Medical Natural Language Processing

6.872/HST950

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

Structure in a CHB ED Note

Patient seen: 11:45 AM 21 year old male patient injured his right knee. The injury occurred when he was tackled while playing football 2 days ago. He complains of pain and swelling along the medial aspect of the right patella, medial collateral ligament of the right knee and medial collateral ligament of the right knee. He has been able to bear weight. His symptoms are exacerbated by bending his knee.. He has used a knee immobilizer. With some relief. **CURRENT MEDICATIONS:** None. **ALLERGIES:** Denies known allergies. **IMMUNIZATIONS:** Up to date. **PE:** Alert. In no acute distress. Well-developed. Well-nourished. **Right Knee:** Positive for tenderness and swelling involving the medial condyle of the distal right femur. There is no effusion or ecchymosis. Full range of motion. Slight limp. Normal bulk, tone, and strength. Sensation intact. The examination of the other knee is unremarkable. There is no evidence other trauma. **Other PE:** No other injuries. **TREATMENT & COURSE:** Knee immobilizer applied. **DISPOSITION/PLAN:** Discharged in good condition. **ASSESSMENT:** 1. Sprain of the medial collateral ligament 844.1. **ATTENDING NOTE:** Discussed with me agree with plan.

Bulk of Valuable Data are in Narrative Text

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

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

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

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

A arterial blood gases on 100 percent face mask showed an oxygen of 205, CO2 57 and PH 7.3. An electrocardiogram showed ST depressions in V2 through V4 which improved with sublingual and intravenous nitroglycerin. The patient was transferred to the Coronary Care Unit for management of his congestive heart failure , ischemia and probable aspiration pneumonia.

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
- discourse analysis for resolution of pronouns, anaphora

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

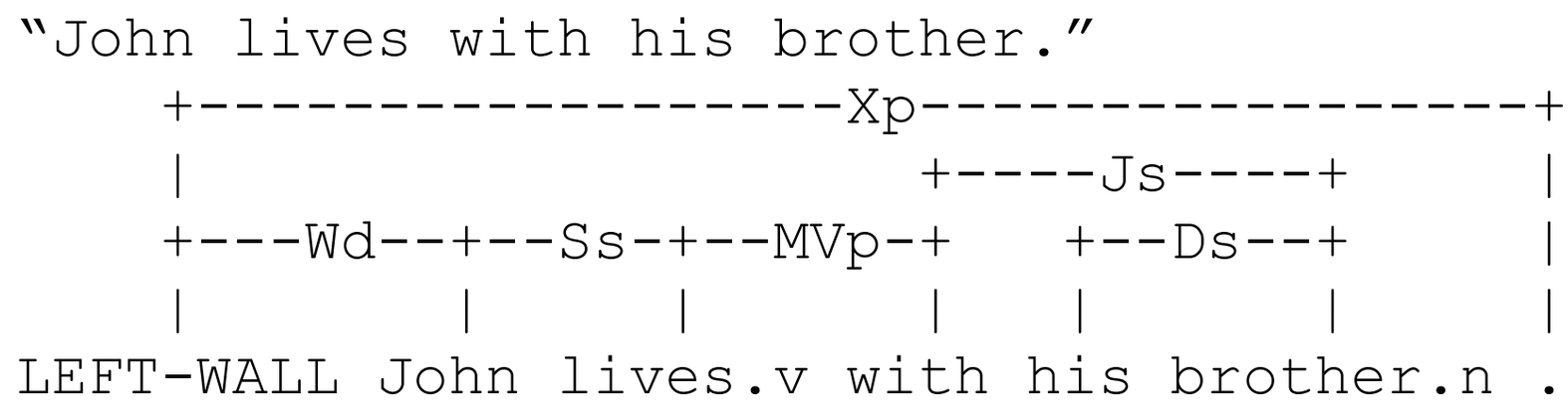
- Tasks:
  - De-identification: find all of
    - Patients' and doctors' first & last names
    - Id numbers
    - Phone, fax, pager numbers
    - Hospital names
    - Geographic locations
    - Dates
    - Try to resolve ambiguity:
      - E.g., "Mr. Huntington, who has Huntington's Disease"
  - Extract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

Classifier for De-Id

- 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”



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): $F = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$
  - –Asymmetry can be modeled by changing β

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

PHI in four corpora

Category	Re-identified	Ambiguous	Non-Dictionary	Authentic
Non-PHI	17,874	19,275	17,875	112,669
Patient	1,048	1,047	1,037	294
Doctor	311	311	302	738
Location	24	24	24	88
Hospital	600	600	404	656
Date	735	736	735	1,953
ID	36	36	36	482
Phone	39	39	39	32

De-Id Results
Authentic Corpus

Method	Class	Precision	Recall	F-measure
Stat De-ID	PHI	98.46%	95.24%	96.82%
iFinder	PHI	26.17%	61.98%	36.80%
H + D	PHI	82.67%	87.30%	84.92%
Stat	Non-PHI	99.84%	99.95%	99.90%
iFinder	Non-PHI	98.68%	94.19%	96.38%
H+D	Non-PHI	99.58%	99.39%	99.48%

Method	Class	Precision	Recall	F-measure
Stat De-ID	PHI	98.40%	93.75%	96.02%
SNoW	PHI	96.36%	91.03%	93.62%
Stat De-ID	Non-PHI	99.90%	99.98%	99.94%
SNoW	Non-PHI	99.86%	99.95%	99.90%

Most important features (considered independently)

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

i2b2 Workshop on Challenges in NLP for Clinical Data, 2006

Ozlem Uzuner, Peter Szolovits, and Isaac Kohane
SUNY Albany, MIT, and Partners Healthcare

Challenge Questions

- Automatic de-identification of clinical data (*de-identification challenge*)
- Automatic evaluation of smoking status of patients based on medical records (*smoking challenge*)

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 Challenge Data

- PHI categories marked while scrubbing
  - Five annotators
  - Agreement 100%
- Realistic surrogates substituted for each PHI type
  - Surrogates obtained by permuting the letters of existing names obtained from the US Census bureau
    - Followed the format of the authentic PHI
    - Most surrogate names are out-of-vocabulary
    - Can generate authentic-looking names which are kept
    - Dates in the same record are all offset by the same number
  - Ambiguity injected into PHI to make the task more challenging
    - Ambiguous PHI lexically overlap with medical terms such as diseases, treatments, and medical tests

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

System ID	Micro-averaged F-measure
Wellner,3, Mitre	**0.997434578**
Szarvas,2, Szeged	0.997413881
Aramaki,1, U Tokyo	0.996031341
Hara,3, Nara	0.993694909
Remaining Systems	0.9767-0.9931

System ID	Macro-averaged F-measure
Wellner,3	**0.941419964**
Szarvas,1	0.940304335
Hara,3	0.922680373
Aramaki,1	0.91541839
Remaining Systems	0.5974-0.8940

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

De-id (2-way)

System ID	Macro-averaged F-measure	Micro-averaged F-measure
Wellner,3, Mitre	**0.989693751**	**0.997774522**
Szarvas,2, Szeged	0.989634637	0.997767856
Aramaki,1, U Tokyo	0.983954094	0.996559061
Hara,3, Nara	0.972942838	0.99417348
Remaining Systems	0.9518-0.9714	0.9786-0.9938

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

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

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

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

	Baseline UMLS lookup			Statistical Classifier		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
None	0.828	0.883	0.855	0.938	0.962	0.950
Disease	0.656	0.707	0.680	0.911	0.899	0.905
Treatment	0.548	0.726	0.625	0.924	0.901	0.912
Test	0.764	0.560	0.646	0.931	0.913	0.922
Result	0.404	0.358	0.380	0.857	0.809	0.832
Dosage	0.901	0.597	0.718	0.966	0.941	0.954
Symptom	0.653	0.334	0.442	0.901	0.815	0.856
Practitioner	0.486	0.733	0.584	0.978	0.934	0.956
Substance	0.685	0.128	0.215	0.934	0.853	0.892

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

Class	Precision	Recall	F-Measure
Present	0.929	0.967	0.947
Absent	0.947	0.900	0.923
Uncertain	0.723	0.556	0.629
Alter-Associati	1.000	0.810	0.895

Semantic Relation Recognition

- Relations of interest:
  - Symptom <==> treatment
  - Uncertain symptom <==> treatment
  - Disease <==> test
  - Uncertain disease <==> test
  - Disease <==> treatment
  - Uncertain disease <==> 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 <==> 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*

HST.950J / 6.872 Biomedical Computing
Fall 2010