**The Dream**

- Develop a comprehensive, precise language of expression for all clinical data
  - It's the language that is precise. Thus, it must be able to state imprecision, uncertainty, etc.
- Translate all actual clinical text into this language
- Develop reasoning/inference methods to draw consequences within this language
- Get clinicians (and others) to use this

**The Reality**

- Most clinical records of observations, interpretations and procedures are stated in free-form natural language
- There are many sources of error and ambiguity
- Language is infinitely varied
- Computers are still poor at doing most text analysis tasks
- But, with significant exceptions, especially for narrow tasks
- Different approaches work best for different tasks -- no universal methods

**Some Typical Tasks**

- Information retrieval -- usually, find an article relevant to x
- Question answering -- answer specific questions from information represented in text
- Learn and generalize -- find and categorize all protein-protein interactions reported in research literature
- Case selection -- find patients based on their clinical characteristics; e.g., find asthmatics who don’t smoke
- Extract diagnoses, symptoms, tests, results, medications, outcomes, etc., from clinical records
- Extract relations among the above: e.g., x was done to rule out y
- Find (and suppress) identifying information to make data safe for public release
Methods

- grep
  - Search for specific words, simple patterns
  - Good for some things: smok.*
- 25 mg Lasix PO QD
- \d+ \[um\]g \[-A-Za-z\]+ (PO|IV|IM) (QD|BID|TID|Q6H|Q4H)
- dictionary + rules
  - E.g., names of people, towns, streets, hospitals, clinics, wards, companies; Mr. xxx.
- supervised training using single word, bigram, etc., features
  - mostly leads to probabilistic models that recover the most likely interpretation
- parsing to recover syntactic structure of sentences
- semantic interpretation in terms of medical vocabularies, taxonomies

Example: Simple text matching

- UMLS contains >1M medically meaningful phrases
- vocabularies from ~150 sources
  - e.g., “heart attack”, “myocardial infarction”, “acute MI”, etc.
  - synonym, antonym, generalization, specialization, co-occurrence links
- 189 semantic types in taxonomy of entities and relations
  - normalizer, all terms indexed by their normalized versions
- Search each of n^2 substrings for match in UMLS; then search for best cover by resulting matches

Example:

Tawanda Sibanda’s MEng thesis, 2006

http://groups.csail.mit.edu/medg/ftp/tawanda/THESIS.pdf

Classifier for De-Id

- Tasks:
  - De-identification: find all of
    - Patients’ and doctors’ first & last names
    - Id numbers
    - Phone, fax, pager numbers
    - Hospital names
    - Geographic locations
    - Dates
  - Try to resolve ambiguity:
    - E.g., "Mr. Huntington, who has Huntington’s Disease"
  - Extract semantic categories
  - Extract semantic relations

- Features:
  - Target word to be classified
  - Words up to 2 words left/right of target
  - Words up to 2 syntactic links left/right of target (using Link Parser, vide infra)
  - Target part of speech
  - Target capitalization
  - Target length
  - MeSH ID of noun phrase containing the target
  - Presences of target ± 1 word in name, location, hospital and month dictionaries
  - Heading of document section where target appears
  - Whether “-” or “/” characters are in target
- Support Vector Machine (linear kernel)
“Secret Sauce”: Syntax

• Link Grammar Parser
  – Lexical database of constraint formulas for each word (many inherit by category)
  – Hundreds of feature pairs; e.g., “plural”

“John lives with his brother.”

+------------------Xp-----------------+
|                    +----Js----+     |
+---Wd--+--Ss-+--MVp-+   +--Ds--+     |
|       |     |      |   |      |     |
LEFT-WALL John lives.v with his brother.n

Evaluation

• Precision = # instances of x correctly classified/total # classified as x (=PPV)
• Recall = # instances of x correctly classified/total # of x in data (=sensitivity)
• F-measure = harmonic average(precision, recall):
  \[ F = \frac{1 + \beta^2}{\frac{\beta^2}{P} + \frac{1}{R}} \times P \times R \]
  – Asymmetry can be modeled by changing \( \beta \)

Test on Four Corpora

1. Re-identified with randomly selected dictionary names and numbers, retaining original formats; e.g., “Szolovits, Peter” ==> “Smith, John”
2. Ambiguous: all names selected from disease, treatment & test dictionaries
3. Non-dictionary: synthesized names; e.g., “O. Ymfgi was admitted …”
4. Authentic: genuine PHI

<table>
<thead>
<tr>
<th>Category</th>
<th>Re-identified</th>
<th>Ambiguous</th>
<th>Non-Dictionary</th>
<th>Authentic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-PHI</td>
<td>17,874</td>
<td>19,275</td>
<td>17,875</td>
<td>112,669</td>
</tr>
<tr>
<td>Patient</td>
<td>1,048</td>
<td>1,047</td>
<td>1,037</td>
<td>294</td>
</tr>
<tr>
<td>Doctor</td>
<td>311</td>
<td>311</td>
<td>302</td>
<td>738</td>
</tr>
<tr>
<td>Location</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>88</td>
</tr>
<tr>
<td>Hospital</td>
<td>600</td>
<td>600</td>
<td>404</td>
<td>656</td>
</tr>
<tr>
<td>Date</td>
<td>735</td>
<td>736</td>
<td>735</td>
<td>1,953</td>
</tr>
<tr>
<td>ID</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>482</td>
</tr>
<tr>
<td>Phone</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>32</td>
</tr>
</tbody>
</table>
## De-Id Results

### Authentic Corpus

<table>
<thead>
<tr>
<th>Method</th>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stat De-ID</td>
<td>PHI</td>
<td>98.46%</td>
<td>95.24%</td>
<td>96.82%</td>
</tr>
<tr>
<td>iFinder</td>
<td>PHI</td>
<td>26.17%</td>
<td>61.98%</td>
<td>36.80%</td>
</tr>
<tr>
<td>H + D</td>
<td>PHI</td>
<td>82.67%</td>
<td>87.30%</td>
<td>84.92%</td>
</tr>
<tr>
<td>Stat</td>
<td>Non-PHI</td>
<td>99.84%</td>
<td>99.95%</td>
<td>99.90%</td>
</tr>
<tr>
<td>iFinder</td>
<td>Non-PHI</td>
<td>98.68%</td>
<td>94.19%</td>
<td>96.38%</td>
</tr>
<tr>
<td>H+D</td>
<td>Non-PHI</td>
<td>99.58%</td>
<td>99.39%</td>
<td>99.48%</td>
</tr>
</tbody>
</table>

### Challenge Questions

- Automatic de-identification of clinical data
- Automatic evaluation of smoking status of patients based on medical records

### Most important features (considered independently)

- Target word
- Syntactic bigrams
- Lexical bigrams
- POS
- Dictionary
- MeSH
- Orthography (punctuation)

### Challenge Questions

- Automatic de-identification of clinical data
- Automatic evaluation of smoking status of patients based on medical records
Data

• ~1000 medical discharge summaries from Partners HealthCare
• Scrubbed semi-automatically
  – One system pass
  – Three manual passes
• Train and test sets representing similar distributions of relevant classes

De-identification Challenge Data

• Focus on the PHI present in discharge summaries
  – Patient: first and last names of patients, their health proxies, and family members. Exclude titles.
  – Doctors: medical doctors and other practitioners; for transcribed records, the transcribers, and their initials. Excludes titles, such as Dr. and MD.
  – Hospitals: hospital names, names of nursing homes where patients are treated and may also reside, room numbers of patients, and buildings and floors related to doctors’ affiliations. Some hospitals, morgues, or nursing homes are described with their street address. These are included in the hospital category.
  – IDs: Any combination of numbers and letters identifying medical records, patients, doctors, or hospitals. All reports start with an id number.
  – Dates: excludes years.
  – Location: Geographic locations such as cities, states, street names, zip codes, and building names and numbers. The professional affiliations of patients and their families are also considered locations.
  – Phone numbers: Telephone, pager, and fax numbers.
  – Ages: Ages over 90.
  – None: none of the above.

De-identification Evaluation

• Metrics
  – Precision, recall, and f-measure (B=1) at token level
  – Micro- and macro-averaged metrics for system-level performance
• Reports on
  – Overall performance 9-way and 2-way (PHI vs. non-PHI)
  – Performance on ambiguous PHI
  – Performance on out-of-vocabulary PHI
## De-id (9-way) F-measure

<table>
<thead>
<tr>
<th>System ID</th>
<th>Micro-averaged F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellner,3, Mitre</td>
<td>0.997434578</td>
</tr>
<tr>
<td>Szarvas,2, Szeged</td>
<td>0.997413881</td>
</tr>
<tr>
<td>Aramaki,1, U Tokyo</td>
<td>0.996031341</td>
</tr>
<tr>
<td>Hara,3, Nara</td>
<td>0.993694909</td>
</tr>
<tr>
<td>Remaining Systems</td>
<td>0.9767-0.9931</td>
</tr>
</tbody>
</table>

## De-id (9-way)(ambiguous)

<table>
<thead>
<tr>
<th>System ID</th>
<th>Micro-averaged F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellner,3</td>
<td>0.941419964</td>
</tr>
<tr>
<td>Szarvas,1</td>
<td>0.940304335</td>
</tr>
<tr>
<td>Hara,3</td>
<td>0.922680373</td>
</tr>
<tr>
<td>Aramaki,1</td>
<td>0.91541839</td>
</tr>
<tr>
<td>Remaining Systems</td>
<td>0.5974-0.8940</td>
</tr>
</tbody>
</table>

* Systems are identified by the last name of the first author and the submission number

## De-id (9-way)(OoV)

<table>
<thead>
<tr>
<th>System ID</th>
<th>Micro-averaged F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellner,3, Mitre</td>
<td>0.983298</td>
</tr>
<tr>
<td>Szarvas,2, Szeged</td>
<td>0.978888</td>
</tr>
<tr>
<td>Aramaki,1, U Tokyo</td>
<td>0.97223</td>
</tr>
<tr>
<td>Hara,3, Nara</td>
<td>0.961143</td>
</tr>
<tr>
<td>Remaining Systems</td>
<td>0.7830-0.9565</td>
</tr>
</tbody>
</table>

## De-id (2-way)

<table>
<thead>
<tr>
<th>System ID</th>
<th>Macro-averaged F-measure</th>
<th>Micro-averaged F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellner,3, Mitre</td>
<td>0.989693751</td>
<td>0.997774522</td>
</tr>
<tr>
<td>Szarvas,2, Szeged</td>
<td>0.989634637</td>
<td>0.997767856</td>
</tr>
<tr>
<td>Aramaki,1, U Tokyo</td>
<td>0.983954094</td>
<td>0.996559061</td>
</tr>
<tr>
<td>Hara,3, Nara</td>
<td>0.972942838</td>
<td>0.99417348</td>
</tr>
<tr>
<td>Remaining Systems</td>
<td>0.9518-0.9714</td>
<td>0.9786-0.9938</td>
</tr>
</tbody>
</table>
De-identification Systems

• Ranking remains almost the same on ambiguous and out-of-vocabulary PHI

General Patterns

• Diverse set of approaches
  – Systems varied in their use of rules and machine learning
  – Systems varied in the features they used for identifying PHI

• Interesting ideas from one or more systems
  – Many made use of rules to recognize PHI with unique format
  – Some systems were adapted from other Natural Language Processing tasks to de-identification
    • Named Entity Recognition systems are easily adapted to this task (though dictionary dependencies cause problems)
    • Text segmentation (parts of a text),
    • Sentence classification,
    • clause chunking
  – Constraints--every mention of a phrase interpreted similarly

Quo Vadis?

• Anecdote:
  – Shawn was admitted to Brigham and Women’s on March 3, 2006.
  – Shawn was admitted to BWH on March 3, 2006.
  – Shawn was admitted to Mass General on March 3, 2006.
  – Mr. Smith was admitted to Massachusetts General Hospital on March 3, 2006.
  – Instance of overtraining

• Much more data should help
  – But annotation is very costly

General Patterns

• General observations
  – Clinical records vary from data traditionally used in Natural Language Processing
  – Despite the difference in the nature of data, systems used for well-studied NLP problems were successfully adapted to de-identification of clinical records
  – Many systems made use of structure of the documents, e.g., headers and footers
    • Szarvas et al.
    • Aramaki et al.
    • Guillen et al.
  – Regular expressions for the structured PHI
  – On this data, surface features and context help de-identification
  – Ambiguities and absence of names from dictionaries make this data more challenging than real data
    • Even on this deliberately more challenging data, performance of systems is impressive
Extracting Assertions

- **Semantic Category Recognition:** identify semantic category of each word in a discharge summary
  - Diseases
  - Treatments
  - Abusive (sic) substances
  - Dosages
  - Practitioners
  - Diagnostic tests
  - Results
  - Signs and symptoms
  - "none"

- **Assertion classification:**
  - Patient definitely has this
  - Someone other than the patient has this
  - Patient may have this
  - Patient does not have this

Semantic Category Recognizer

- 8-way + none Support Vector Machine (linear) classifier
- **Features:**
  - Target
  - Left/right lexical bigrams
  - Section heading
  - Left/right syntactic bigrams
  - Head of noun phrase + syntactic bigrams of head
  - Parts of Speech of target and words ± 2 left/right
  - UMLS semantic type of noun phrase containing target
  - Capitalized?
  - Contains numerals?
  - Contains punctuation?

Comparison

<table>
<thead>
<tr>
<th>Class</th>
<th>Baseline UMLS lookup</th>
<th>Statistical Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>None</td>
<td>0.628</td>
<td>0.683</td>
</tr>
<tr>
<td>Disease</td>
<td>0.656</td>
<td>0.707</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.548</td>
<td>0.726</td>
</tr>
<tr>
<td>Test</td>
<td>0.764</td>
<td>0.560</td>
</tr>
<tr>
<td>Result</td>
<td>0.404</td>
<td>0.358</td>
</tr>
<tr>
<td>Dosage</td>
<td>0.901</td>
<td>0.597</td>
</tr>
<tr>
<td>Symptom</td>
<td>0.653</td>
<td>0.334</td>
</tr>
<tr>
<td>Practitioner</td>
<td>0.486</td>
<td>0.733</td>
</tr>
<tr>
<td>Substance</td>
<td>0.685</td>
<td>0.128</td>
</tr>
</tbody>
</table>

Assertion Classifier

- Rule-based, using regular expressions on common phrases that precede or succeed a problem (± 4 words):
  - “Alter-association” phrases: imply that the problem is someone else’s
  - Negation phrases
  - Uncertainty phrases
- Greedy algorithm, in above order
- If none of the above match, then assert as present.
### Assertion Classification

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>0.929</td>
<td>0.967</td>
<td>0.947</td>
</tr>
<tr>
<td>Absent</td>
<td>0.947</td>
<td>0.900</td>
<td>0.923</td>
</tr>
<tr>
<td>Uncertain</td>
<td>0.723</td>
<td>0.556</td>
<td>0.629</td>
</tr>
<tr>
<td>Alter-Association</td>
<td>1.000</td>
<td>0.810</td>
<td>0.895</td>
</tr>
</tbody>
</table>

### Semantic Relation Recognition

- Relations of interest:
  - Symptom $\leftrightarrow$ treatment
  - Uncertain symptom $\leftrightarrow$ treatment
  - Disease $\leftrightarrow$ test
  - Uncertain disease $\leftrightarrow$ test
  - Disease $\leftrightarrow$ treatment
  - Uncertain disease $\leftrightarrow$ treatment
- Mode of relation
  - Test reveals disease
  - Test conducted to investigate disease
  - none

### Semantic Relations

- For each relation, T. S. developed a k-way SVM classifier to get the mode.
- E.g., disease $\leftrightarrow$ test features
  - # words between concepts
  - Whether disease precedes test
  - Whether other concepts occur in between
  - Verbs between disease and test
  - Two verbs before/after disease and test
  - Head words of disease and test phrases
  - Right/left lexical bigrams of disease and test
  - Right/left syntactic bigrams of disease and test
  - Words between disease and test
  - Path of syntactic links between disease and test
  - Path of syntactically connected words between disease and test