In the United States there are on the order of 2,000,000,000 doctor-patient encounters per year; that's over 200,000 an hour.

“The Electronic Health Record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports...”

WHAT'S IN AN EHR?

- Patient demographic data
- Procedure/Treatment/Diagnosis Codes
- Vital signs
- Radiology reports
- Test results
- Clinical text notes

April 14, 2007

CHIEF COMPLAINT: Shortness of breath.

HISTORY OF PRESENT ILLNESS: This 68-year-old female presents to the emergency department with shortness of breath that has gone on for 4-5 days, progressively getting worse. It comes on with any kind of activity whatsoever. She has had a nonproductive cough. She has not had any chest pain. She has had chills but no fever.

EMERGENCY DEPARTMENT COURSE: The patient was admitted. She has had intermittent episodes of severe dyspnea. Lungs were clear. These would mildly respond to breathing treatments and morphine. Her D-dimer was positive. We cannot scan her chest; therefore, a nuclear V/Q scan has been ordered. However, after consultation with Dr. C, it is felt that she is potentially too unstable to go for this. Given the positive D-dimer and her severe dyspnea, we have waved the risks and benefits of anticoagulation with her heme-positive stools. She states that she has been constipated lately and doing a lot of straining. Given the possibility of a PE, it was felt like anticoagulation was very important at this time period; therefore, she was anticoagulated. The patient will be admitted to the hospital under Dr. C.
WHY CLINICAL NOTES MATTER?

SECONDARY USE

If we have access to thousands / millions of medical records representing millions of patients what can we learn?

SECONDARY USE

- Make research advances in medicine
- Discover negative medication interactions
- Identify a cohort of patients for a research study
- Ensure patient safety procedures

ACCESS CLINICAL TEXT

- Most EHR data is structured
- Relatively easy to export and use
- Clinical text is unstructured?
- How do we use it?
- Natural Language Processing
CLINICAL NATURAL LANGUAGE PROCESSING

• The task of applying NLP tools to clinical free text

WHY CLINICAL NLP?

• News
• Social media
• Conversations
• Blogs
• Books

CLINICAL NLP CHALLENGES

• Complex clinical terminology
  • mitral regurgitation, Left ventricular systolic dysfunction, Subarachnoid hemorrhage
• Numerous medical domains
  • Radiology vs. pediatrics vs. oncology vs. …
• Institutional specialization
  • Johns Hopkins Medicine vs. Cheyenne Regional Medical Center
• Differ in patients, hospital policies and standards
• EHR clinical text storage
  • Different EHR systems/deployments store text data differently

CLINICAL NLP CHALLENGES

• Mixed data types
  • Clinical text notes contain tables, lists, bullets, full paragraphs
• Varying clinical note types
  • Discharge notes: long, boiler plate summaries
  • ER notes: detailed descriptions, wide variety of issues
  • Progress notes: very brief updates (a sentence of less)
  • Radiology reports: descriptive analysis of imagery
• Records in context
  • Clinical record needs to be taken in context of structured data and previous notes
DATA CHALLENGES

- English Wikipedia: 3.7 billion words
- Common Crawl: billions of pages
- Social media: 500 million tweets per day
- Europarl: tens of millions of words

DATA CHALLENGES

- Available Clinical data corpora
- MIMIC 3: 1 ICU million notes
- Data restricted due to HIPAA concerns
- HIPAA: US law that protects medical records

COMMON CLINICAL NLP TASKS

XXX | XXX | XXX | | XXX | 12/01/1998 12:00:00 AM | INCARCERATED UMBILICAL HERNIA | Signed | DIS |
Admission Date: 11/7/1998 Report Status: Signed
Discharge Date: 1/25/1999
PRINCIPAL DIAGNOSIS: INCARCERATED UMBILICAL HERNIA.

HISTORY: Jane Doe is a 76 year old woman with a complex past medical history including coronary artery disease with a history of MIs times two in the past, a history of DVT back in 1974, hypertension, rheumatoid arthritis, gout and history of atrial fibrillation and atrial flutter as well as onset adult diabetes mellitus. She presented to the XXX Community Hospital on the day of admission complaining of an umbilical bulge over the past several weeks. This umbilical bulge had been increasing somewhat in size, but had not bothered her and was always reducible. However, over the preceding weekend it became incarcerated and then became somewhat painful. It was not associated with any nausea or vomiting and she reported that she was having normal bowel movements even in the face of this problem. She presented initially to the XXX County Health Center and was admitted with the diagnosis of incarcerated umbilical hernia.


MEDICATIONS ON ADMISSION: Lasix 80 mg a day , sublingual nitroglycerin p.r.n. , Propafenone 225 mg t.i.d. , Lopressor 150 mg t.b.i.d. , Lisinopril 10 mg a day and Micronase 10 mg b.i.d. , Isordil 40 mg t.i.d. , Coumadin 5 mg a day with 2-1/2 mg every Sunday.

ALLERGIES: She is allergic to aspirin and penicillin.

PHYSICAL EXAMINATION: She is an extremely pleasant elderly woman in no acute distress. HEENT - showed extracranial movements intact. Pupils equally round and reactive to light. NECK - supple. HEART - regular rhythm. LUNGS - clear. ABDOMEN - soft, nontender, nondistended with approximately 1.5 cm in diameter umbilical hernia to the left of her umbilicus. This hernia was somewhat tender to palpation, but showed no overlying erythema or evidence of necrosis. She had normal bowel sounds.
SEGMENTATION
- Section boundary detection
  - Chief complaint, medical history, assessment and plan, tables, lists
- Sentence boundary detection
  - Identify coherent sentences
    - Many sentence fragments and notes
- Tokenization
  - Complex because of abbreviations, technical terms
  - 75 lbs. P.O. P.O. x-ray U/ml left/right
- Abbreviations

NEGATION
- When is a concept/statement negated?
  - Additionally, there was no evidence of extension of his infected pseudocyst into the psoas muscle.
  - There is no significant interval change in the 2 large pancreatic pseudocysts.
  - Acute pancreatitis with pseudocyst, with no obvious complications of the pseudocyst at this point in time.

SYNTAX
- Part of speech tagging
- Chunking
- Parsing

DE-IDENTIFICATION
- Use and sharing of clinical notes requires de-identification (HIPAA)
  - Removal of information that could identify the patient
  - “Sarah reported feeling dizzy towards the end of the school day; left teaching to come to ER.”
NAMED ENTITY RECOGNITION

- Identify mentions of named entities in text
  - Medications
  - Symptoms
  - Comorbidities
  - Treatments

SUMMARIZATION

- Produce summaries of clinical notes
  - Single note summarization
  - Cross-note summarization
  - Integrating summaries with structured data

PATIENT SIMILARITY METRICS

- Find other patients like this one
- Applications: precision medicine

AUTOMATED CODING

- Medical data is hand coded
  - ICD 10: diseases, symptoms, abnormal findings, complaints, social circumstances, …
  - Creates structured data from free clinical text

<table>
<thead>
<tr>
<th>Table 2: Revised Wound Care-Related ICD-10 Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1: Disorders of Metabolism &amp; Lipids</td>
</tr>
<tr>
<td>832.219 Cerebral infarction due to unspecified occlusion or stenosis of unspecified vertebral artery</td>
</tr>
<tr>
<td>832.239 Cerebral infarction due to unspecified occlusion or stenosis of unspecified carotid artery</td>
</tr>
<tr>
<td>832.333 Cerebral infarction due to thrombosis of bilateral posterior cerebral arteries</td>
</tr>
<tr>
<td>832.343 Cerebral infarction due to thrombosis of bilateral cerebellar arteries</td>
</tr>
<tr>
<td>L98.495 Non-pressure chronic ulcer of other sites, with muscle involvement, without evidence of necrosis</td>
</tr>
<tr>
<td>L98.496 Non-pressure chronic ulcer of other sites, with bone involvement, without evidence of necrosis</td>
</tr>
<tr>
<td>L98.498 Non-pressure chronic ulcer of other sites, with other specified severity</td>
</tr>
<tr>
<td>M99.021 Other chronic osteomyelitis, right humerus</td>
</tr>
</tbody>
</table>
SENTENCE SIMILARITY

• Are two sentences saying the same thing?
  • No clinically relevant adverse events, such as urinary retention, respiratory disturbances, or wound infections were reported in the M-ADL group.
  • Neither intraoperative nor postoperative clinically relevant adverse events, such as urinary retention, respiratory disturbances, or wound infections, were observed.

CONCEPT LINKING

The patient underwent a CT scan in April which did not reveal lesions in his liver.

PHENOTYPING

• Extract patient conditions from free clinical text:
  387055    | 26563    | 18146    | 2009-01-26 15:08:00.0    | CHEST (PORTABLE AP)    | C12 CHEST (PORTABLE AP)    | Clip # 282-0776  Actual report    | DATE: [**2009-01-26**] 3:08 PM    | CHEST (PORTABLE AP)    | **Clip Number (Radiology) 7881** | Reason: r/o ptx, s/p aicd removal

UNDERLYING MEDICAL CONDITION:
57 year old man with pleural effusion.
REASON FOR THIS EXAMINATION:
r/o ptx, s/p aicd removal.

FINAL REPORT

INDICATION: 57 y/o male with pleural effusion, rule out pneumothorax, status post AICD removal.

FINDINGS: A single portable chest radiograph is compared with a portable study done earlier today. The right subclavian line is redemonstrated and in appropriate position. There is no evidence for pneumothorax. The AICD has been removed in the interim. The heart is mildly enlarged when adjusting for technique. There is evidence for mild pulmonary vascular redistribution. There has been interval increase in the size of the right pleural effusion. The patient is status post sternotomy. The patient has no prior history of smoking.

IMPRESSION: 1) Status post AICD removal without evidence for pneumothorax.
2) Findings consistent with worsening CHF.
3) Interval increase in the size of the right pleural effusion.
METHODS

ONTOLOGIES / DICTIONARIES

- Medicine has extensive ontologies
- Structured knowledge repositories
- List of synonyms, relations, definitions, etc.
- Many clinical NLP incorporate these resources
- Not typically available in other domains

RULES!

- Extensive use of rule based methods in clinical NLP
- Easier to implement, deploy and understand
- Medicine requires domain experts, easier for them to understand and create rules
- Statistical methods require training data: often not present in medicine

STATISTICAL METHODS

- Standard statistical NLP methods
- These work well (often better than rules) but face some challenges:
  - Robustness
  - Training data
  - Interpretability
  - Ease of implementation by domain experts
Abstract
Contextual word embedding models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) have dramatically improved performance for many natural language processing (NLP) tasks in recent months. However, these models have been minimally explored on specialty corpora, such as clinical text; moreover, in the clinical domain, no publicly-available pre-trained BERT models yet exist. In this work, we address this need by exploring and releasing BERT models for clinical text: one for generic clinical text and another for discharge summaries specifically. We demonstrate that using a domain-specific model yields performance improvements on three common clinical NLP tasks as compared to nonspecific embeddings. These domain-specific models are not as performant on two clinical de-identification tasks, and argue that this is a natural consequence of the differences between de-identified source text and synthetically non de-identified task text.

1 Introduction
Natural language processing (NLP) has been shaken in recent months with the dramatic successes enabled by transfer learning and contextual word embedding models, such as ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), and BERT (Devlin et al., 2018). These models have been primarily explored for general domain text, and, recently, biomedical text with BioBERT (Lee et al., 2019). However, clinical narratives (e.g., physician notes) have known differences in linguistic characteristics from both general text and non-clinical biomedical text, motivating the need for specialized clinical BERT models.

In this work, we build and publicly release exactly such an embedding model. Furthermore, we demonstrate on several clinical NLP tasks the improvements this system offers over traditional BERT and BioBERT alike.

In particular, we make the following contributions:
1. We train and publicly release BERT-Base and BioBERT-finetuned models trained on both all clinical notes and only discharge summaries.
2. We demonstrate that using clinical specific contextual embeddings improves both upon general domain results and BioBERT results across 2 well established clinical NER tasks and one medical natural language inference task (i2b2 2010 (Uzuner et al., 2011), i2b2 2012 (Sun et al., 2013a,b), and MedNLI (Romanov and Shivade, 2018)). On 2 de-identification (de-ID) tasks, i2b2 2006 (Uzuner et al., 2007) and i2b2 2014 (Stubbs et al., 2015; Stubbs and Uzuner, 2015), general BERT and BioBERT outperform clinical BERT and we argue that fundamental facets of the de-ID context motivate this lack of performance.
**UNDERLYING MEDICAL CONDITION:**
57 year old man with pleural effusion.

**REASON FOR THIS EXAMINATION:**
r/o ptx, s/p aicd removal

**FINAL REPORT**

**INDICATION:** 57 y/o male with pleural effusion, rule out pneumothorax, status post AICD removal.

**FINDINGS:** A single portable chest radiograph is compared with a portable study done earlier today. The right subclavian line is redemonstrated and in appropriate position. There is no evidence for pneumothorax. The AICD has been removed in the interim.

The heart is mildly enlarged when adjusting for technique. There is evidence for mild pulmonary vascular redistribution. There has been interval increase in the size of the right pleural effusion. The patient is status post sternotomy.

**IMPRESSION:**
1) Status post AICD removal without evidence for pneumothorax.
2) Findings consistent with worsening CHF.
3) Interval increase in the size of the right pleural effusion.
UNDERLYING MEDICAL CONDITION:
57 year old man with pleural effusion.

REASON FOR THIS EXAMINATION:
r/o ptx, s/p aicd removal

FINAL REPORT
INDICATION: 57 y/o male with pleural effusion, rule out pneumothorax, status post AICD removal.

FINDINGS: A single portable chest radiograph is compared with a portable study done earlier today. The right subclavian line is redemonstrated and in appropriate position. There is no evidence for pneumothorax. The AICD has been removed in the interim.

The heart is mildly enlarged when adjusting for technique. There is evidence for pulmonary vascular redistribution. There has been interval increase in the size of the right pleural effusion. The patient is status post sternotomy.

IMPRESSION: 1) Status post AICD removal without evidence for pneumothorax.
2) Findings consistent with worsening CHF.
3) Interval increase in the size of the right pleural effusion.

Unified Medical Language System - UMLS

CUI: C0018802
Preferred Name: Congestive heart failure
Semantic Type: Disease or Syndrome
Definition: Heart failure accompanied by EDEMA, such as swelling of the legs and ankles and congestion in the lungs.
Alternative Names: Congestive heart disease, cardiac failure congestive, ccf, chf

CUI: C0155582
Name: Congestive rheumatic heart failure
Neural Ranker

We treat our concept linking task as a learning-to-rank task:
Neural Ranker

\[ L(\theta) = \max\{0, e^{-S(m, c_1)}; \theta \} - \max\{S(m, c_2); \theta \} \ldots S(m, c_n); \theta) \} \]

Dataset - N2C2 Task 3

- 50 training clinical notes with ~6,700 training (mention, concept) annotations.
- Concept labels consisted of any CUI among ~430,000 in 2017AB SNOMED, RxNorm subsets.

Lexical Rules

- First, we apply two lookup procedures:
  - If a mention was previously seen in the training set, we assign it the same concept
  - If a mention name matches exactly to one entry in the UMLS, we assign it that concept
- These rules work well for this dataset due to annotation guidelines
  - Annotators were requested to try to break apart mentions to be easily matched
  - If none of these rules match to a single concept, we use the neural mention ranker to identify the best concept.
### Results

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenge Rules</td>
<td>0.77%</td>
<td>0.77%</td>
</tr>
<tr>
<td>Train Lookup</td>
<td>51.83%</td>
<td>53.69%</td>
</tr>
<tr>
<td>UMLS Lookup</td>
<td>71.23%</td>
<td>79.22%</td>
</tr>
<tr>
<td>MetaMap</td>
<td>73.43%</td>
<td>83.57%</td>
</tr>
<tr>
<td>Linker</td>
<td>76.75%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

### Outline

1. Introduction to Clinical NLP
2. Concept Linking
3. Clinical Semantic Textual Similarity
4. Phenotyping

### N2C2 2019 STS Data

<table>
<thead>
<tr>
<th>Procedure Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 hours as needed.</td>
<td>5</td>
</tr>
<tr>
<td>Cardiovascular assessment findings include heart rate normal, Heart rhythm, atrial fibrillation with controlled ventricular response.</td>
<td>3</td>
</tr>
<tr>
<td>The risks and benefits of the procedure were discussed, and the patient consented to this procedure</td>
<td>1</td>
</tr>
<tr>
<td>Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 hours as needed #1 each.</td>
<td>5</td>
</tr>
<tr>
<td>Cardiovascular assessment findings include heart rate, bradycardic, Heart rhythm, first degree AV Block.</td>
<td>3</td>
</tr>
<tr>
<td>The content of this note has been reproduced, signed by an authorized</td>
<td>1</td>
</tr>
</tbody>
</table>
Clinical STS

- Task: Given a pair of sentences from a clinical note, predict the degree of content overlap
- Current well-known solutions
  - Include hand-crafted features
  - InferSent - A sentence encoder from GloVe embeddings

Model

From “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Devlin et al. 2018

BERT for Clinical Text

<table>
<thead>
<tr>
<th>Model</th>
<th>MedNLI</th>
<th>12b2 2006</th>
<th>12b2 2010</th>
<th>12b2 2012</th>
<th>12b2 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>77.6%</td>
<td>93.9</td>
<td>83.5</td>
<td>75.9</td>
<td>92.8</td>
</tr>
<tr>
<td>BioBERT</td>
<td>80.8%</td>
<td>94.8</td>
<td>86.5</td>
<td>78.9</td>
<td>93.0</td>
</tr>
<tr>
<td>Clinical BERT</td>
<td>80.8%</td>
<td>91.5</td>
<td>86.4</td>
<td>78.5</td>
<td>92.6</td>
</tr>
<tr>
<td>Discharge Summary BERT</td>
<td>80.6%</td>
<td>91.9</td>
<td>86.4</td>
<td>78.4</td>
<td>92.8</td>
</tr>
<tr>
<td>Bio+Clinical BERT</td>
<td>82.7%</td>
<td>94.7</td>
<td>87.2</td>
<td>78.9</td>
<td>92.5</td>
</tr>
<tr>
<td>Bio+Discharge Summary BERT</td>
<td>82.7%</td>
<td>94.8</td>
<td>87.8</td>
<td>78.9</td>
<td>92.7</td>
</tr>
</tbody>
</table>

Table 2: Accuracy (MedNLI) and Exact F1 score (12b2) across various clinical NLP tasks.

From “Publicly Available Clinical BERT Embeddings”, Alsentzer et al. 2019

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Pearson Correlation ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein Distance</td>
<td>--</td>
<td>.680</td>
</tr>
<tr>
<td>Clinical BERT</td>
<td>STS</td>
<td>.771</td>
</tr>
<tr>
<td>Clinical BERT</td>
<td>MedSTS</td>
<td>.849</td>
</tr>
<tr>
<td>Clinical BERT</td>
<td>STS + MedSTS</td>
<td>.854</td>
</tr>
</tbody>
</table>
Outline

1. Introduction to Clinical NLP
2. Concept Linking
3. Clinical Semantic Textual Similarity
4. Phenotyping

Phenotyping

- Task: Identify patient trait from a clinical note
- We cannot rely on only the first part of the document as in some tasks (e.g. topic detection)
- Current well-known solutions
  - For the i2b2 2008 Smoking dataset, the best performing system uses handcrafted regular expressions, rules, and features to train a SVM binary classifier
  - For the i2b2 2008 Obesity dataset, the best performing system consisted only of hand-engineered rules
  - Other approaches have looked at using neural architectures (e.g. CNNs) with worse results
Classification models

- We explored several inputs into our FNN for our document given a set of [CLS] embeddings.
  - Dimension-wise max over all CLS embeddings
  - Input all embeddings into a FNN with padding
  - Input all embeddings into Transformer
  - The final state of an LSTM over all embeddings

Dataset

- 2006 Smoker Identification
  - Past Smoker
  - Current Smoker
  - Non-Smoker
  - Unclear
- 2008 Obesity Identification
  - Obesity
  - 14 Co-Morbidities (e.g. congestive heart failure)
  - As a note can have more than one label, we train a classifier for each label.

Results

Table 1: Phenotyping results (micro-averaged $F_1$) with our architectures compared to the top shared task system and recent deep learning based systems.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{max}$</td>
<td>60.9</td>
<td>74.7</td>
</tr>
<tr>
<td>$f_1$</td>
<td>82.9</td>
<td>81.3</td>
</tr>
<tr>
<td>Transformer</td>
<td>75.9</td>
<td>87.9</td>
</tr>
<tr>
<td>$f_{LSTM}$</td>
<td>97.5 (97.1 ± .48)</td>
<td>94.5 (93.9 ± .59)</td>
</tr>
<tr>
<td>Shared Task 1st Place</td>
<td>90.0</td>
<td>95.0</td>
</tr>
<tr>
<td>CNN [16]</td>
<td>77.0</td>
<td>-</td>
</tr>
<tr>
<td>CNN + Rules [7]</td>
<td>-</td>
<td>96.2</td>
</tr>
</tbody>
</table>

THANK YOU

- This presentation based on slides from
  - Brant Chee, Masoud Rouhizadeh, Elliot Schumacher, Philip Resnik