NLP For Health (HLP) Monthly Gathering

January 25, 2018
Program & Speaker List

Welcome/Introduction – Graciela Gonzalez, PhD – Associate Professor, Division of Informatics, Department of Biostatistics, Epidemiology and Informatics. gragon@upenn.edu
Tweet @UPennHLP #HLPMMeeting

Short Presentations

1. Michael Becker – Senior Data Scientist, Penn Medicine
2. David Birtwell – Director of Informatics, Biobank; Technical Director, TURBO
3. Ari Klein – Postdoctoral Researcher, Health Language Processing (HLP) Lab, DBEI
4. Ani Nenkova – Associate Professor, CIS
5. David Nicholson – 2nd Yr., Genomic and Computational Biology
6. Ravi B. Parikh, MD, MPP – Fellow, Hematology and Oncology, HUP
7. Li Shen – Informatics, DBEI & IBI
8. Davy Weissenbacher – Research Associate, HLP Lab, DBEI
9. Gary Weissman, MD, MSHP – Clinical Associate of Medicine, PSOM
10. Y. Jesse Zlatsin – Data Analytics Center, Corporate Information Services
Long Presentation

Abeed Sarker– Research Associate, HLP Lab, DBEI
Natural Language Processing for Health (HLP)

January 2018 Gathering

Tweet @UPennHLP #HLPMeeting

Graciela Gonzalez-Hernandez
Contact: gragon@upenn.edu
Twitter: @gracielagon

https://healthlanguageprocessing.org
We Do
✦ Develop algorithm-driven applications that have a positive impact on patients and hospital operations.
✦ Partner with clinicians and researchers to expand the ways patient data is used.
✦ Provide insights into the effectiveness of business processes.
✦ Incorporate clinical notes and other text into predictive applications.

We Need
✦ Clinical and research partners to help identify and champion new use cases.
✦ Input on approaches and methods.

Programs

Continuity of Care

- Heart Failure Readmission Prediction
- Predict cancer hospitalization
- Predict need for Goals of Care discussion
- Reduce patient hospitalization

Acute Care

- Predict inpatient Severe Sepsis/Shock
- Detection for ICU Vent Management
- Risk of maternal decline
- Reduce inpatient decline

Penn Medicine
Predictive Healthcare
Linguamatics I2E – Classic NLP and Search

• Sentence deconstruction via grammars and term hierarchies
• Tailored for the pharma and healthcare spaces
• Text searching with Boolean operators and term expansion
• Comes with pre-packaged term hierarchies such as MedDRA
• Designed to return quick results over large document sets
• Fairly intuitive user interface
• Features include morphological variants and negation
• I2E has yet to be used at Penn Medicine since the evaluation
• Potential intersection with TURBO and PMBB

I2E Evaluation – Trileaflet Aortic Stenosis

• Inclusion and exclusion phrases over echocardiography reports
• I2E outperformed PennSeek with advanced features

<table>
<thead>
<tr>
<th>Human Genes</th>
<th>biomarker</th>
<th>Breast cancer Evidence</th>
<th>Link</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNCG</td>
<td>biomarker is expected to be a useful marker for breast cancer progression and a potential target for breast cancer treatment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNCG</td>
<td>biomarker is expected to be a useful marker for breast cancer progression and a potential target for breast cancer treatment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIM25</td>
<td>Ezh2 immunoreactivity is a significant prognostic factor in breast cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>the TP53 gene is a well-documented strong prognostic factor in breast cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERBB2</td>
<td>HER-2 is an important prognostic factor in breast cancer, and its overexpression is observed in 20-60% of cases with micrometastases in the bone marrow.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCA1</td>
<td>MT is a potential prognostic biomarker for breast cancer, supported by many reports in the literature.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCFL</td>
<td>BORIS is a valuable early marker of breast cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Detection of BORIS in a high proportion of patients with various types of breast tumors indicates that BORIS can be a valuable early blood marker of breast cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test + by PennSeek: 3,189
Test + by ICD9: 3,926

Linguamatics: 6,627
PennSeek: 4,419

1,022
489
A Social Media Mining Approach to Birth Defects Surveillance and Research

- **Research Interest**
  - Although birth defects are the leading cause of infant mortality in the U.S., human data on teratogenic risks remains scarce, making it difficult to provide evidence-based clinical care for pregnant women. I am interested in mining user-posted social media data, through NLP applications, as a potential source of information on birth defects.

- **Project Overview**
  - Our approach shows promise in detecting rare reports of birth defects in large amounts of noisy social media (Twitter) data, but now we face the NLP and machine learning challenges of automating the annotation and analysis processes (e.g., handling imbalanced training data for tweet-level classification; mining timelines for user-level classification).

<table>
<thead>
<tr>
<th>Birth Defect (# of Cases)</th>
<th>Sample Tweets</th>
<th>Class</th>
</tr>
</thead>
</table>
| Down Syndrome (17)       | • Girl at Sprouts, as I push stroller w/1 arm & carry full basket in other: 'you're so strong!' You have no idea. #downsyndromemom #cancermom<br>• @[username] help spread awareness for Down Syndrome. We are walking in just two weeks. This is my sons donation page.<br>• Sally Phillips: My son has Down’s syndrome - but I wouldn’t want to live in a world without it via @[username] | +
|                          |               | ?     |
| Cleft Lip and/or Palate (16) | • yes he was born with a little cleft lip but he is still perfect in my eyes #motherslove #name 😊
• For [name] 😊😊 #cleftlipandpalate weekk2017 [URL]
• I fell asleep like 3 hours ago on this couch 😊😊. Can't watch this cleft palette infomercial though, I can't wake up depressed | +
|                          |               | ?     |

Ari Z. Klein, Ph.D.
421A Blockley Hall
ariklein@pennmedicine.upenn.edu
Health Language Processing Lab (https://healthlanguageprocessing.org)
Department of Biostatistics, Epidemiology, and Informatics
PI: Graciela Gonzalez, Ph.D.
Ani Nenkova: Assoc. prof of CIS

Interests:
- search for evidence-based medicine
- connecting primary literature and informal discussion
- argumentation for improving patient compliance/understanding

Past work:
- *Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls*, K Hong, A Nenkova, ME March, AP Parker, R Verma, CG Kohler, *Psychiatry research* **225** *(1)*, 40-49

- *Aggregating and Predicting Sequence Labels from Crowd Annotations*, AT Nguyen, BC Wallace, JJ Li, A Nenkova, M Lease, *ACL 2017*
HLP Introduction

Research Interests

Hetnets

Know
Machine Learning
Python Programming
NLP (Learning)

Need
Sentence Semantics
Biomedical Knowledge

Contact
David Nicholson
2nd year GCB
 Github: danich1
dnicholson329@gmail.com
Sentiment Analysis of Outpatient Oncology Clinical Notes

Problem:
• Prognosis is documented in <40% of clinical oncology notes
• Important prognostic clues, such as an oncologists’ overall impression of a patient and fluctuating performance status, may not be captured by existing risk models
• Improved prognostic modeling in the outpatient oncology setting can encourage earlier discussions of goals of care, limit unnecessary hospitalizations, and encourage earlier use of palliative care

Area of interest:
• Predictive impact of overall sentiment of outpatient oncology notes using available sentiment analysis tools on 1) mortality, 2) physician practice patterns, 3) hospitalizations
• Development of an oncology-specific NLP-based sentiment lexicon

Resources:
• Penn Data Store – access to unstructured oncology notes

Needs
• Programming support
• Cohort definition

Contact: Ravi.Parikh@uphs.upenn.edu
Li Shen
Informatics, DBEI & IBI
404 Brockley
215-573-2956
li.shen@pennmedicine.upenn.edu

Medical image computing, bioinformatics, machine learning, network science, visual analytics, applications to brain disorders

Medical Image Computing
Map of Science

Biomarker Discovery

Bioinformatics and Imaging Genomics


Created by Sungeun Kim
**Interests:** Information extraction, Machine learning, Reasoning

- Find drug mentions in tweets
- Find places mentioned in Phylogeographic literature
- Estimate syntactic and semantic quality of seniors’ writings (Alzheimer’s disease detection)

**Collaborations:** anything involving texts and semi-structured data

- Working with medical ontologies
- Extracting information from large data set
- Educating patients / medical students
Improving care for patients at risk for and with critical illness and advanced lung disease

• Phenotyping

• Prediction

• Prevention

• Equity

Contact:
Gary Weissman, MD, MSHP
gary.weissman@uphs.upenn.edu
@garyweissman

NLP Collaborators:
Michael Harhay, PhD
Kate Courtright MD, MSHP
Janae Heath, MD
Jessica Dine, MD, MSHP
Rebecca Hubbard, PhD
Casey Greene, PhD
Blanca Himes, PhD
Lyle Ungar, PhD
Scott Halpern, MD, PhD
Data Analytics Center: NLP Solutions

**PennSeek**
- Promotes data analysis while reducing data sourcing burden
- Intuitive shopping-like self service user interface based on Oracle’s Endeca Information Discovery platform
- Reduces data abstraction turnaround by up to 83%
- Provides answers to unanticipated questions

**PennSeek Pro**
- Developed by Linguamatics—I2E
- Supports deeper analysis of patient cohorts
- Supports advanced users with powerful query tools
Social Media Mining for Toxicovigilance

Automatic Monitoring of Prescription Medication Abuse from Twitter

Abeed Sarker (@sarkerabeed)
Health Language Processing Lab
Research Associate
Department of Biostatistics, Epidemiology and Informatics

01/25/2018
Background

- Prescription medication (PM) abuse and overdose— the fastest growing medication-related problem in the U.S.

- PMs and illicit analogs are the leading causes of death by overdose

- Due to the rapidly escalating morbidity and mortality, the problem is now receiving national attention

- Improving surveillance has been recommended as a major step to understand the trajectory of the PM abuse epidemic [1]

Source: https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

### PM abuse information on social media

<table>
<thead>
<tr>
<th>Abuse indicating tweets</th>
<th>Non-abuse tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did I just snort an <strong>oxy</strong>? Is it like 10:30? What can I say I know a Friday when I smell one.......</td>
<td>I love how sometimes you're a personification of an oxymoron and sometimes minus the &quot;oxy&quot;</td>
</tr>
<tr>
<td>I poured 5 shots of whiskey into my lemonade and I took 4 <strong>OxyContin</strong> and I feel so great because I'm not feeling at all</td>
<td><strong>Seroquel</strong> is prescribed. i use valerian root sometimes too. mostly i don’t sleep</td>
</tr>
<tr>
<td>about to be cracked on <strong>adderall</strong> to survive today</td>
<td>speaking of <strong>oxycodone</strong> .. i need to take mine. This pain is ridiculous</td>
</tr>
<tr>
<td>i’m just gonna shower and overdose on <strong>Seroquel</strong> so I’ll sleep until morning</td>
<td>Y'all got to lay off the <strong>Percs</strong> y'all looking crazy out here</td>
</tr>
<tr>
<td>popped <strong>Adderall</strong> tonight hahahah let’s finish this 100 page paper</td>
<td>Where the <strong>Percs</strong> at</td>
</tr>
</tbody>
</table>
Automated detection of PM abuse

- **Goals:**
  1. To verify that abuse-prone PMs have significantly higher abuse information compared to non-abuse-prone PMs
  2. To assess if abuse information can be automatically detected via natural language processing and machine learning
  3. To assess if automatically detected temporal patterns of abuse match manually verified ones

- **Twitter as the source of data**
  - Chosen medications: oxycodone, quetiapine and Adderall® (plus metformin)

- **Annotation**
  - 6400 tweets
  - 2 annotators
Automatic classification

- Supervised classification using annotated data for training and some feature engineering

- Features
  - N-grams, abuse-indicating lexicons, synonym detection and word clusters

- Evaluation on annotated data via 10-fold cross validation

- Trained classifier run on ~130k tweets over ~1 year period to analyze temporal patterns
Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Abuse F-score</th>
<th>Non-abuse F-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.39</td>
<td>0.84</td>
<td>75%</td>
</tr>
<tr>
<td>wSVM</td>
<td>0.45</td>
<td>0.89</td>
<td>81%</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.46</td>
<td>0.89</td>
<td>82%</td>
</tr>
</tbody>
</table>

- Stacking multiple classifiers produced best performance
- Trendline suggests that ~50k training tweets will lead to abuse class F-scores ~0.80
- Automatically detected patterns of abuse over time resembled manually detected ones (for Adderall®) from Hanson et al. (2013) [2]

Summary and future work

- Social media encapsulates an abundance of data about PM abuse and the abusers
- The question is not *if* social media can play a role, but *what* role it can play and *how* it can be best utilized

Plans

1. Collect longitudinal public posts by detected PM abusers
2. Characterize cohort utilizing metadata and NLP techniques
   - Geolocation, gender, time of abuse and age-group (if possible)
3. Study the long-term consequences
Resources

- https://healthlanguageprocessing.org/software-and-downloads/
- https://healthlanguageprocessing.org/sharedtask2/