Biomedical and Clinical Natural Language Processing

Ozlem Uzuner
Meliha Yetisgen
Amber Stubbs
Welcome

• Ozlem Uzuner
  – Associate Professor, University at Albany

• Meliha Yetisgen
  – Assistant Professor, University of Washington

• Amber Stubbs
  – Assistant Professor, Simmons College
Outline

• Introduction to clinical and biomedical NLP
• Research questions in clinical and biomedical NLP
• Data and annotation processes
• Methods
• Open questions and future directions
Introduction to Clinical and Biomedical NLP
Clinical vs. Biomedical NLP

• Clinical NLP
  – Focus on documentation related to patient care
  – Electronic Health Records
    • Both in-patient and out-patient documentation

• Biomedical NLP
  – Focus on scientific discoveries about biology, physiology, and medicine
  – Journal articles, clinical trials, webpages, ...
Ever-growing Volumes of Data

• Clinical text:
  – Millions of patients*
    • Percent of adults who had contact with a health care professional in the past year: 82.1%,
    • Percent of children who had contact with a health care professional in the past year: 92.8%
  – Number of visits (to physician offices, hospital outpatient and emergency departments): 1.2 billion (actual number reported by CDC in 2010)
  – Hospital inpatient care
    • Number of discharges: 35.1 million
    • Discharges per 10,000 population: 1,139.6
    • Average length of stay in days: 4.8

• Biomedical text:
  – PubMed contains more than 23 million biomedical articles from MEDLINE, life science journals, and online books
  – ~500,000 new records are added each year
  – 13.1 million abstracts, and 14.2 million full-text

*http://www.cdc.gov/nchs/fastats/physician_visits.htm
Text Processing Needs in Clinical Practice

• Cohort selection, phenotyping:
  – finding groups of patients that satisfy particular criteria

• Decision support systems:
  – Is there a common practice for treating the disease I am observing? What were the rates of success and what is the best practice?

• Quality assurance in the hospital:
  – Who saw these patients?
Text Processing Needs in Biomedical Domain

• Cohort selection, phenotyping:
  – finding groups of patients that satisfy particular criteria

• Decision support systems:
  – Is there a common practice for treating the disease I am observing? What were the rates of success and what is the best practice?

• Quality assurance in the hospital:
  – Who saw these patients?
Clinical and Biomedical NLP

• Two closely-related complementary domains

• Focus of this tutorial on clinical NLP
HISTORY OF PRESENT ILLNESS: Mrs. [Huntington] is a 77-year-old-woman with long standing hypertension who presented as a Walk-in to me at the [Bronx] Health Center on [DATE]. Recently had been started q.o.d. on Clonidine since [DATE] to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: Lives alone, has one daughter living in [Spring]. Is a non-smoker, and does not drink alcohol.

HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on IV Heparin, Sotalol 40 mg PO b.i.d. increased to 80 mg b.i.d., and had an echocardiogram. By [DATE] the patient had better rate control and blood pressure control but remained in atrial fibrillation. On [DATE], the patient was felt to be medically stable...
The patient is a 46 year old woman with a history of Q wave myocardial infarction with right ventricular infarct in October 1992. Peak CK's were 2300. Catheterization showed 100% RCA lesion which was treated with angioplasty reduced to 20-30% stenosis. Subsequent catheterization October 92, July 92 and September 92 for atypical chest pain, showed clean coronaries. **Exercise tread mill test in September 92**, the patient went three minutes and 31 seconds with standard Bruce protocol and stopped secondary to atypical chest pain. **Maximum heart rate 162, blood pressure 176/90, no ST or T wave changes.** In April 92 she ruled out for myocardial infarction by enzymes and EKG, after presenting with prolonged chest pain. **VQ scan was low probability.** Chest CT ruled out aortic dissection. The patient now presents to the hospital with 24 hours of right sided chest pain, stating that it was squeezing in her right breast, felt to be between the shoulder blades. She complained of shortness of breath, dizziness, weakness and nausea, **no palpitations were noted...**
Pt recently hospitalized 7/19/06 for chf exacerbation (diastolic dysfunction) 2nd to dietary and medicine noncompliance (salty foods, stopped her HCTZ) and continued to smoke. Pt diuresed and sent home on new lasix 60qam 40qpm regimen. Pt noticed steady decline in functional status during the last 3 weeks because of SOB. at baseline should sat 85% on ra, 95% on 6L02NC at rest and ambulation. (on home o2) but now, can't ambulate, sating 83-89% on 6l at rest. also notes pnd, orthopnea. Pt notes intermittent chest pain on and off lasting 5 minutes not associated with exertion or any other cardiac sx. 8/15 dobuta mibi-> ischemia in d1 territory. 11/19 :echo->ef 60%, Pa pressure 48 + RA. no valve dz. rv enlarged and hypokinetic. A/P: pump: decompeesated CHF (diastolic dysfxn, ? cor pulmonale component) 2nd to diet/med non-compliance. uptitrate captopril, continue iv lasix 60 qd with goal net neg 2 liters, daily weights, strict land O. check cxray. Switched to po lasix 10/06, back to lisinopril for d/c Fri. ischemia: has + mibi in past, but no further workup to d1 lesion. can't get ecasa 2nd to vWD. continue BB, will hold off on statin since not hyperlipidemic. rate:tele. ...
Clinical Language

• Domain-specific, jargon, idioms
• Telegraphic, with misspellings, incomplete sentences
• Speculations, hypotheses, and negations
• Some structure

Clinical Language

• Linguistic variation
  – Derivation
    • mediastinal = mediastinum
  – Inflection
    • opacity = opacities; cough = coughed
  – Synonymy
    • Addison’s Disease: Addison melanoderma, adrenal insufficiency, adrenocortical insufficiency, asthenia pigmentosa, bronzed disease, melasma addisonii, ...
    • Chest wall tenderness: chest wall did demonstrate some slight tenderness when the patient had pressure applied to the right side of the thoracic cage

Clinical Language

• Polysemy
  – General polysemy
    • Patient was prescribed codeine upon discharge
    • The discharge was yellow and purulent
  – Acronyms and Abbreviations
    • APC: activated protein c, adenomatosis polyposis coli, adenomatous polyposis coli, antigen presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, alfalfa protein concentrated, allophycocyanin, anaphase promoting complex, anoxic preconditioning, anterior piriform cortex, antibody producing cells, atrial premature complex, ...

Clinical Language

• Negation and uncertainty
• Approximately half of all clinical concepts in dictated reports are negated*
  – Explicit negation
    • The mediastinum is not widened
      – Mediastinal widening: absent
  – Implicit negation
    • Lungs are clear upon auscultation
      – Rales/crackles: absent
      – Rhonchi: absent
      – Wheezing: absent
  – Uncertainty
    • Treated for a presumptive sinusitis

Clinical Language

• Hypotheses
  – *It was felt that the patient probably had a cerebrovascular accident involving the left side of the brain. Other differentials entertained were perhaps seizure and the patient being post-ictal when he was found, although this consideration is less likely.*
  – R/O out pneumonia.

Clinical Language

• Implication
  – Audience for patient reports is physicians
    • That puts lay people at a disadvantage when interpreting these records.

• Requires inference
  – Sentence level inference
    • There were hazy opacities in the lower lobes → Localized infiltrate
  – Report level inference
    • Localized infiltrates → Probable pneumonia

Clinical Language

• More inference
  – Fever
    • Temperature 38.5°C
  – Oxygen desaturation
    • Oxygen saturation low
    • Oxygen saturation 85% on room air

• Temporality
  – Past medical history
    • History of CHF presenting with shortness of left-sided chest pain.
  – Hypothetical or non-specific mentions
    • He should return for fever or increased shortness of breath.
  – Temporal course of disease
    • Patient presents with chest pain ... After administration of nitroglycerin, the chest pain resolved.
Clinical Language

• Report structure
  – Anatomic Location sometimes in section header
    • NECK: no adenopathy.
  – Some sections carry more weight
    • IMPRESSION: atelectasis
  – Some reports contain pasted (or templated) text difficult to process
    • Cardiovascular: [ ] Angina [ ] MI [x ]
      HTN [ ] CHF [ ] PVD [ ] DVT [ ]
      Arrhythmias [ ] Previous PTCA [ ]
      Previous Cardiac Surgery [ ] Negative - Denies CV problems

Biomedical Documents

• Grows in size at a dramatic pace

• 2014 Medline baseline contains:
  – 22,376,811 citations
  – 13,214,810 abstracts
Biomedical Language

• Contains domain specific rich and evolving vocabulary
  – Concepts introduced when new discoveries are presented

• Very structured
  – Journal abstracts and articles usually follow similar section structure

• Sentences are very grammatical but include highly ambiguous terms
  – Neurofibromatosis 2 [disease]
  – Neurofibromin 2 [protein]
  – Neurofibromatosis 2 gene [gene]
Research Problems in Clinical and Biomedical NLP
Clinical and Biomedical Applications

• Applications focus on making information hidden in vast amounts of clinical data available in biomedical literature and clinical records more accessible to improve biomedical research and patient care.
Sample Clinical NLP Application

• Does the patient drink?
  – Classify patient into 3 classes: Heavy consumption, Moderate consumption, None

• Application:
  – Retrospective cohort study of high B12 levels as ICU mortality predictor

• Hypotheses:
  – High B12 levels are associated with liver function
  – Alcohol consumption impacts liver

• Data:
  – The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) database
  – B12 measurements for ~2,000 adult patients

• Structured data:
  – ICD9 codes of alcohol-based illness, e.g., delirium tremens (291*)

Callaghan F.M., Leishear K., Abhyankar S., Demner-Fushman D., McDonald C.J. (2014) High vitamin B12 levels are not associated with increased mortality risk for ICU patients after adjusting for liver function: a cohort study. ESPEN J. 2014 Apr 1;9(2):e76-e83.
Clinical NLP Application

- Search for alcohol OR drink OR...
- Retrieval results:
  - Yes: The patient is known to have a history of alcohol abuse
  - Maybe: tox screen only significant for alcohol level of 273
  - No: She denied any intravenous drug use, tobacco, or alcohol
  - Maybe: He was counseled against the use of alcohol
  - Maybe: He stated that he would not drink alcohol in the future
  - No: Major Surgical or Invasive Procedure: Alcohol septal ablation.
  - Yes: Quit smoking 'many' yrs ago. smoked 1 ppweek. used to drink 10-12 hard drinks every other day. last drink before going to OSH. resolved to quit now. no IVDU.
  - No: her husband is a heavy drinker and often isolates himself to drink alone
  - No: the patient was seen to drink a tremendous amount of water

Clinical NLP Application

- Is ‘alcohol’ used in its beverage sense?
  - Polysemy (concept extraction and word sense disambiguation)
- What are synonyms, hypernyms, and hyponyms of ‘alcohol’?
  - Synonymy (concept extraction)
- Does the patient drink alcohol or not?
  - Negation
- Is alcohol use asserted, equivocal, modal or hypothetical?
  - Uncertainty
- Who is drinking?
  - Experiencer, relation extraction
- What is the patient drinking?
  - Relation, event extraction
- Is the patient drinking currently and regularly?
  - Temporal relations, timeline generation

Clinical NLP Research Questions

- Open questions at all levels of NLP research.
- Low-level tasks
  - Sentence boundary detection
  - Tokenization
  - Part-of-speech tagging
  - Stemming
  - Shallow parsing (chunking)
  - Text segmentation
- High-level tasks
  - Spelling/grammatical error identification and recovery
  - Named entity recognition and information extraction
  - Word sense disambiguation
  - Negation and uncertainty
  - Relationship extraction
  - Temporal reasoning

Low-level Clinical NLP Research Questions

• Sentence boundary detection: complicated by
  – abbreviations and titles, e.g., m.g., Dr.
  – lists and templates, e.g., MI [x], SOB[]

• Tokenization: complicated by
  – characters typically used as token boundaries, e.g.,
    $10\text{ mg/days}$, $N$-acetylcysteine.

• Morphological decomposition: complicated by
  – compound words, e.g., nasogastric

• Text segmentation: complicated by
  – problem-specific needs, e.g., sections, including Chief
    Complaint, Past Medical History, HEENT, etc.

• In general, systems developed for non-clinical
  text often work less well on clinical narratives.

High-level Clinical NLP Research Questions

• Spelling/grammatical error identification and recovery: complicated by
  – highly synthetic phrases
  – incorrectly used homophones, e.g., sole/soul, their/there

• Named entity recognition (NER): complicated by
  – Word/phrase order variation:
    • perforated duodenal ulcer vs. duodenal ulcer, perforated
    • chest tenderness vs. chest wall shows slight tenderness on pressure
  – Derivation:
    • Mediastinum vs. mediastinal
  – Inflection:
    • opacity vs. opacities
  – Synonymy:
    • Addison’s disease vs. adrenocortical insufficiency
  – Acronyms:
    • APC vs. activated protein C vs. adenomatous polyposis coli vs. ...

High-level Clinical NLP Research Questions

• Negation and uncertainty identification: complicated by
  – Explicit and implicit assertions, e.g., *Patient denies chest pain, Lungs are clear upon auscultation*
  – Uncertainty, e.g., *the ill-defined density suggests pneumonia.*
  – Uncertainty in the reasoning processes, e.g., *The patient probably has a left-sided cerebrovascular accident; postconvulsive state is less likely.*

High-level Clinical NLP Research Questions

- Relationships are complicated by
  - The nature of the entities, which affect:
    - Relationship extraction, e.g., causal relations
    - Coreference resolution
    - Temporal reasoning
    - ...

Datasets and the Annotation Processes
Biomedical and Clinical Corpora

• There are various biomedical corpora annotated for syntax and semantics
  – **MedTag**: A collection of biomedical annotations (MEDLINE abstracts): the AbGene corpus of annotated sentences of genes and protein named entities, the MedPost corpus of part of speech tagged sentences and the GENETAG corpus for named entity identification used for BioCreAtIvE I.
  – **TREC Genomics Track**: A set of data collections provided by TREC Genomics Track useful for development and evaluation of retrieval and text categorization strategies in the biomedical domain.
  – **BioCreative corpus**: Dataset produced by the BioCreative assessment, text passages relevant for GO annotations of human proteins.
  – **GENIA corpus**: Annotated corpus of literature related to the MeSH terms: Human, Blood Cells, and Transcription Factors.
  – **Yapex corpus**: Training and test data for the protein tagger (NER) YAPEX.
  – **PASBio**: Predicate-argument structures of biomedical literature.
  – **LLL05 dataset**: Genic Interaction Extraction Challenge: protein/gene interactions IE data set
  – **IEPA corpus**: The Interaction Extraction Performance Assessment corpus
  – **BioText Data**: Dataset for extraction of disease/treatment entities relations
  – **BioText NC Semantics Dataset**: Dataset of Noun Compound Semantics used in experiments described in articles
  – **PennBioIE**: UPenn Biomedical Information Extraction datasets of annotated PubMed abstracts: CYP450 domain and oncology domain
  – **Medstract corpus**: Biomedical annotation corpus useful for acronym definition and coreference resolution
  – **Medstract corpus**: Biomedical annotation corpus useful for acronym definition and coreference resolution
  – **OHSUMED text collection**: Document collection used for the TREC-9 contest.
  – **BMC corpus**: Open access corpus of full text articles provided by BioMed Central.
  – **FetchProt corpus**: Full text journal articles from the biological domain analyzed for experiments on proteins.
  – **PDG Bio-sentence splitter corpus**: Small collection of text data sets derived from PubMed abstracts to develop and assess sentence splitting tools.
  – **Bio1 corpus**: annotated corpus, same field as GENIA, but annotated to small top-level ontology.
  – ...
9 years of i2b2 NLP Challenges

All data available at http://www.i2b2.org/NLP/ DataSets with a data use agreement
i2b2 NLP Shared-task Challenges

Goals:
– Understanding the key information in narrative patient records
– Through extraction and classification tasks
– To enable phenotype extraction from narrative patient records

Run as shared-tasks
– Training data made available
– Testing set held out for evaluation
– Evaluation performed by i2b2
# i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenge 2006 – De-identification

Automatic de-identification

• Patient,
• Doctor,
• Location,
• Hospital,
• Date, ...

HISTORY OF PRESENT ILLNESS: Mrs. [Huntington] is a 77-year-old-woman with long standing hypertension who presented as a Walk-in to me at the [Bronx] Health Center on [DATE]. Recently had been started q.o.d. on Clonidine since [DATE] to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: Lives alone, has one daughter living in [Spring]. Is a non-smoker, and does not drink alcohol.

HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on IV Heparin, Sotalol 40 mg PO b.i.d. increased to 80 mg b.i.d., and had an echocardiogram. By [DATE] the patient had better rate control and blood pressure control but remained in atrial fibrillation. On [DATE], the patient was felt to be medically stable...
## i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td><strong>Smoking Status</strong></td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td>42</td>
</tr>
</tbody>
</table>
Challenge 2006 – Smoking Status

- Document classification into the following classes:
  - Smoker
  - Current smoker
  - Past smoker
  - Non-smoker

HISTORY OF PRESENT ILLNESS: Mrs. [Huntington] is a 77-year-old-woman with long standing hypertension who presented as a Walk-in to me at the [Bronx] Health Center on [DATE]. Recently had been started q.o.d. on Clonidine since [DATE] to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: Lives alone, has one daughter living in [Spring]. Is a non-smoker, and does not drink alcohol.

HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on IV Heparin, Sotalol 40 mg PO b.i.d. increased to 80 mg b.i.d., and had an echocardiogram. By [DATE] the patient had better rate control and blood pressure control but remained in atrial fibrillation. On [DATE], the patient was felt to be medically stable...
## i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td>45</td>
</tr>
</tbody>
</table>
Challenge 2008 – Obesity Diagnosis

• Obesity and 15 of its co-morbidities
  – Asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus (DM), gallstones / cholecystectomy, gastroesophageal reflux disease (GERD), gout, hypercholesterolemia, hypertension (HTN), hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency

• To be classified on a record level
  – Present (Y): the patient has the disease
  – Absent (N): the patient does not have the disease
  – Questionable (Q): the patient may have the disease
  – Unmentioned (U): the disease is not mentioned in the record
<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td><strong>Community annotation</strong></td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenge 2009 – Medication extraction

- Medications and information related to them from medical discharge summaries
  - Medication names
  - Dosages
  - Modes
  - Frequencies
  - Durations
  - Reasons
  - List/narrative
Medication extraction

• Extraction task
  – Medications and information related to them

• Classification task
  – Whether a piece of information is related to a medication


### i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenge 2010 – Relation Extraction

• Three tier task
  – Clinical concepts
  – Assertions on concepts
  – Relations of concepts
## i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# i2b2 NLP Datasets, from i2b2 Challenges

<table>
<thead>
<tr>
<th>Task</th>
<th>Classification (C) or Extraction (E)</th>
<th>Data Sources</th>
<th>Annotation Protocol</th>
<th>Training/Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification</td>
<td>E</td>
<td>Partners Healthcare (PH)</td>
<td>Standard</td>
<td>669 / 220</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>398 / 104</td>
</tr>
<tr>
<td>Obesity Diagnosis</td>
<td>C</td>
<td>PH</td>
<td>Standard by experts</td>
<td>730 / 507</td>
</tr>
<tr>
<td>Medication Extraction</td>
<td>E &amp; C</td>
<td>PH</td>
<td>Community annotation</td>
<td>17 / 251</td>
</tr>
<tr>
<td>Concepts, Assertions, and Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, University of Pittsburgh Medical Center (UPMC)</td>
<td>Standard, then machine validation</td>
<td>349 / 477</td>
</tr>
<tr>
<td>Coreference Resolution</td>
<td>E &amp; C</td>
<td>PH, MIMIC II, UPMC</td>
<td>Standard then machine validation</td>
<td>492 / 322</td>
</tr>
<tr>
<td>Temporal Relations</td>
<td>E &amp; C</td>
<td>PH, MIMIC II</td>
<td>Standard</td>
<td>190 / 120</td>
</tr>
<tr>
<td>Heart disease risk factors</td>
<td>C</td>
<td>PH Longitudinal records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenge 2014 – Heart Disease Risk Factors

• Four separate tracks under one title
  – Heart disease risk factors
  – De-identification
  – Software evaluation
  – Novel data use
Aims

- Use annotation to answer a complex clinical question:
  - “How does heart disease (specifically coronary artery disease) progress in diabetic patients?”
- Question developed in consultation with i2b2 MD/PhDs who study diabetes and cardiology
- De-identify a new corpus of clinical narratives
Corpus selection

- 297 patients
  - 2-5 records per patient
  - 1304 records total
- Postdoc MD selected records
- Cohort selection
  - All patients diagnosed with diabetes
  - 1/3 start with CAD
  - 1/3 develop CAD over course of records
  - 1/3 never develop CAD
Task description for heart disease track

- Identify risk factors for CAD in diabetic patients:
  - hyperlipidemia/hypercholesterolemia,
  - hypertension,
  - obesity,
  - a family history of premature CAD, and
  - being a smoker

- Categorize temporality of risk factors
  - Present before, during, after, or some combinations of those relative to date of record (document creation time, DCT)
Gold standard creation

- All annotators with medical backgrounds
  - 1 medical doctor
  - 6 registered nurses
  - 1 medical assistant

- Each file triple-annotated

- Gold standard: any tags appearing in 2/3 or more of the annotations
Training/testing

- 60-40 split between training and testing
  - 790 records in the training set
  - 514 records in the testing set
How to obtain the data?

- Data available with a use agreement
- Download data use agreement from

http://www.i2b2.org/NLP/DataSets
Methods
Methods

• Depending on the problem a wide range of methods are applied.
  – Rule based
  – Statistical
  – Hybrid
Example clinical projects from UW-BioNLP lab*

• Extracting structure and semantics from clinical text
  – Section segmentation
  – Assertion analysis

• Clinical applications
  – Phenotype modeling in the ICU
    • Pneumonia predictor
    • Acute lung injury predictor
  – Information extraction from radiology notes
    • Clinically important incidental recommendation extractor

*UW-BioNLP Lab: http://depts.washington.edu/bionlp/index.html
HISTORY OF PRESENT ILLNESS:
This is an 85 year old man initially admitted to the Plastic Surgery Service for evaluation of a left facial mass. Subsequently, CMED CCU was consulted and he was transferred to our Service postoperatively.

PAST MEDICAL HISTORY:
His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non-Hodgkin’s lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur. Last colonoscopy, five years ago. Dementia.

ALLERGIES:
No known drug allergies.

MEDICATIONS:
1. Levothyroxine.
2. Lasix.
3. Proscar.
4. Aeroseb.
5. Ancef.

PHYSICAL EXAMINATION:

HOSPITAL COURSE:
He was initially admitted to CMED for resection and repair of this left facial lesion. He also had consults from Urology for his hematuria as well as Medicine preoperatively and CMED CCU. He went to the Operating Room on 2016-03-10 with Urology for hematuria where he had a cystoscopy transurethral resection of prostate placement. He then went to the Operating Room on 2016-03-14 where he had ...
HISTORY OF PRESENT ILLNESS:
This is an 85 year old man initially admitted to the Plastic Surgery Service for evaluation of a left facial mass. Subsequently, CMED CCU was consulted and he was transferred to our Service postoperatively.

PAST MEDICAL HISTORY:
His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non Hodgkin ’s lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur. Last colonoscopy, five years ago. Dementia.

ALLERGIES:
No known drug allergies.

MEDICATIONS:
1. Levothyroxine.
2. Lasix.
3. Proscar.
4. Aeroseb.
5. Ancef.

PHYSICAL EXAMINATION:

HOSPITAL COURSE:
He was initially admitted to CMED for resection and repair of this left facial lesion. He also had consults from Urology for his hematuria as well as Medicine preoperatively and CMED CCU. He went to the Operating Room on 2016-03-10 with Urology for hematuria where he had a cystoscopy transurethral resection of prostate placement. He then went to the Operating Room on 2016-03-14 where he had ...
Framework

• We create a section header ontology for a given note type (e.g., radiology reports, discharge summaries)

• We annotate a small set of document for each of the report types

• We train a two level classifier
  – First classifier identifies the boundaries of the sections
  – Second classifier identified the section type
Radiology report ontology

• Clinical Information
  – Clinical History
• Exam Details
  – Exam
  – Comparison
  – Contrast
  – Procedure
• Findings
  – Findings
• Impression
  – Impression
  – Attending Statement
Discharge report ontology

• General patient information
  – Admit date
  – Discharge date
  – Service

• Provider information
  – Admit physician
  – Attending physician
  – Discharge physician
  – Attending surgeon

• Condition before admission
  – Admission diagnoses
  – History
  – Reason for admission
  – Medications

• Conditions as discharge
  – Discharge diagnoses
  – Other diagnoses
  – Physician exam on discharge
  – Disposition
  – Other diagnoses
  – Condition

• Medical history
  – Allergies
  – Past medical history
  – Past surgical history
  – Family history
  – Gynecological history
  – Social history

• Hospital course
  – Consultation
  – Procedures
  – Hospital course
  – Studies
  – Physical

• Discharge instructions
  – Discharge instructions
  – Discharge medications
  – Follow-up

• Addenda
  – Attending statement
  – Note
HISTORY OF PRESENT ILLNESS:
This is an 85 year old man initially admitted to the Plastic Surgery Service for evaluation of a left facial mass. Subsequently, CMED CCU was consulted and he was transferred to our Service postoperatively.

MEDICAL HISTORY:
His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non Hodgkin 's lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur. Last colonoscopy, five years ago. Dementia.

ALLERGIES:
No known drug allergies.

MEDICATIONS:
1. Levothyroxine.
2. Lasix.
3. Proscar.
4. Aeroseb.
5. Ancef.

PHYSICAL EXAMINATION:

HOSPITAL COURSE:
He was initially admitted to CMED for resection and repair of this left facial lesion. He also had consults from Urology for his hematuria as well as Medicine preoperatively and CMED CCU. He went to the Operating Room on 2016-03-10 with Urology for hematuria where he had a cystoscopy transurethral resection of prostate placement. He then went to the Operating Room on 2016-03-14 where he had ...
Annota0on Process

### HISTORY OF PRESENT ILLNESS:
This is an 85 year old man initially admitted to the Plastic Surgery Service for evaluation of a left facial mass. Subsequently, CMED CCU was consulted and he was transferred to our Service postoperatively.

### MEDICAL HISTORY:
His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non Hodgkin 's lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur. Last colonoscopy, five years ago. Dementia.

### ALLERGIES:
No known drug allergies.

### MEDICATIONS:
1. Levothyroxine.
2. Lasix.
3. Proscar.
4. Aeroseb.
5. Ancef.

### PHYSICAL EXAMINATION:

### HOSPITAL COURSE:
He was initially admitted to CMED for resection and repair of this left facial lesion. He also had consults from Urology for his hematuria as well as Medicine preoperatively and CMED CCU. He went to the Operating Room on 2016-03-10 with Urology for hematuria where he had a cystoscopy transurethral resection of prostate placement. He then went to the Operating Room on 2016-03-14 where he had ...
Two-level classifier

• Line Labeling
  – Labels each line of the report to one of the following three labels
    • B – Beginning of section
    • I – Inside of section
    • O – Outside of section

• Section Type Labeling
  – Labels each section to a section type based on the content of an identified section (e.g. Impression)

• Classifier: MaxEnt
# Features

## Features for line labeling

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text features</td>
<td><code>isAllCaps, isTitleCaps, containsNumber, beginsWithNumber, numTokens, numPreBlanklines, numPostBlanklines, firstToken, secondToken, unigram</code></td>
</tr>
<tr>
<td>Tag features</td>
<td><code>prevTag, prevTwoTags, tagChainLength</code></td>
</tr>
</tbody>
</table>

## Features for section labeling

<table>
<thead>
<tr>
<th>Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header features</td>
<td>Same as tag features, only the header line is used</td>
</tr>
<tr>
<td>Body features</td>
<td><code>avgLineLength, numLines, docPosition, containsList, unigram</code></td>
</tr>
<tr>
<td>Tag features</td>
<td><code>prevTag</code></td>
</tr>
</tbody>
</table>
Results*

Discharge Summaries
Dataset: 430 notes
Performance:
  - Precision: 91.1%
  - Recall: 92.4%
  - F: 91.8%

Radiology Reports
Dataset: 100 notes
Performance:
  - Precision: 93.1%
  - Recall: 91.1%
  - F: 92.1%

Software released at: http://depts.washington.edu/bionlp/software.htm

Assertion analysis

Assertion classification: Given a medical problem concept mentioned in a clinical note (e.g., chest pain), the purpose of a system solving this task to determine whether the concept is:

- present
- absent
- conditional
- hypothetical
- possible
- not associated with the patient (not patient).
Assertion classification (examples)

- The patient was then followed in the cardiac critical care unit where he had evidence of *anoxic encephalopathy*. (present)
- Heart was regular with a I/VI systolic ejection murmur without *jugular venous distention*. (absent)
- He does become *slightly short of breath* when lifting furniture. (conditional)
- If you have *fevers* please contact your PCP or return to the emergency room. (hypothetical)
- The patient was continued on antibiotics for possible *pneumonia*. (possible)
- Father had *coronary artery disease*. (not patient)
Annotated corpus

• Part of 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran’s Affairs (VA) shared-task challenge:
  – 21 teams competed for this task

• Dataset:
  – 826 clinical reports (mostly discharge summaries):
    • 349 training documents (11,968 instances)
    • 477 test documents (18,550 instances)

• Distribution of instances per assertion category:
  – 69% present
  – 20% absent
  – 4.5% hypothetical and possible
  – <1% conditional and not patient
Related work

• De Burjin et al. (2011):
  – state-of-the-art at the 2010 i2b2/VA challenge
  – 93.62 micro-averaged F-measure (primary metric)
  – two layers of classifiers for computing concept predictions
  – wide range of engineered features

• Roberts and Harabagiu (2011):
  – 93.94 micro-averaged F-measure
  – various feature type optimization techniques

• Kim et al. (2011):
  – 94.17 micro-averaged F-measure
  – 79.76 macro-averaged F-measure
  – Implemented additional features with a focus on improving the performance of minority classes
Assertion classification framework

• SVM-based classification framework* (LIBLINEAR)
• SPLAT preprocessing: tokenization, lemmatization, Porter stemming, and constituent and dependency parsing
• Optimization of feature types over the training set (greedy forward/backward feature selection)
• Extracted a wide diversity of features: basic features (lexical, NegEx, ConText, etc.)
  – section features (statistical section chunker*)
  – category specific (N-grams highly correlated with an assertion category)
  – assertion focus features

Basic features

• Encode the surrounding contextual information of the medical concept at the sentence level:
  – word, lemma, and stem uni/bi/tri-grams occurring before and after the concept
  – right sparse stem trigram (e.g., “* lift funitur”)
    • while lifting furniture, if lifting furniture, after lifting furniture
  – tests the presence of special tokens (question mark, conjunction)
  – concept stems
  – the left closest preposition
  – tests the presence of negative prefixes (ab, de, di, il, im, in, ir, re, un, no, mel, mal, and mis)
  – output of ConText and NegEx (window size = 6)
Category specific features

- N-gram features highly correlated with a specific assertion category

<table>
<thead>
<tr>
<th>Ngram</th>
<th>L/R</th>
<th>Absent</th>
<th>NP</th>
<th>Cond.</th>
<th>Hypo.</th>
<th>Possible</th>
<th>Present</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Present</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>found to have</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>1.0</td>
</tr>
<tr>
<td>which showed</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>1.0</td>
</tr>
<tr>
<td>showed</td>
<td>L</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>153</td>
<td>.99</td>
</tr>
<tr>
<td>revealed</td>
<td>L</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>83</td>
<td>.97</td>
</tr>
<tr>
<td>developed</td>
<td>L</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>54</td>
<td>.96</td>
</tr>
<tr>
<td><strong>Conditional</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on exertion</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>with exertion</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>when</td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>.58</td>
</tr>
</tbody>
</table>
## Category specific features

<table>
<thead>
<tr>
<th>Ngram</th>
<th>L/R</th>
<th>Absent</th>
<th>NP</th>
<th>Cond.</th>
<th>Hypo.</th>
<th>Possible</th>
<th>Present</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Absent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no evidence *</td>
<td>L</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>no * or</td>
<td>L</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>did not *</td>
<td>L</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>no</td>
<td>L</td>
<td>1000</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>.99</td>
</tr>
<tr>
<td><strong>Possible</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r/o</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>questionable</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>possible</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>.98</td>
</tr>
<tr>
<td>question of</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>1</td>
<td>.94</td>
</tr>
</tbody>
</table>
The patient was then followed in the cardiac critical care unit where he had evidence of anoxic encephalopathy. (present)

Heart was regular with a I/VI systolic ejection murmur without jugular venous distention. (absent)

He does become slightly short of breath when lifting furniture. (conditional)

If you have fevers please contact your PCP or return to the emergency room. (hypothetical)

The patient was continued on antibiotics for possible pneumonia. (possible)

Father had coronary artery disease. (not patient)
Semantic features

• The patient was then followed in the cardiac critical care unit where he had evidence of anoxic encephalopathy. (present)
• Heart was regular with a I/VI systolic ejection murmur without jugular venous distention. (absent)
• He does become slightly short of breath when lifting furniture. (conditional)
• If you have fevers please contact your PCP or return to the emergency room. (hypothetical)
• The patient was continued on antibiotics for possible pneumonia. (possible)
• Father had coronary artery disease. (not patient)
Semantic features

• Assertion cues:
  – Bioscope *negation cues* (e.g., *not*, *without*, *absence of*)
  – Bioscope *speculative* (or hedge) cues (e.g., *suggest*, *possible*, *might*)
  – *temporal signals* from TimeBank (e.g., *after*, *while*, *on*, *at*)
  – *kinship terms* from Longman English Dictionary (e.g., *mother*, *brother*)

• Semantic features:
  – encode the connection between assertion cues and medical concepts
  – capture the meaning of assertion cues
  – help classifiers decide whether or not a concept is within the *focus* of an assertion cue
Semantic features

- the closest negative cue in the left token context window (size=8)
- the first assertion cue on the path in the dependency tree between the concept and root
- the first verb on the dependency tree path between the medical concept and root
- the modal auxiliary verb associated with the first verb on the dependency tree path between the medical concept and the closest assertion cue
- the sequence of POS labels between the closest left assertion cue and the medical concept
## Assertion classification results*

<table>
<thead>
<tr>
<th></th>
<th>absent 20%</th>
<th>not patient &lt;1%</th>
<th>conditional &lt;1%</th>
<th>hypothetical 4.5%</th>
<th>possible 4.5%</th>
<th>present 69%</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>training set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>basic</td>
<td>95.77</td>
<td>95.66</td>
<td>85.48</td>
<td>57.61</td>
<td>76.19</td>
<td>31.07</td>
<td>94.26</td>
</tr>
<tr>
<td>+sect</td>
<td>96.20</td>
<td>95.78</td>
<td>92.68</td>
<td>82.61</td>
<td>73.33</td>
<td>32.04</td>
<td>95.36</td>
</tr>
<tr>
<td>+spec</td>
<td>96.50</td>
<td>95.78</td>
<td>92.94</td>
<td>85.87</td>
<td>80.39</td>
<td>39.81</td>
<td>95.51</td>
</tr>
<tr>
<td>+focus</td>
<td>96.87</td>
<td>96.37</td>
<td>95.18</td>
<td>85.87</td>
<td>82.35</td>
<td>40.78</td>
<td>95.54</td>
</tr>
<tr>
<td>test set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al 2011</td>
<td>96.31</td>
<td>94.71</td>
<td>97.52</td>
<td>81.38</td>
<td>81.25</td>
<td>30.41</td>
<td>92.07</td>
</tr>
<tr>
<td>Our system</td>
<td>95.71</td>
<td>93.88</td>
<td>91.79</td>
<td>84.83</td>
<td>80.00</td>
<td>30.41</td>
<td>92.42</td>
</tr>
</tbody>
</table>

Open source software

• Available at: http://depts.washington.edu/bionlp/software.htm

• Applied on various note types:
  – UW Medicine
    • Radiology reports, microbiology notes, operative notes, admit notes, discharge summaries, ICU progress notes, ...
  – Fred Hutch:
    • Pathology notes
Examples of Clinical Application Projects

1. Phenotype modeling in the ICU
   – Pneumonia predictor
   – Acute lung injury predictor

2. Information extraction from radiology notes
   – Clinically important incidental recommendation extractor
Application #1 – Phenotype modeling in ICU

- Aim: Developing an automated screening tool that accurately identifies critical illness phenotypes in the ICU
  - Cohort selection for phenotype-genotype correlation (community acquired pneumonia (CAP))
  - Phenotype surveillance: Predicting the likelihood of a patient acquiring a given phenotype the next day in ICU (ventilator associated pneumonia (VAP))
Data - unstructured

• All free-text notes generated during ICU stay
  – Admit notes
  – ICU daily progress notes
  – Acute care daily progress notes
  – Transfer notes
  – Cardiology daily progress notes
  – Respiratory therapy notes
  – Radiology notes (chest x-rays)
  – Microbiology notes
  – Discharge summary
Data – structured (ventilator associated pneumonia)

- Vital signs (temperature, blood pressure, heart rate, oxygen saturation)
- Ventilator settings (tidal volume, FiO2, peep, respiratory rate), laboratory values (white blood cell count, pH, PaO2, PCO2)
- Frequency and character of endotracheal aspirates
- Ventilator bundle elements (e.g., head of bed position)
- Compliance with oral hygiene and chlorhexidine mouthwash
- Date, time, and type of antibiotic therapy administrated
**CAP cohort selection**  
**System architecture (baseline*)**

![Diagram](image)

- **Patient records**
- **MetaMap**
- **Feature Extractor**
- **Pneumonia Learner**
- **Pneumonia Predictor**
- **Training Data**
- **Test Data**

*F-score = 50.70*

---


CAP cohort selection
System architecture (extension*)

*F-score=85.71

Patient records
−> Ranked words
−> Ranked concepts
−> MetaMap
−> Feature Extractor
−> Assertion Classifier
−> Pneumonia Learner
−> Pneumonia Predictor
−> Training Data
−> Test Data
−> Positive
−> Negative

Time-of-onset prediction* (VAP)

- **Instance** (*patient*, *prediction timepoint*)

  Patient A

  InstanceLabel(*Patient A*, Day 0) = −
  InstanceLabel(*Patient A*, Day 1) = −
  InstanceLabel(*Patient A*, Day 2) = +

- **Lookback period (*lp*)**

  *lp* = 2

  Patient A

  F-score=76.46

Other experiments

• Acute lung injury prediction from chest x-rays
  

• CPIS score prediction from chest x-rays
  
Application #2 – Information extraction from radiology notes

• Goal: Clinically important recommendation extraction.

• Setting: patient, clinician, radiologist
  – The clinician orders a radiology test for the patient
  – The radiologist takes X-ray and writes a radiology report, which is sent back to the clinician
Example radiology report

Reason for test: Prostate cancer surveillance

Incidental finding: 6-mm left lung nodule


Architecture
### Features

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>Unigram, bigram, tense, stemmedVerb, includesModalVerb, includesTemporalPhrase</td>
</tr>
<tr>
<td>Knowledge-based</td>
<td>umlsConcept, umlsSemanticType, radlexConcept</td>
</tr>
<tr>
<td>Structural</td>
<td>sectionType</td>
</tr>
</tbody>
</table>

- **Performance:**
  - P: 0.74 R: 0.78 F: 0.76
- **Next steps:** identify details of recommendation
  - Reason for recommendation
  - Recommended test
  - Time frame
- **New NIH grant with start date March 1<sup>st</sup>, 2014.**
Error analysis for both applications

• Current text representation is not sufficient to capture the meaning.
  – Assertion classes are not powerful enough
    • (1) present, (2) absent, (3) conditional, (4) hypothetical, (5) possible, (6) not associated with the patient
  – False Positive Examples from ALI:
    • Report #1: Diffuse lung opacities consistent with pulmonary edema.
    • Report #2: There has been **gradual improvement** of diffuse lung opacities consistent with pulmonary edema.
Error analysis (Cont.)

• Some reports include very complex information that can not be represented with bag-of-words approach
  – Example microbiology note:
    • 2+ Stenotrophomonas maltophilla: Timentin (Ticarcillin/Clavulanic Acid) MIC is 128 Mcg/ml Resistant, Moxifloxacin MIC is 4.0 Mcg/ml. No CLSI interpretive criteria available for this organism antibiotic combination.
Summary of error analysis

• To extract meaningful information more sophisticated representation approaches are needed for clinical text!
Current research

• Events in clinical text
  – How are events represented in clinical text?
  – How can we extract event?

• Report types:
  – Microbiology notes
  – Longitudinal chest x-ray notes
Microbiology notes

• Microbiology laboratory culture tests ordered to (1) identify sources of bacterial infection, (2) determine between differential diagnoses, and (3) adjust antibiotic treatments

• Unlike other report types, microbiology notes change over time as more information is available about the culture
Example microbiology findings

- 60,000 Col/mL Staphylococcus aureus, coagulase positive: See other cultures of Bronchoalveolar Lavage from the same day for sensitivities > 100,000 Col/mL Beta hemolytic Streptococcus not A, C, or G.
- 2+ Stenotrophomonas maltophilia: Timentin (Ticarcillin/Clavulanic Acid) MIC is 128 Mcg/ml Resistant, Moxifloxacin MIC is 4.0 Mcg/ml No CLSI interpretive criteria available for this organism antibiotic combination.
- No Methicillin Resistant Staphylococcus aureus is isolated.
Event definition

• Main attributes:
  – **Organism**: bacteria, flora, fungus, yeast (e.g., *Stenotrophomonas maltophilla*)
  – **Organism quantity**: quantity (e.g. >10000 col/ml)
  – **Rating**: cell confluence rating (e.g. 1+, 2+)
  – **Drug**: drug name that was tested (e.g., Moxifloxacin)
  – **Drug resistance**: susceptibility of the organism to the drug (e.g., resistant)
  – **MIC**: minimum inhibitory concentration (e.g. 2.0 Mcg/ml)

• Additional attributes:
  – **No-growth mention**: no growth mention (e.g., no growth)
  – **No-growth measurement**: time measurement of no growth (e.g., 2 days)
  – **Specimen description**: reference specimen (e.g., 2 days)
  – **Specimen date**: minimum span that identifies the reference specimen date (e.g., same day)
  – **Specimen attribute**: attributes that the reference uses (e.g. reference sensitivities)
Relations

- **equivalentRefOf**: organism-to-organism, or drug-to-drug
- **hasQuantity**: organism-to-organism quantity
- **hasRating**: organism-to-rating
- **measuredBy**: no growth-to-no growth measurement
- **hasDrugDesc**: organism-to-drug, organism-to-drug resistance
- **hasResistance**: drug-to-drug resistance, organism-to-drug resistance
- **hasMIC**: drug-to-MIC
- **hasAttrRefIn**: organism-to-specimen description, organism quantity-to-specimen description
- **timestamp**: specimen description-to-specimen date
- **attr**: specimen description-to-reference item
Annotation examples

Staphylococcus aureus, coagulase positive: See other culture of Bronchoalveolar Lavage from same day for sensitivities > 100,000 Col/mL. Beta hemolytic Streptococcus not A, C, or G.

**Stenotrophomonas maltophilia**: Timentin (Ticarcillin / Clavulanic Acid) MIC is 128 Mcg/ml Resistant, Moxifloxacin MIC is 4.0 Mcg/ml No CLSI interpretive criteria available for this organism antibiotic combination.

Acinetobacter species isolated.
Corpus*

• 1442 microbiology reports from UW Medical center
• 100 reports were double annotated by a medical student and a biomedical informatics PhD student
  – Entity level
    • Kappa: 0.977
    • F-score: 0.964
  – Event level (Exact entity and relation)
    • F-score: 0.960

Entity extraction

- **Rule based:**
  - **Rating:** cell confluence rating (e.g. 1+, 2+)
  - **MIC:** minimum inhibitory concentration (e.g. 2.0 Mcg/ml)
  - **No-growth mention:** no growth mention (e.g., no growth)

- **Hybrid:** (rules + logistic regression to prune false positives)
  - **Organism quantity:** quantity (e.g. >10000 col/ml)
  - **Drug:** drug name that was tested
  - **Drug resistance:** susceptibility of the organism to the drug
  - **No-growth measurement:** time measurement of no growth (e.g., 2 days)
  - **Specimen date:** minimum span that identifies the reference specimen date (e.g., same day)
  - **Specimen attribute:** attributes that the reference uses (e.g. reference sensitivities)

- **Statistical:** (Sequential classification with conditional random fields)
  - **Organism:** bacteria, flora, fungus, yeast
  - **Specimen description:** reference specimen (e.g. lower respiratory culture from endotracheal tube)
## Entity extraction performance

<table>
<thead>
<tr>
<th>RULE</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>453</td>
<td>2</td>
<td>0</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>MIC</td>
<td>83</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>No-growth</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No-growth measure</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Specimen</td>
<td>124</td>
<td>1</td>
<td>3</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Reference</td>
<td>134</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Drug resistance</td>
<td>262</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Organism quantity</td>
<td>738</td>
<td>5</td>
<td>18</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Drug</td>
<td>304</td>
<td>1</td>
<td>7</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Organism</td>
<td>1281</td>
<td>94</td>
<td>123</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>Specimen</td>
<td>109</td>
<td>13</td>
<td>27</td>
<td>0.89</td>
<td>0.80</td>
<td>0.85</td>
</tr>
</tbody>
</table>
## Event extraction performance

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>All entities</td>
<td>0.968</td>
<td>0.952</td>
<td>0.960</td>
</tr>
<tr>
<td>Relations</td>
<td>0.915</td>
<td>0.860</td>
<td>0.886</td>
</tr>
</tbody>
</table>
Events in Radiology Reports

• Radiology reports contain rich semantic vocabulary designed to interpret imaging data with free-text

• Events in this context are descriptions of relevant disease processes found by radiologists on imaging studies
Events with change of state

- ICU Day #1: Diffuse lung opacities consistent with pulmonary edema.
- ICU Day #1: No change in diffuse lung opacities consistent with pulmonary edema.
- ICU Day #2: Diffuse lung opacities consistent with pulmonary edema have worsened.
- ICU Day #2: No change in diffuse lung opacities consistent with edema.
- ICU Day #3: There has been gradual improvement of diffuse lung opacities consistent with pulmonary edema.
Event description

• Main attributes
  – loc: anatomical location
  – attr: something being measured or observed (e.g., volume, opacity)
  – val: a possible value for the attr (e.g., clear)
  – cos: change of state compared to other reports for the same patient (e.g., unchanged)
  – ref: a link to the report(s) that the change of state compared to (e.g., prior examination)
Example annotations

A snippet featuring an event annotation connecting all five fields of the COS tuple.

A snippet featuring shared entities between events
Corpus*

• 1008 sentences from 1344 chest x-ray notes
  – 7173 entities
  – 4128 relations
  – 2101 event tuples
• Agreement:
  – 3 annotators annotated 100 snippets
    • Entity annotation: Kappa = 0.902
    • Event annotation: Kappa = 0.716


Event extraction

• Sequential classification for entity recognition
• SVM for relation classification

• Entity extraction performance:
  – P: 0.94 R: 0.95 F: 095

• Relation extraction experiments are on-going
Future steps

• Running experiments with the VAP data to see whether features extracted from microbiology and radiology events improve VAP classification
Role of Local Context in Understanding Clinical Sublanguage

Özlem Uzuner
MIT CSAIL and SUNY, Albany

Joint Work with: Tawanda Sibanda, Tian He, Jonathan Mailoa, and Peter Szolovits
Contributions and Take Home Message

• We can extract key information from such narratives using a statistical representation of local context

• Local context improves system performance in information extraction and goes a long way towards creating an accurate interpretation of the information buried in clinical sublanguage
Local Context

• Characteristics of the target word (TW) and of the words immediately surrounding the TW
  – Lexical and orthographic features
  – Syntactic features
  – Semantic features
  – ...

• Characteristics of the target word (TW) and of the words immediately surrounding the TW
  – Lexical and orthographic features
  – Syntactic features
  – Semantic features
  – ...
De-identification

• Privacy concerns related to medical records

• Health Information Portability and Accountability Act (HIPAA)
  – 17 pieces of textual Personal Health Information (PHI)
  – PHI in discharge summaries:
    • First and last names of patients, their health proxies, family members
    • Identification numbers
    • Telephone, fax, pager numbers
    • Geographic locations
    • Dates
    • ...
    • Hospital names
    • First and last names of doctors
Discharge Summaries

HISTORY OF PRESENT ILLNESS: Mrs. [Huntington] is a 77-year-old-woman with long standing hypertension who presented as a Walk-in to me at the [Bronx] Health Center on [DATE]. Recently had been started q.o.d. on Clonidine since [DATE] to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: Lives alone, has one daughter living in [Spring]. Is a non-smoker, and does not drink alcohol.

HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on IV Heparin, Sotalol 40 mg PO b.i.d. increased to 80 mg b.i.d., and had an echocardiogram. By [DATE] the patient had better rate control and blood pressure control but remained in atrial fibrillation. On [DATE], the patient was felt to be medically stable...

What if the patient had Huntington’s disease?

Misspelled or foreign name?

What if the patient has to take Tylenol?
Related Work

• Named Entity Recognition (NER)
  – Exploit both the characteristics of the names of the entities and contextual clues related to these entities (Bikel et al.; McCallum et al.; Riloff and Jones; ...)

• Bio-Named Entity Recognition
  – Exploit various feature sets including surface and syntactic features, word formation patterns, morphological patterns, POS tags, etc. (Collier et al.; Yu et al.; ...)

• De-identification
  – Combinations of statistical and rule-based approaches
  – Most statistical approaches focused on sub-categories of PHI (Taira et al.; Thomas et al.; ...)
  – Approaches that target full de-identification use dictionaries, rules, and patterns (Gupta et al.; Douglass; ...)

126
Local Context for De-identification

- **Local context**: Characteristics of the target word (TW) and of the words immediately surrounding the TW
  - Lexical and orthographic features:
    - The target word (TW) itself
    - The word before and after the TW
    - The bigram before and after the TW
    - Capitalization, punctuation, numbers, word length
  - Syntactic features:
    - Part of speech (POS) of TW, of the word before, and of the word after
    - Syntactic bigrams
  - Semantic features:
    - Presence of TW, of the word before, and of the word after in relevant dictionaries
    - MeSH ID
  - The heading of the section in which TW appears
Local Context for De-identification

- **Local context:** Characteristics of the target word (TW) and of the words immediately surrounding the TW
  - Lexical and orthographic features:
  - Syntactic features:
    - Part of speech (POS) of TW, of the word before, and of the word after
    - Syntactic bigrams
      - Transferred to
      - Transferred immediately to
      - Transferred later to
      - ...
  - Semantic features:
  - The heading of the section in which TW appears
Syntactic Information...

• From the output of the Link Grammar Parser (Sleator and Temperly, 1991)

![Diagram showing syntactic links]

- Op links verbs to their plural objects
- Dmc links determiners to their plural nouns
- MVa connects verbs to their modifiers
- Xp links periods to words

LEFT-WALL  John  smokes  two  packs  daily .
Syntactic Bigrams

LEFT-WALL John smokes two packs, NONE → right syntactic bigram
Local Context for De-identification

- **Local context**: Characteristics of the target word (TW) and of the words immediately surrounding the TW
  - Lexical and orthographic features:
  - Syntactic features:
  - Semantic features:
    - Presence of TW, of the word before, and of the word after in relevant dictionaries
    - MeSH ID
  - The heading of the section in which TW appears
Stat De-id

- Multi-class SVM (linear kernel) with local context
- Determine if a word is:
  - PHI
    - Patient name
    - Doctor name
    - Date
    - Phone
    - ID
    - Hospital name
    - Location
  - Non-PHI
Evaluation

• Compare with an Heuristic + Dictionary (H+D) approach that benefits from dictionaries, rules, and patterns (Douglass)

• Compare with approaches that benefit from wider context, i.e., IdentiFinder (Bikel et al.) and a Conditional Random Field (CRF) De-identifier
  – **Wider context**: Characteristics and dependencies of the entities in the sentence containing the target
Methods

• Stat De-id
  – Cross-validated
• CRFD
  – Cross-validated
• H+D
  – Rule-based
• IdentiFinder
  – Obtained pre-trained on newswire
# Data

Table 1: Number of words in each PHI category in the corpora.

<table>
<thead>
<tr>
<th>Category</th>
<th>Random corpus</th>
<th>Authentic corpus</th>
<th>Challenge corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-PHI</td>
<td>17,874</td>
<td>112,669</td>
<td>444,127</td>
</tr>
<tr>
<td>Patient</td>
<td>1,048</td>
<td>294</td>
<td>1,737</td>
</tr>
<tr>
<td>Doctor</td>
<td>311</td>
<td>738</td>
<td>7,697</td>
</tr>
<tr>
<td>Location</td>
<td>24</td>
<td>88</td>
<td>518</td>
</tr>
<tr>
<td>Hospital</td>
<td>600</td>
<td>656</td>
<td>5,204</td>
</tr>
<tr>
<td>Date</td>
<td>735</td>
<td>1,953</td>
<td>7,651</td>
</tr>
<tr>
<td>ID</td>
<td>36</td>
<td>482</td>
<td>5,110</td>
</tr>
<tr>
<td>Phone</td>
<td>39</td>
<td>32</td>
<td>271</td>
</tr>
</tbody>
</table>
Table 2: Distribution of words, i.e., tokens, that are ambiguous between PHI and non-PHI.
## Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>86.45%</td>
<td>86.50%</td>
<td>87.5%</td>
<td>87.5%</td>
<td>12.65%</td>
</tr>
<tr>
<td>Authentic</td>
<td>78.57%</td>
<td>70.33%</td>
<td>54.55%</td>
<td>80.18%</td>
<td>21.97%</td>
</tr>
<tr>
<td>Challenge</td>
<td>14.10%</td>
<td>17.20%</td>
<td>11.40%</td>
<td>26.59%</td>
<td>5.15%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>15.87%</td>
<td>9.19%</td>
<td>14.10%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Authentic</td>
<td>16.12%</td>
<td>10.19%</td>
<td>12.74%</td>
<td>0.02%</td>
</tr>
<tr>
<td>Challenge</td>
<td>15.36%</td>
<td>11.32%</td>
<td>8.61%</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

Table 3: Percentage of words that appear in name, location, hospital, and month dictionaries used by Stat De-id and by the H+D approach.
Evaluation Metrics

- Precision
- Recall
- F-measure
  - Only on the PHI
  - Aggregate over all PHI
Results

F-measures on PHI in the Randomly Redid’ed Corpus

<table>
<thead>
<tr>
<th>Method</th>
<th>Stat De-id</th>
<th>IFinder</th>
<th>H+D</th>
<th>CRFD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>97.63</td>
<td>68.35</td>
<td>77.82</td>
<td>81.55</td>
</tr>
</tbody>
</table>

F-measures on PHI in the Authentic Corpus

<table>
<thead>
<tr>
<th>Method</th>
<th>Stat De-id</th>
<th>IFinder</th>
<th>H+D</th>
<th>CRFD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>96.82</td>
<td>36.80</td>
<td>84.92</td>
<td>87.83</td>
</tr>
</tbody>
</table>

F-measures on PHI in the Challenge Corpus

<table>
<thead>
<tr>
<th>Method</th>
<th>Stat De-id</th>
<th>IFinder</th>
<th>H+D</th>
<th>CRFD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>98.03</td>
<td>33.20</td>
<td>43.95</td>
<td>85.57</td>
</tr>
</tbody>
</table>

All differences significant at alpha=0.05.
F-measure Comparison of Local Context Features on PHI

Random: 97.63, 95.33
Authentic: 96.82, 94.28
Challenge: 98.03, 95.08

All differences significant at alpha=0.05.
We can:
- De-identify medical discharge summaries using a statistical representation of local context

We showed that:
- Stronger the local context, better the performance
  - When using an SVM
Open Questions and Future Directions