patient (C0030705)		Carboplatin (C0079083)	Intra-arterial (C1561451)			Triple-neg. breast cancer (C3539878)		Discontinue (C1706472)

Identifying clinical entities

- Example: de-identification (you'll also see this in PS2). Identify *spans* and label as 'AGE', 'NAME', 'LOCATION', etc.

```
Mary Jane is 97 and lives in Brookline with her sister
```

- Desired output:
  "Mary Jane" (chars 0-8): NAME
  "97" (chars 13-14): AGE
  "Brookline" (chars 29-37): LOCATION

- By distinguishing among PHI types and finding the full span of each mention, we can replace with all mentions of "Mary Jane" with [***NAME (1)***] and tell it apart from other names and other types of PHI

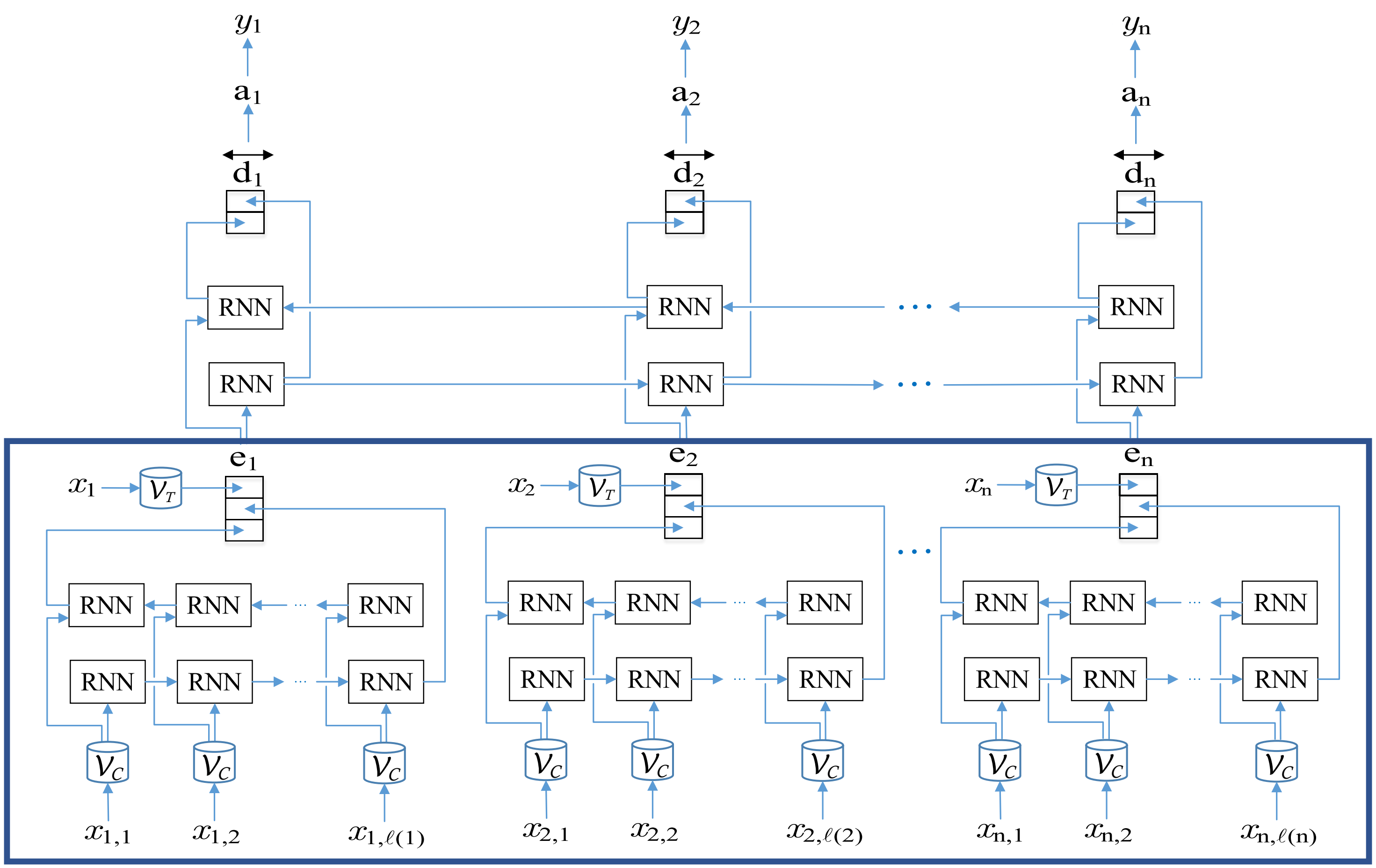
Identifying clinical entities: I/O/B tagging

- How can our labeling scheme distinguish between two adjacent words that refer to the same clinical entity (e.g. "abdominal pain") from two distinct mentions (e.g., "pain fever")?
- One way is to use the I/O/B ('inside', 'outside', 'beginning') tagging format

- Example:

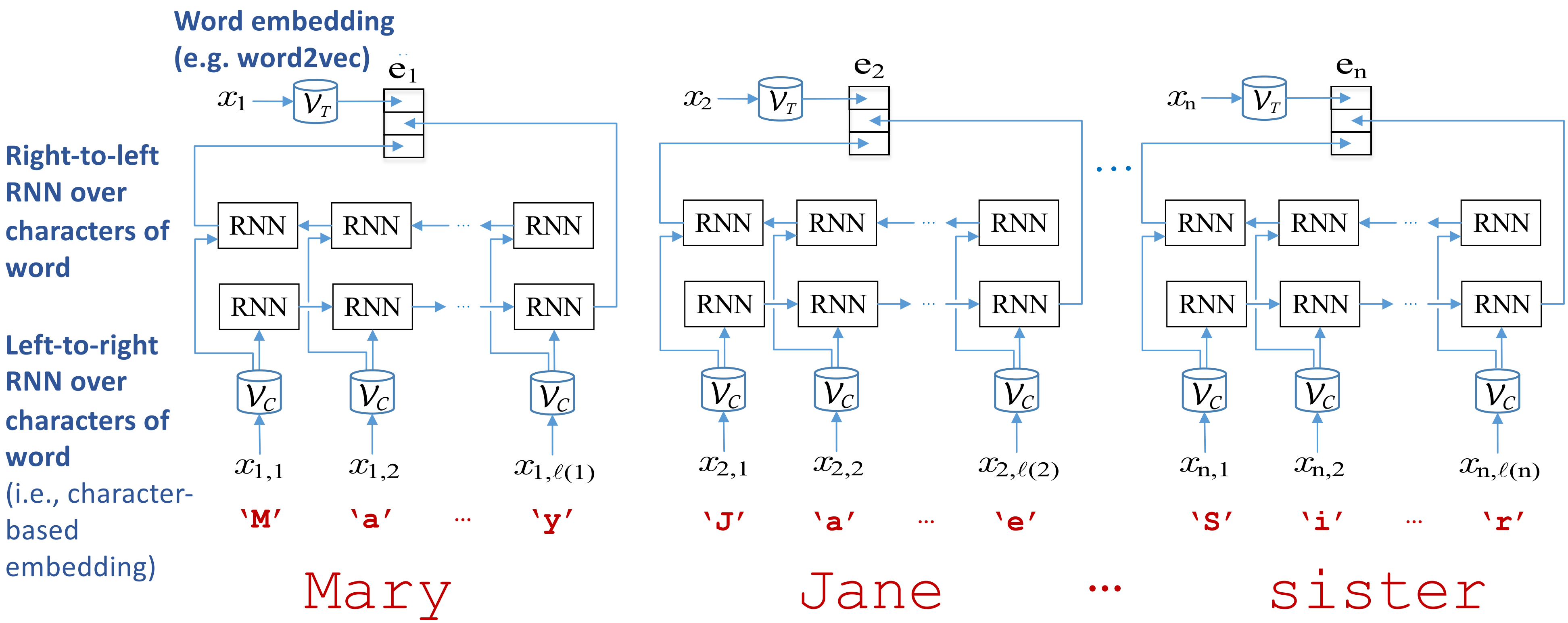
Mary	Jane	is	97	and	lives	in	Brookline	with	her	sister
B-NAME	I-NAME	O	B-AGE	O	O	O	B-LOCATION	O	O	O

De-identification using recurrent neural networks



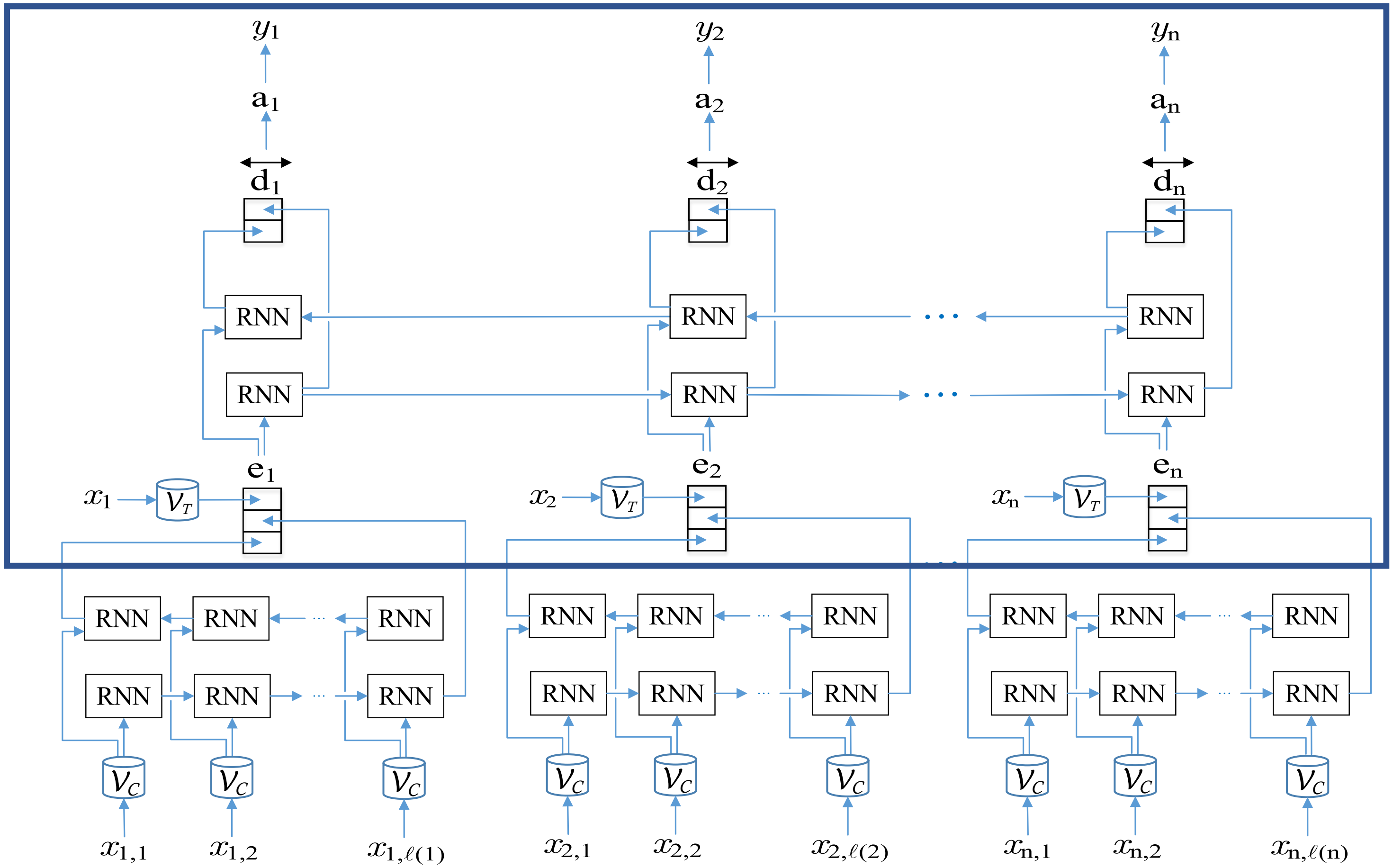
Dernoncourt, Lee, Uzuner, Szolovits. De-identification of Patient Notes with Recurrent Neural Networks. Journal of the American Medical Informatics Association, Volume 24, Issue 3, May 2017

De-identification using recurrent neural networks



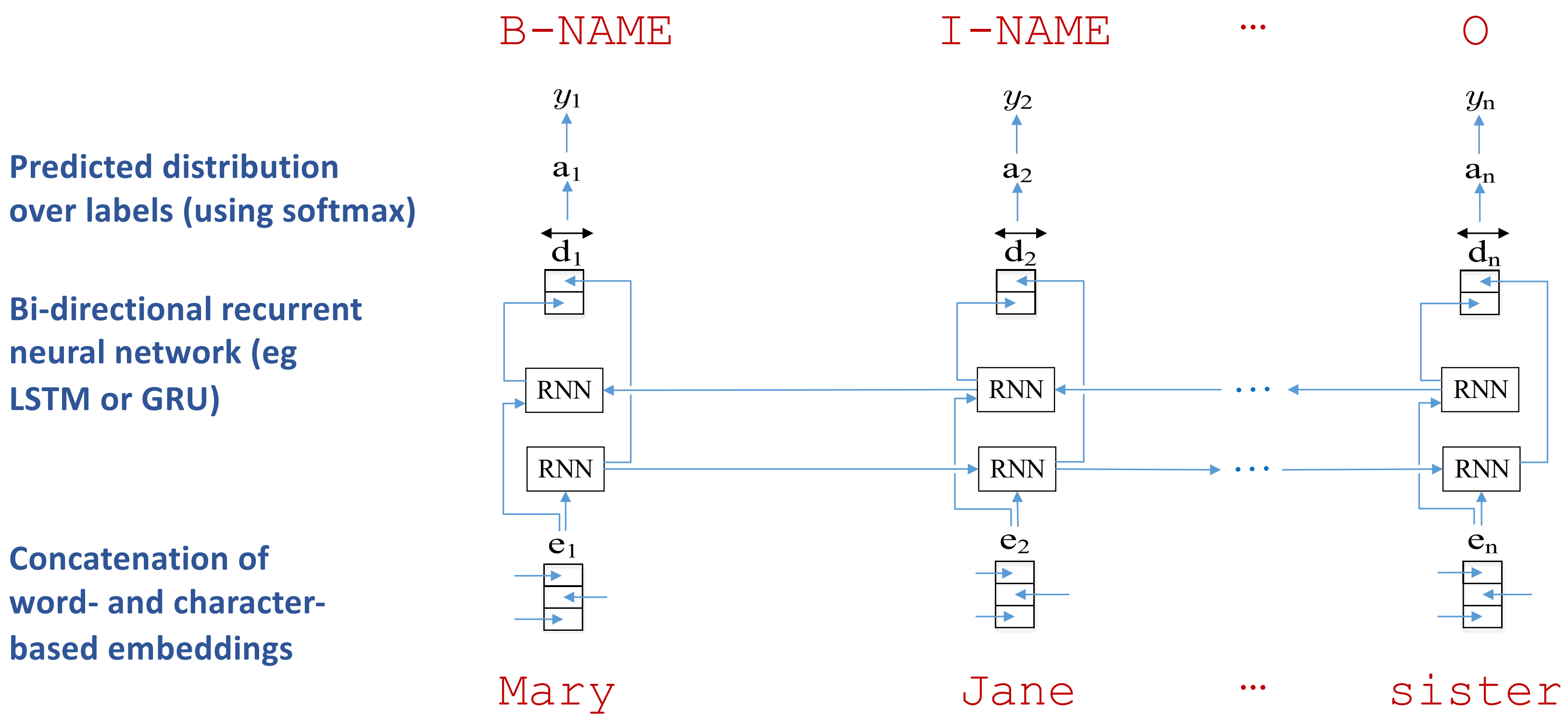
Lample, Ballesteros, Subramanian, Kawakami, Dyer. Neural Architectures for Named Entity Recognition. NAACL 2016.

De-identification using recurrent neural networks



Dernoncourt, Lee, Uzuner, Szolovits. De-identification of Patient Notes with Recurrent Neural Networks. Journal of the American Medical Informatics Association, Volume 24, Issue 3, May 2017

De-identification using recurrent neural networks



Dernoncourt, Lee, Uzuner, Szolovits. De-identification of Patient Notes with Recurrent Neural Networks. Journal of the American Medical Informatics Association, Volume 24, Issue 3, May 2017

More subtle clinical entity detection

- How could we do something more general, e.g. identify mentions that are split across multiple words or interweaved?
- We need a better token scheme.

B	First token of the mention
I	Other tokens of the mention
O	Everything else
OD	Within scope of a mention but not part of the mention itself
ID	Tokens which are part of a discontinuous mention
In	Identifying token in overlapping mentions
Bn	Identifying token in overlapping mentions, first word of the mention
Ip	Part of only one of two overlapping mentions, but not the identifying token

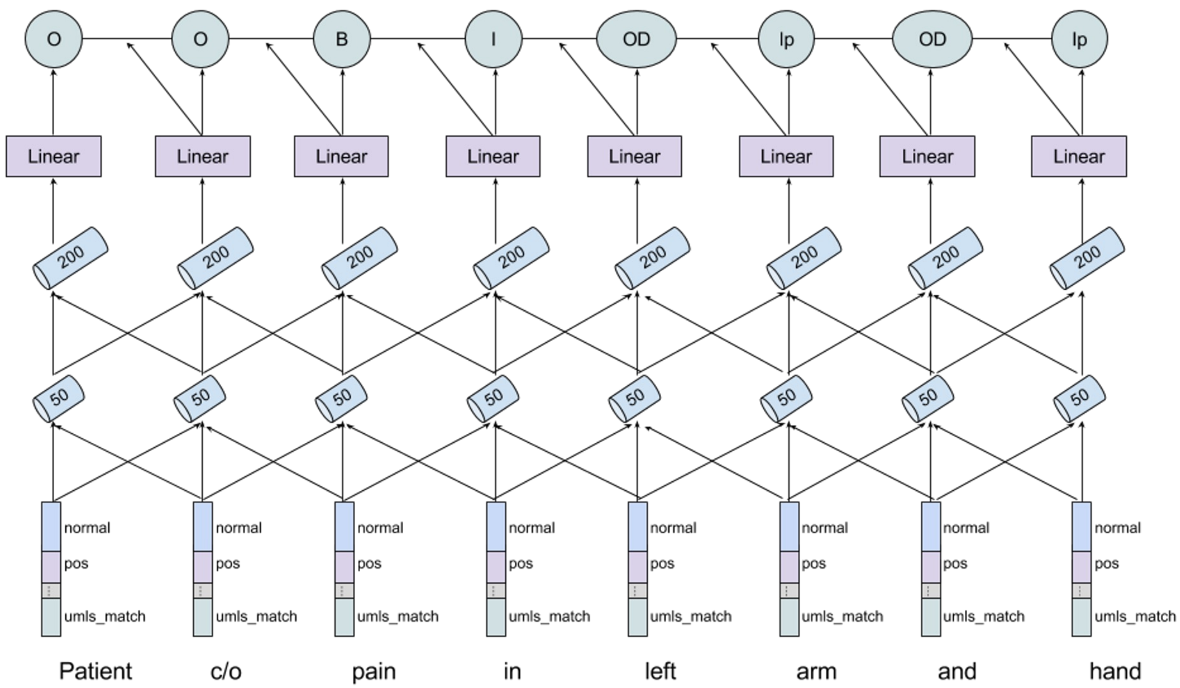
More subtle clinical entity detection

- Examples:

```
the patient suffers from a broken jaw .
 O     O       O       O  O   B     I  O

the pain is strongest in the arm .
 O    B  OD      OD     OD  OD   ID O

left arm and shoulder are swollen
 B    In  OD    In      OD    ID

elbow and wrist broken
 Bn   OD   Bn      ID

inflammation of left kidney and spleen
     B        OD  In    Ip     OD   In
```

Prediction with a (deep) conditional random field

Model the joint distribution over the output tokens using a conditional random field (CRF)



Predict the single-node and edge-potentials of the CRF. Then decode with dynamic programming

Two-layers of 1-D convolution with length-3 filters

Concatenation of bag-of-words representation and expert-derived features *(could alternatively use embeddings)*

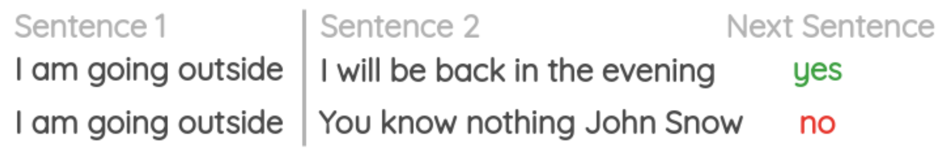
References:
McCallum & Li NAACL '03

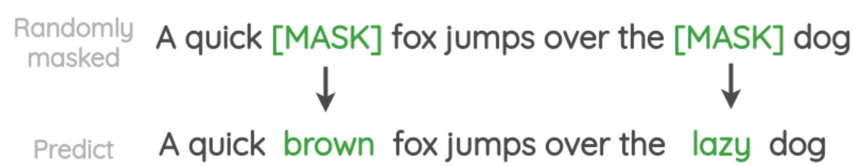
Strubell, Verga, Balanger, McCallum EMNLP '17

(Figure credit: Ankit Vani and Yacine Jernite, NYU)

Detour: self-supervision

Self-supervision: leveraging the underlying structure to learn representations useful for downstream tasks

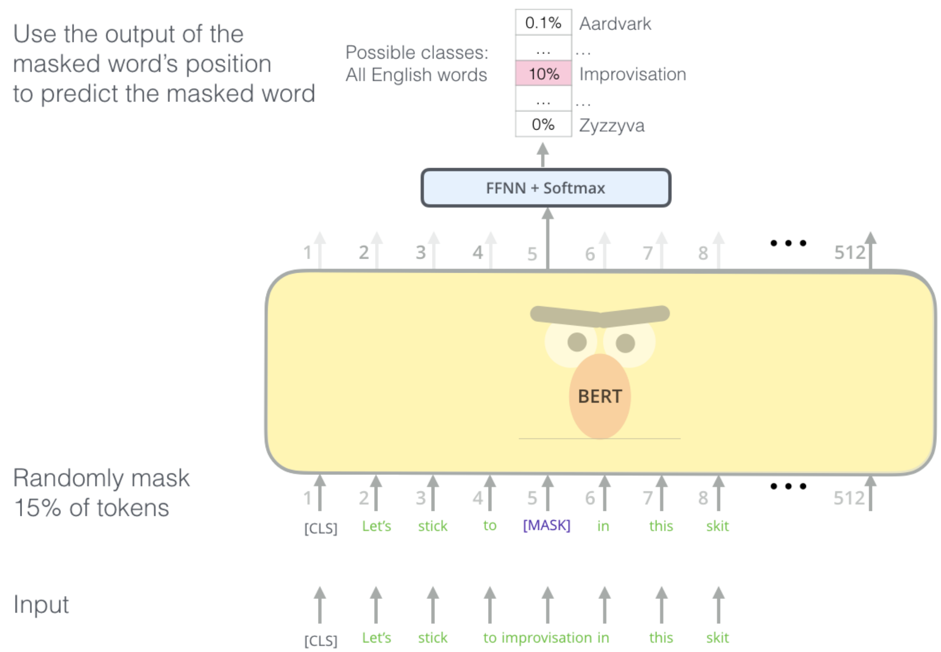




Text (Amit Chaudhary, blog)

Fine-tuning a BERT model

Earlier we showed how to use word embeddings as input.
Alternatively, we could predict using a pre-trained language model (e.g. BERT)



Devlin et al, *BERT*, NAACL '19

Figure reference: https://jalammar.github.io/illustrated-bert/

Fine-tuning a BERT model

- Earlier we showed how to use word embeddings as input. Alternatively, we could predict using a pre-trained language model (e.g. BERT)

[Devlin, Chang, Lee, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL '19]

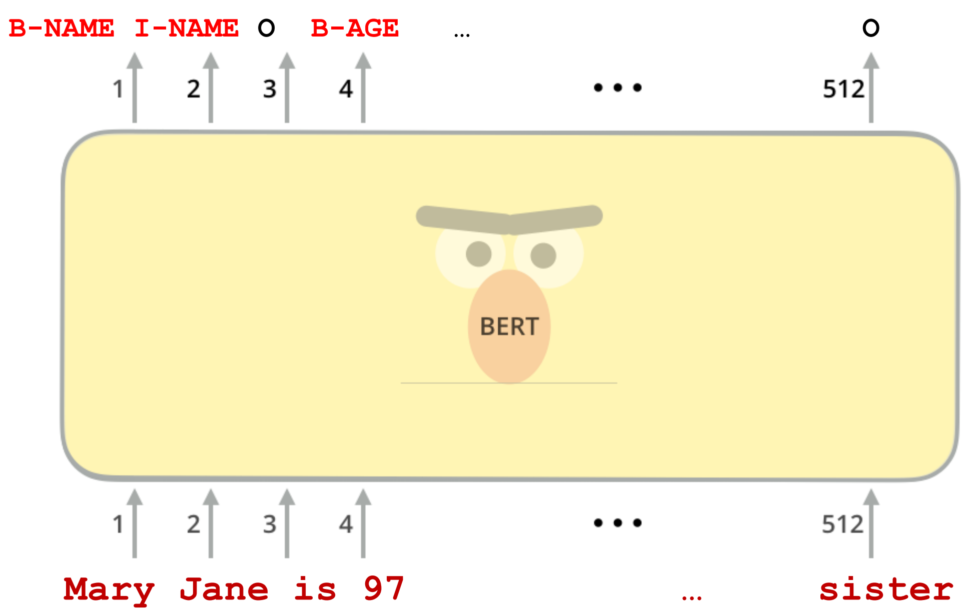


Figure reference: https://jalammar.github.io/illustrated-bert/

Simple methods can get you far.
Example: Negation Detection

NegEx Algorithm

1. Find all UMLS terms in each sentence of a discharge summary
    - “The patient denied experiencing chest pain on exertion” ⇒ “The patient denied experiencing *S1459038* on exertion”
2. Find patterns
    - <negation phrase> *{0,5} <UMLS term>
        - "no signs of", "ruled out unlikely", "absence of", "not demonstrated", "denies", "no sign of", "no evidence of", "no", "denied", "without", "negative for", "not", "doubt", "versus"
    - <UMLS term> *{0,5} <negation phrase>
        - “declined”, “unlikely”
    - Pseudo-negation: "gram negative", "no further", "not able to be", "not certain if", "not certain whether", not necessarily", "not rule out", "without any further", "without difficulty", "without further”

Chapman, et al. *A simple algorithm for identifying negated findings and diseases in discharge summaries.* J Biomed Inform. 2001

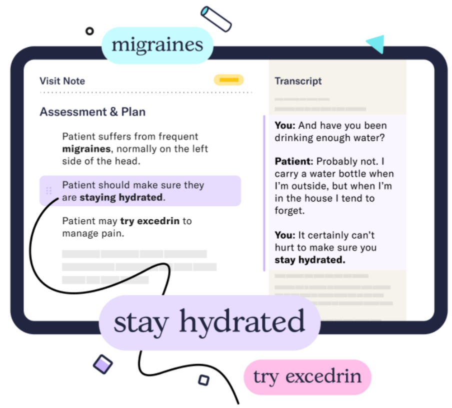
Outline

- Uses of clinical text
- Idiosyncrasies of clinical text
- Extraction of relevant variables
- Text generation

Generation

Automatic generation of text from other sources:

- From conversations
  - Abridge, EmpowerMD
- From imaging
  - Captioning literature

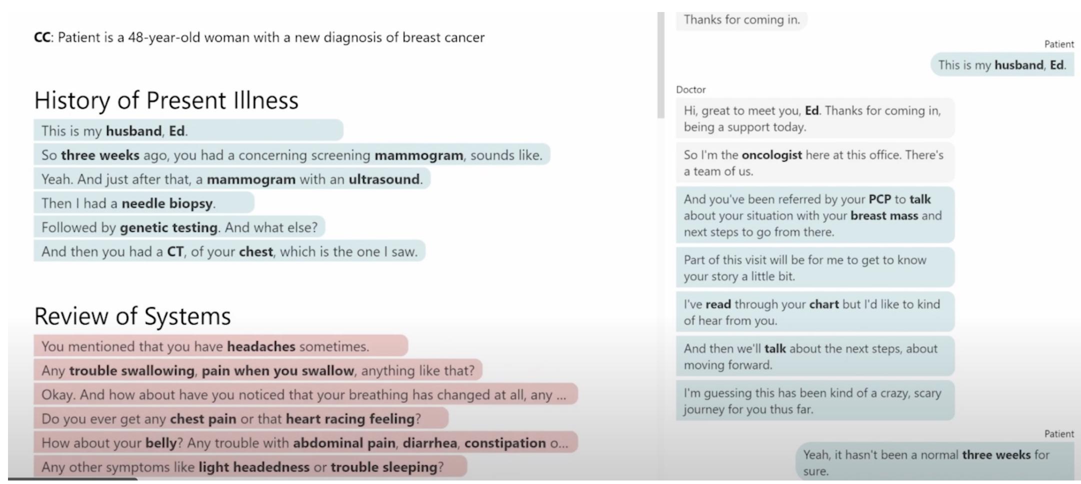


03

Receive a SOAP outline

Our algorithms quickly produce a draft of the documentation for your conversation.

Abridge



EmpowerMD, Microsoft

Generation

Automatic generation of text from other sources:

- From conversations
  - Abridge, EmpowerMD
- From imaging
  - Captioning literature

"We use categorical variables for location and character to generate new sentences …with general structure:

"There is a [degree of displacement], [+/- comminuted][+/- impacted] fracture of the [location] neck of femur [+/- with an avulsed fragment]."

Negative cases (i.e., those without fractures) had – *"No fracture was identified on this study"*.

Gale et al 2019, *Producing Radiologist-Quality Reports for Interpretable Deep Learning*

Summarization: Why?

- Information overload → delays, errors of omission
- Most chronically ill → complex most notes
  - E.g. for chronic kidney disease, 300+ notes on average over a decade plus, some have >4000
- Goal: Aid cognition, minimize superfluous text
  - For refamiliarization
    - Can take form of QA
  - For handoff
  - For discharge summaries

Pivoravov and Elhadad, *JAMIA 2015*

Summarization: How?

- *Extractive summaries:* selecting a subset of phrases from the original input text
- *Abstractive summaries:* generate novel synthesis of text

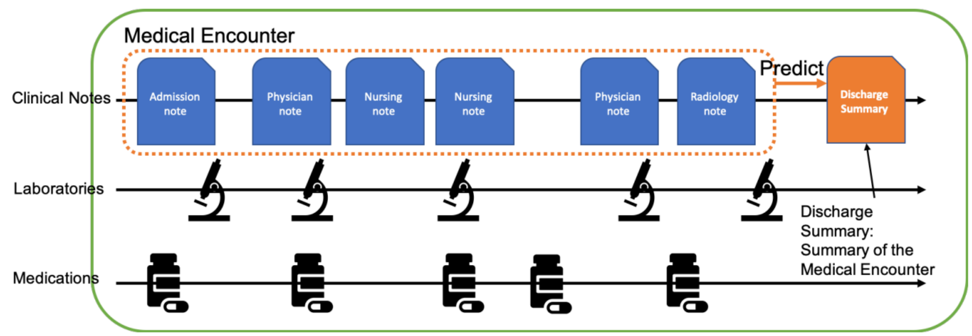
Pivoravov and Elhadad, *JAMIA 2015*

Summarization: To what end?

- *Indicative summaries:* point reader towards relevant knowledge, for use alongside the full patient record.

- *Informative summaries:* replacement of original text, to stand alone as shorter representation

Pivoravov and Elhadad, *JAMIA 2015*



Shing et al, *Towards Clinical Encounter Summarization*

Summarization: Evaluation

- Intrinsic: Measure the quality of the summary
  - Factual accuracy
  - Comparison to a gold standard summary
- Extrinsic: Measure the utility of the summary, for task of interest
  - Tasks like question answering, clinical trial matching, etc
  - Time-to-task completion
  - User satisfaction

Pivoravov and Elhadad, *JAMIA 2015*

Summarization Challenges

1. Accounting for similarity
2. Correct temporality
3. Missing data (interoperability)
4. Relevance: task-specific and disagreement
5. Incorporation of domain knowledge
6. Deployment and real-world evaluation

Pivoravov and Elhadad, *JAMIA 2015*

CLIP: An Extractive Approach

What does the PCP need to know from a hospital visit?

Action Type	Description	Example
Appointment	Appointments to be made by the PCP, or monitored to ensure the patient attends them.	The patient requires a neurology consult at XYZ for evaluation.
Lab	Laboratory tests that either have results pending or need to be ordered by the PCP.	We ask that the patients' family physician repeat these tests in 2 weeks to ensure resolution.
Procedure	Procedures that the PCP needs to either order, ensure another caregiver orders, or ensure the patient undergoes.	Please follow-up for EGD with GI.
Medication	Medications that the PCP either needs to ensure that the patient is taking correctly, e.g. time-limited medications or new medications that may need dose adjustment.	The patient was instructed to hold ASA and refrain from NSAIDs for 2 weeks.
Imaging	Imaging studies that either have results pending or need to be ordered by the PCP.	Superior segment of the left lower lobe: rounded density which could have been related to infection, but follow-up for resolution recommended to exclude possible malignancy
Patient Instructions	Post-discharge instructions that are directed to the patient, so the PCP can ensure the patient understands and performs them.	No driving until post-op visit and you are no longer taking pain medications.
Other	Other actionable information that is important to relay to the PCP but does not fall under existing aspects (e.g. the need to closely observe the patient's diet, or fax results to another provider).	Since the patient has been struggling to gain weight this past year, we will monitor his nutritional status and trend weights closely.

Mullenbach et al, ACL 2021

Question: What approach might you take?

CLIP

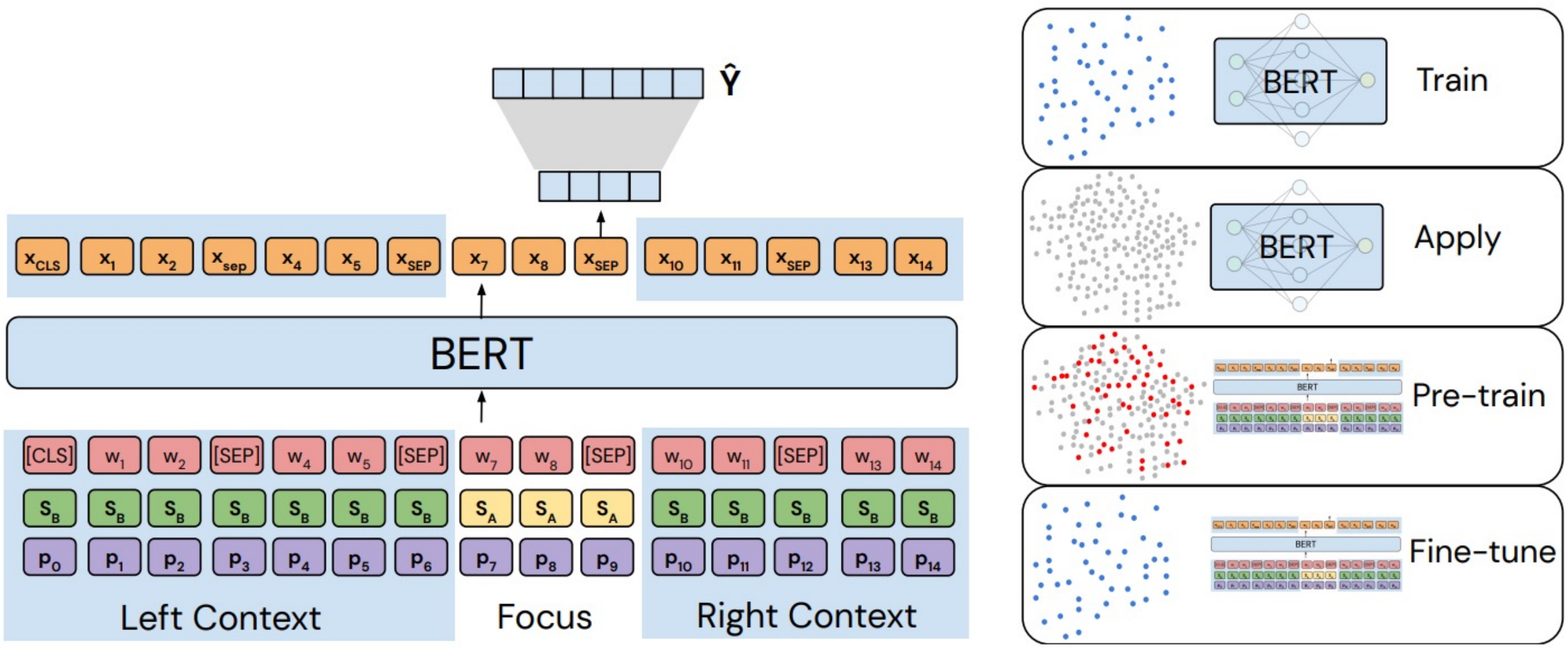
- CLInicalfollowuP
- Dataset on top of MIMIC-III
  - 718 labeled discharge summaries
  - 100k+ sentence-level annotations
  - 0.925 inter-rater reliability

"Given perfect performance, this would reduce the number of sentences a PCP may need to read by 88%. The best model's binary F1 is near 0.9, compared to the human benchmark of 0.93. "

https://github.com/asappresearch/clip

Mullenbach et al, ACL 2021

CLIP



(a) (b)

Mullenbach et al, ACL 2021

CLIP

Model	Patient	Appt	Medication	Lab	Procedure	Imaging	Other
Bag-of-words	0.741	0.792	0.546	0.625	0.302	0.343	0.236
CNN	0.759	0.824	0.595	0.629	0.315	0.431	0.228
BERT	0.780	0.855	0.635	0.719	0.415	0.474	0.275
MIMIC-DNote-BERT	0.783	0.854	0.656	0.741	0.524	0.567	0.294
MIMIC-DNote-BERT+Context	0.830	0.882	0.659	0.744	0.597	0.567	0.349
TTP-BERT+Context (250k)	0.841	0.887	0.668	0.745	0.548	0.566	0.365

Table 5: Average balanced F1 scores on the test set for each label across 10 runs.

Mullenbach et al, ACL 2021

Disclaimer: bag-of-words is a strong baseline

Table 3: ClinicalBERT accurately predicts 30-day readmission using discharge summaries. The mean and standard deviation of 5-fold cross validation is reported. ClinicalBERT outperforms the bag-of-words model, the BI-LSTM, and BERT deep language models.

Model	AUROC	AUPRC	RP80
ClinicalBERT	0.714 ± 0.018	0.701 ± 0.021	0.242 ± 0.111
Bag-of-words	0.684 ± 0.025	0.674 ± 0.027	0.217 ± 0.119
BI-LSTM	0.694 ± 0.025	0.686 ± 0.029	0.223 ± 0.103
BERT	0.692 ± 0.019	0.678 ± 0.016	0.172 ± 0.101

(Huang et al, CHIL 2020)

Table 2. Model AUC means and standard deviations over five data splits for IPV and injury prediction using radiology reports. Bold rows indicate best performance for task.

Model	IPV	Injury
Logistic Regression	0.841 ± 0.033	0.866 ± 0.016
Random Forest	**0.852 ± 0.022**	**0.887 ± 0.019**
Gradient Boosted Trees	0.842 ± 0.027	0.858 ± 0.030
Neural Network (Bag of Words)	0.849 ± 0.026	0.879 ± 0.010
Neural Network (clinicalBERT[42])	0.843 ± 0.022	0.852 ± 0.021

(Chen et al, PSB 2021)

Model	Accuracy	M-AUC	M-F1	m-AUC	m-F1	Precision-at-1
Input agnostic baseline	0.9189	0.5000	0.1414	0.7434	0.3109	0.2027
UMLS Medical Entity Matching	0.9122	0.8147	0.5121	0.8420	0.5833	0.5034
Logistic Regression	0.9417	0.8930	0.2510	0.9317	0.5004	0.6064
LinearSVC	0.9395	0.8959	0.2113	0.9354	0.4603	0.6199
Multinomial NaiveBayes	0.9269	0.7171	0.0615	0.8296	0.1938	0.4848
Random Forest	0.9212	0.8868	0.0155	0.8795	0.0541	0.5304
Gradient Boosting Classifier	0.9467	0.9181	0.5024	0.9447	0.6514	0.5861
BERT	0.9452	0.8953	0.4413	0.9365	0.6009	0.6199
CLINICALBERT (CBERT)	0.9476	0.9040	0.4573	0.9413	0.6029	0.6300

...

Table 2: Aggregate results for the medical diagnosis prediction task. AN: predicted noteworthy utterances, DN: utterances predicted to be noteworthy specifically concerning a summary passage discussing diagnoses, F2K: UMLS-extracted noteworthy utterances with added top predicted AN/DN utterances to get K total utterances, M-: macro average, m-: micro average

(Krishna et al, AAAI WS 2020)

Question: How would you automatically measure the quality of a summary?

Radiology: An Abstractive Approach

Common Neural abstractive summarization models

Goal: generate summaries which have high overlap with human references

Metrics:
ROUGE – check for overlap via 'n-grams', 'longest-common-subsequences', 'weighted-subsequences', 'skip-bigrams'

Background: radiographic examination of the chest. clinical history: 80 years of age, male ...

Findings: frontal radiograph of the chest demonstrates repositioning of the right atrial lead possibly into the ivc. ... a right apical pneumothorax can be seen from the image. moderate right and small left pleural effusions continue. no pulmonary edema is observed. heart size is upper limits of normal.

Human Summary: pneumothorax is seen. bilateral pleural effusions continue.

Summary A (ROUGE-L = 0.77):
no pneumothorax is observed. bilateral pleural effusions continue.

Summary B (ROUGE-L = 0.44):
pneumothorax is observed on radiograph. bilateral pleural effusions continue to be seen.

Figure 1: A (truncated) radiology report and summaries with their ROUGE-L scores. Compared to the human summary, Summary A has high textual overlap (i.e., ROUGE-L) but makes a factual error; Summary B has a lower ROUGE-L score but is factually correct.

Zhang et al, Optimizing the Factual Correctness of a Summary, ACL 2020

Radiology: An Abstractive Approach

x

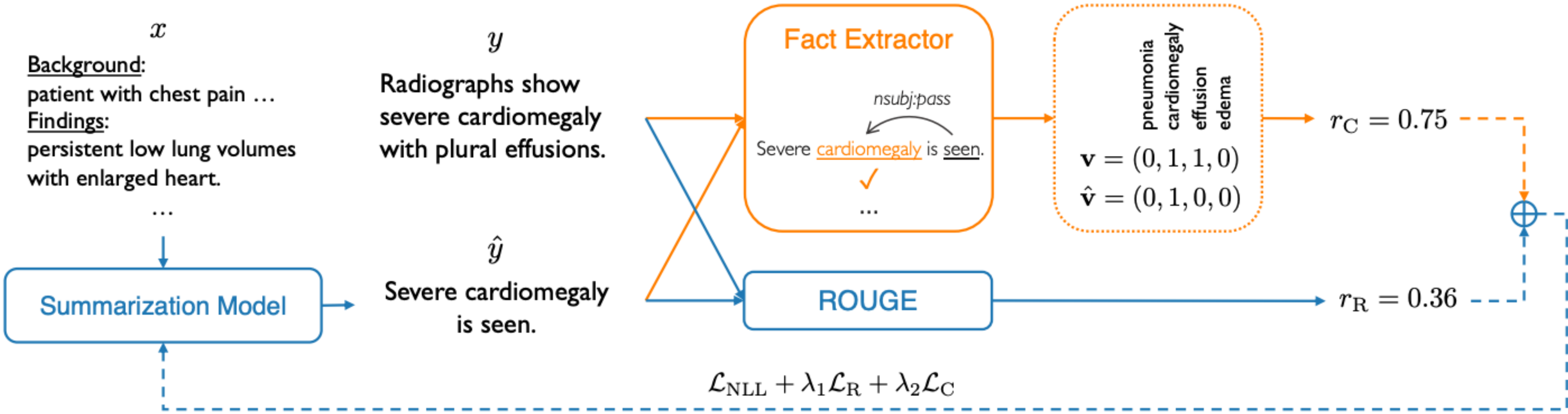
<u>Background</u>:
patient with chest pain …
<u>Findings</u>:
persistent low lung volumes
with enlarged heart.
…

y

Radiographs show
severe cardiomegaly
with plural effusions.

Zhang et al, ACL 2020

Radiology: An Abstractive Approach



Trained via reinforcement learning

Zhang et al, ACL 2020

Radiology: An Abstractive Approach

Variable	PG Baseline	RL_{R+C}	Δ
No Finding	77.3	81.5	+4.2*
Cardiomegaly	29.5	40.4	+10.9*
Airspace Opacity	64.6	74.9	+10.3*
Edema	58.4	70.9	+12.5*
Consolidation	46.3	53.2	+6.9*
Pneumonia	46.7	46.8	+0.2
Atelectasis	48.8	56.3	+7.5*
Pneumothorax	69.5	82.9	+13.4*
Pleural Effusion	62.0	73.4	+11.4*
Macro Avg.	55.9	64.5	+8.6*

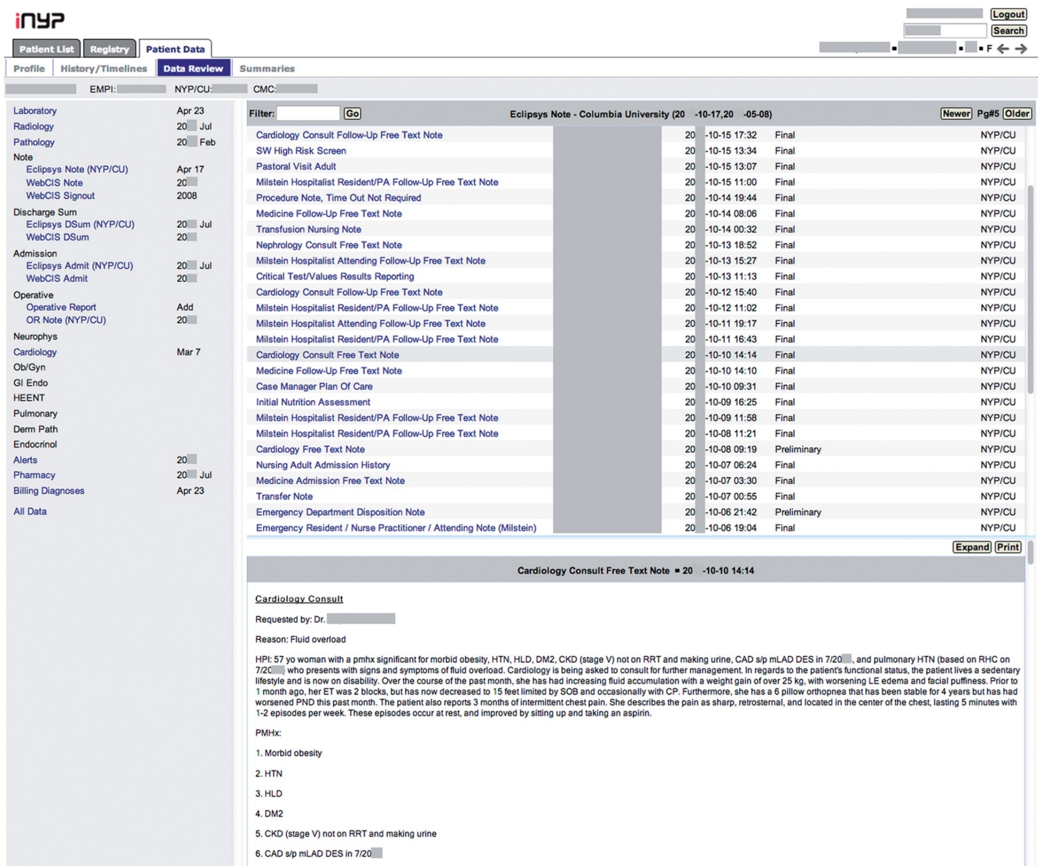
Table 3: Test set factual F_1 scores for each variable on the Stanford dataset. $*$ marks statistically significant improvements with $p < .01$ under a bootstrap test.

Metric	Win	Tie	Lose
Our Model vs. PG Baseline			
Fluency	7%	60%	33%
Factual Correctness	31%	55%	14%
Overall Quality	48%	24%	28%
Our Model vs. Human Reference			
Fluency	17%	54%	29%
Factual Correctness	23%	49%	28%
Overall Quality	44%	17%	39%

Table 4: Results of the radiologist evaluation. The top three rows present results when comparing our RL_{R+C} model output versus the baseline model output; the bottom three rows present results when comparing our model output versus the human-written summaries.

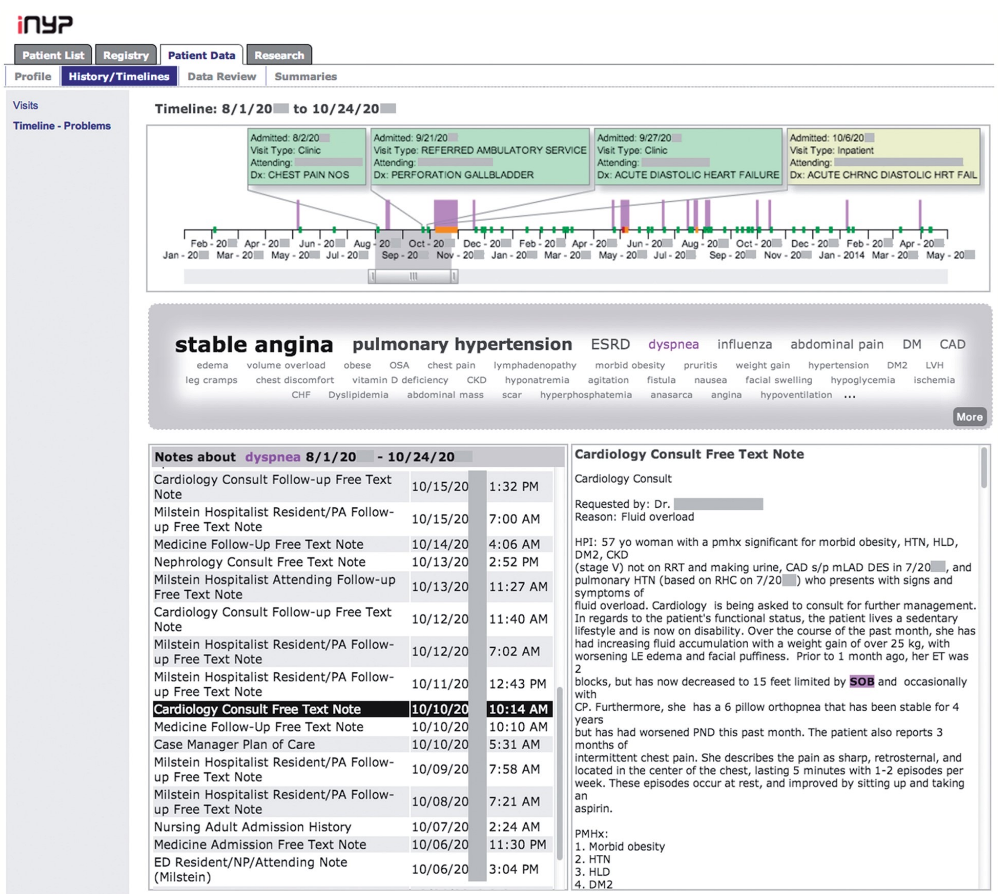
Zhang et al, ACL 2020

HARVEST: a visualization approach



Hirsch et al, JAMIA 2015

HARVEST: a visualization approach



Term identification looking at UMLS disorders

Salience determined by tf-idf weighting

Hirsch et al, JAMIA 2015

HARVEST: evaluation

20 min-scenarios where participants answered a 35-item questionnaire

- **Date finding:** *How soon after discharge did the patient have follow up?*
- **Clinical fact or event finding:** *Does this patient have a history of X?*
- **Clinical comparisons:** *What accounted for the change in X over Y period?*
- **Clinical synthesis:** What were the five most prominent problems over Y period?

Table 1: Electronic health record representation of the 12 study patients

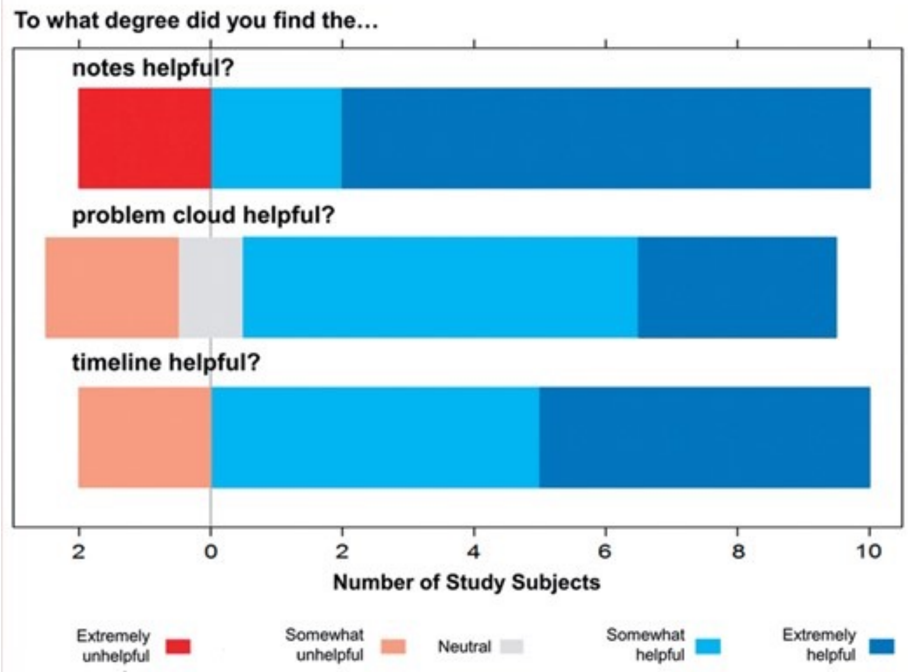
	Mean	Median	Range
Total clinical follow-up in the hospital system (months)	48 .5	57 .5	9 –62
In 2 years preceding study, number of:			
Healthcare visits	65 .7	57 .5	20 –133
Provider types	12 .75	11 .5	4 –20
Documents	266 .25	226 .5	75 –874
Document types	62 .25	56 .5	12 –137
Document authors	111	90 .5	22 –379
Problem concepts	5129 .5	3877	1001 –16 502
Unique problem concepts	253 .75	271	94 –462

Hirsch et al, JAMIA 2015

HARVEST: evaluation

20 min-scenarios where participants answered a 35-item questionnaire

- **Date finding:** *How soon after discharge did the patient have follow up?*
- **Clinical fact or event finding:** *Does this patient have a history of X?*
- **Clinical comparisons:** *What accounted for the change in X over Y period?*
- **Clinical synthesis:** What were the five most prominent problems over Y period?



Hirsch et al, JAMIA 2015

Summary

- Clinical text contains a lot of important information found nowhere else in the medical record
- Clinical text can be leveraged to impact patient outcomes, both at the point-of-care and via retrospective research
- These problems are challenging and of longstanding interest
- Now is a ripe time to solve them!