Using Natural Language Processing to Uncover Signals of Mental State

(accompanied by a brief rant about data)

Philip Resnik
University of Maryland
resnik@umd.edu

NLP State of Science Conference
Veteran’s Administration
September 9, 2015
“Magic is a rich and largely untapped source of insight into perception and awareness. Insofar as the understanding of behaviour and perception goes, there are specific cases in which the magician's intuitive knowledge is superior to that of the neuroscientist.”

S. Macknik et al., “Attention and awareness in stage magic: turning tricks into research”, Nature Reviews Neuroscience 9, 871-879 (November 2008)
“Mistakes were made.”
Ronald Reagan, January 27, 1987

“My toy broke.”
Jay Resnik, frequently
Kentucky clerk walks free

By Emma Margolin

HUMANITARIAN CRISIS
Migrants overwhelm European nations
Watch now

WOMEN'S RIGHTS
Is Clinton's message breaking through?
What's working

NOW ON SHIFT
Sports Matters: Legalized sports gambling
Watch now

DAVIS OUT OF JAIL: Ky. clerk who denied gay marriage licenses is freed

KENTUCKY CLERK KIM DAVIS, who refused to issue marriage licenses to gay couples, was released from jail Tuesday and greeted by a crowd of cheering supporters, including GOP presidential candidate Mike Huckabee (left).

- READ THE OFFICIAL ORDER: U.S. District Court order releasing Kim Davis from custody
- VIDEO: Kim Davis' lawyer: 'She has no regrets'
- VIDEO: Davis joined by attorney, Huckabee at jail release
- VIDEO: Davis’ attorney says last week's licenses are void

CONCERNS CONFIRMED
Intel review backs claim Clinton emails 'top secret'

- MAKING IMPROVEMENTS: Kerry taps State Dept 'transparency' czar to oversee records
- MEDIA BUZZ: Why Hillary won't apologize for email fiasco
- VIDEO: Clinton says no email apology

VICTORY FOR COAL
Colo. mine avoids closure over group's green gambit

- VIDEO: EPA on the hot seat after mine spill disaster
- VIDEO: Clinton doubles down on Obama's green energy agenda

MOTORCYCLE BAR BLAZE
'World's largest biker bar' burns to the ground in SD

- DEADLY BLAZE: Fire at popular nightclub in Cambodia's capital kills 5 women
Accused of sex acts with children, he walked free. Here's why:

walked free - French translation - bab.la English-French
walked free - Dictionary Definition : Vocabulary.com

Apple refused to wiretap an iMessage account for the ...

12 Homeowners Who Refused To Be Forced Out Of Their ...

Refuse - definition of refuse by The Free Dictionary
Refuse implies determination and often brusqueness: “The commander ... refused to discuss questions of right” (George Bancroft). “I'll make him an offer he can't ...

The Cannibal that Walked Free (TV Movie 2007) - IMDb

The Cannibal that Walked Free (also known as Cannibal Superstar) is a British documentary produced by Visual Voodoo for Channel Five which explores ...
Multi-level modeling

[The] press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about. The world will look different to different people depending on the map that is drawn for them by writers, editors, and publishers of the paper they read."


What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

Latent Dirichlet Allocation (LDA), Blei et al. (2003)

For every document, pick the mixture of topics that will be used.

For every word position, pick a topic for that word.

Prior probabilities for document-topic distributions

Prior probabilities for topic-word distributions

Generate a word associated with that topic.
### Table 2  Ten most discussed topics

<table>
<thead>
<tr>
<th>Label</th>
<th>Identifying stems</th>
<th>% Press releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriations/grants</td>
<td>fund, project, 000, million, water, transport, develop, improv, airport, citi</td>
<td>8.6</td>
</tr>
<tr>
<td>Honorary</td>
<td>honor, servic, school, serv, american, veteran, academi, famili, student, world</td>
<td>8.2</td>
</tr>
<tr>
<td>Iraq war</td>
<td>iraq, troop, war, iraqi, american, militari, polit, secur, support, countri</td>
<td>6.6</td>
</tr>
<tr>
<td>Health grants</td>
<td>health, program, educ, children, school, fund, student, care, servic, 000</td>
<td>6.3</td>
</tr>
<tr>
<td>Homeland security</td>
<td>secur, homeland, port, border, depart, fund, guard, air, servic, transport</td>
<td>5.3</td>
</tr>
<tr>
<td>Judicial nominations</td>
<td>court, vote, justic, american, judg, case, hous, congress, constitut, protect</td>
<td>4.8</td>
</tr>
<tr>
<td>Hurricanes/disasters</td>
<td>disast, assist, hurrican, fema, flood, damag, fund, katrina, storm, declar</td>
<td>4.5</td>
</tr>
<tr>
<td>Taxes</td>
<td>tax, american, budget, social, secur, wage, famili, worker, increas, benefit</td>
<td>4.4</td>
</tr>
<tr>
<td>Defense projects</td>
<td>million, defens, fund, air, militari, base, facil, guard, armi, project</td>
<td>4.2</td>
</tr>
<tr>
<td>Health policy</td>
<td>health, care, drug, medicar, senior, prescript, plan, medic, program, cost</td>
<td>3.8</td>
</tr>
</tbody>
</table>

#### Iraq War

![Graph showing Iraq War timeline with peaks for al-Zarqawi's Death, Iraq Study Group Report, and Surge Announcement]

The swine flu finding only fuels fears from law enforcement along the border who say the **illegal immigrants are not being properly screened for diseases and contagious sicknesses** before moving along to other facilities for holding across the nation.

“Some of the children who have come to this country may not have a valid legal basis to remain, but some will. Yet, *it is virtually impossible for a child to assert a valid claim under immigration law in the absence of legal representation.* … It is a fantasy to believe that unrepresented children have a fair shot in an immigration proceeding”

Supervised Hierarchical LDA (SHLDA)

1. For each node $k \in [1, \infty)$ in the tree
   (a) Draw topic $\phi_k \sim \text{Dir}(\beta_k)$
   (b) Draw regression parameter $\eta_k \sim \mathcal{N}(\mu, \sigma)$
2. For each word $v \in [1, V]$, draw $\tau_v \sim \text{Laplace}(0, \omega)$
3. For each document $d \in [1, D]$
   (a) Draw level distribution $\theta_d \sim \text{GEM}(m, \pi)$
   (b) Draw table distribution $\psi_d \sim \text{GEM}(\alpha)$
   (c) For each table $t \in [1, \infty)$, draw a path $c_{d,t} \sim \text{nCRP}(\gamma)$
   (d) For each sentence $s \in [1, S_d]$, draw a table indicator $t_{d,s} \sim \text{Mult}(\psi_d)$
      i. For each token $n \in [1, N_{d,s}]$
         A. Draw level $z_{d,s,n} \sim \text{Mult}(\theta_d)$
         B. Draw word $w_{d,s,n} \sim \text{Mult}(\phi_{c_{d,t}d,s,n})$
   (e) Draw response $y_d \sim \mathcal{N}(\eta^T \tilde{z}_d + \tau^T \tilde{w}_d, \rho)$:
      i. $\tilde{z}_{d,k} = \frac{1}{N_{d,s}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[k_{d,s,n} = k]$
      ii. $\tilde{w}_{d,v} = \frac{1}{N_{d,s}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[w_{d,s,n} = v]$
Supervised Hierarchical LDA (SHLDA)

Supervised Hierarchical LDA (SHLDA)

\[ y_d \sim \mathcal{N}(\eta^T \tilde{z}_d + \tau^T \bar{w}_d, \rho) \]
Supervised Hierarchical LDA (SHLDA)
(Rant coming, not there yet.)
What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

“It is not pleasant to experience decay, to find yourself exposed to the ravages of an almost daily rain, and to know that you are turning into something feeble, that more and more of you will blow off with the first strong wind, making you less and less.


“[Despair], owing to some evil trick played upon the sick brain by the inhabiting psyche, comes to resemble the diabolical discomfort of being imprisoned in a fiercely overheated room. And because no breeze stirs this cauldron, because there is no escape from this smothering confinement, it is entirely natural that the victim begins to think ceaselessly of oblivion.”

— Dr. Kay Redfield Jamison, *Night Falls Fast: Understanding Suicide*

“Ptsd is like being nailed to the cross over and over and over until each body surface, organ and area has had its countless repetitive turns at assaults: has been split and punctured wide open and then left to fry alive in tossed blame, and another‘s abandoned shame”

— PTSD discussion board
SLDA topics from undergraduate stream-of-consciousness essays identified by a clinician as most relevant for assessing depression. Supervision (regression) is based on Z-scored Big-5 scores for emotional instability (neuroticism).

John Doe @johndoe  Aug 25

Ethical dilemma, woman I’m dating doesn’t know I’m diagnosed with depression and struggling with anxiety

Introduction to the CLPsych-2015 Shared and Unshared Tasks: Depression vs. PTSD on Twitter.
Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, Margaret Mitchell (2015).
(Rant coming up very, very soon.)
Most extreme and neutral sLDA topics from Twitter dataset containing authors with self-reported depression (positive) and controls (negative).
Mental health in social media

Selected hierarchical topics derived from Twitter training data using supervised nested latent Dirichlet allocation (SNLDA)
Take-aways

• **Bayesian topic models** provide a way to uncover latent structure in text content

• **Hierarchical models** can capture not only topics, but how those topics are *framed* – an indication of underlying mental state

• **Integrating supervision** makes it possible to predict response variables of interest
• The mantra of NLP progress is that *it’s all about the data.*
• The clinical NLP data shortage is **fatally desperate.**

**P Resnik, Paths to success in computer-assisted coding, 3M white paper, 2014.**

• **We absolutely must have** *large scale, representatively variable clinical data* if we are to make rapid progress.

• **RESEARCHERS CANNOT SOLVE THIS PROBLEM!!!**

Can you?
Collaborators and thanks

• Jordan Boyd-Graber
• Viet-An Nguyen
• Deborah Cai
• Amber Boydstun
• Rebecca Glazier
• Matt Pietryka
• Tim Jurka
• Stephan Greene
• Noah Smith
• Justin Gross
• Kris Miler

• Rebecca Resnik
• William Armstrong
• Leonardo Claudino
• Thang Nguyen
• Glenn Coppersmith
• Meg Mitchell
• Kristy Hollingshead
• Mark Dredze
• Jamie Pennebaker
• IARPA SCIL program
• NSF SOCS program
Thanks!
Linguistic Structured Sparsity in Text Categorization

Dani Yogatama and Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
{dyogatama,nasmith}@cs.cmu.edu
Summary

• Words of a feather (should) flock together

• Idea: use linguistic structure to define *feathers* (flocks) instead of *features*

• Math: sparse group lasso regularization

• Results: text classification (sentiment, forecasting, topic)
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor. It's the kind of thing you either like visually and immediately "get" or you don't. That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid-80s, when Nicholson and Turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t.

that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
### Bag of Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>acting</td>
<td>1</td>
</tr>
<tr>
<td>at</td>
<td>1</td>
</tr>
<tr>
<td>back</td>
<td>1</td>
</tr>
<tr>
<td>basics</td>
<td>1</td>
</tr>
<tr>
<td>big</td>
<td>1</td>
</tr>
<tr>
<td>bit</td>
<td>1</td>
</tr>
<tr>
<td>brutality</td>
<td>1</td>
</tr>
<tr>
<td>but</td>
<td>1</td>
</tr>
<tr>
<td>cheek</td>
<td>1</td>
</tr>
<tr>
<td>crudest</td>
<td>1</td>
</tr>
<tr>
<td>dagerous</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>6</td>
</tr>
<tr>
<td>::</td>
<td></td>
</tr>
</tbody>
</table>
# Linear Classifier

\[
\hat{y} = \text{sign} \left( f(\text{document}) \cdot \mathbf{w} \right)
\]

<table>
<thead>
<tr>
<th>1</th>
<th>acting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>at</td>
</tr>
<tr>
<td>1</td>
<td>back</td>
</tr>
<tr>
<td>1</td>
<td>basics</td>
</tr>
<tr>
<td>1</td>
<td>big</td>
</tr>
<tr>
<td>1</td>
<td>bit</td>
</tr>
<tr>
<td>1</td>
<td>brutality</td>
</tr>
<tr>
<td>1</td>
<td>but</td>
</tr>
<tr>
<td>1</td>
<td>cheek</td>
</tr>
<tr>
<td>1</td>
<td>crudest</td>
</tr>
<tr>
<td>1</td>
<td>dangerous</td>
</tr>
<tr>
<td>1</td>
<td>danger</td>
</tr>
<tr>
<td>6</td>
<td>the</td>
</tr>
</tbody>
</table>
Text is Not a Bag of Words!

• Sentences

This film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor. It’s the kind of thing you either like visually and immediately “get” or you don’t. That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when Nicholson and Turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
Text is Not a Bag of Words!

- Sentences
- Phrases

This film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

It’s the kind of thing you either like viserally and immediately ”get” or you don’t.

That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dagerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t. that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes
- Thematic topics

(and many more!)

this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

It’s the kind of thing you either like viserally and immediately ”get” or you don’t. That is a matter of taste and expectations.

I enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

The acting is very good, if a bit obviously tongue-in-cheek.
Learning the Weights $\mathbf{w}$

"fit the data"
(e.g., log-likelihood of $y_n$ given $d_n$, hinge loss, ...)

\[
\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w}) + R(\mathbf{w})
\]

"generalize"
(e.g., $\lambda \| \mathbf{w} \|_2^2$; $\lambda \| \mathbf{w} \|_1$)
Group Lasso (Yuan & Lin ‘06)

\[ R(w) = \sum_{g} \lambda_g \|w_g\|_2 \]
Group Lasso (Yuan & Lin ‘06)

\[ R(w) = \sum_{g} \lambda_{g} \| w_{g} \|_{2} \]

In NLP:
• chunking and parsing (Martins et al., 2011)
• language modeling (Nelakanti et al., 2013)
Learning the Weights $\mathbf{w}$

\[ \hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w}) + R(\mathbf{w}) \]
Learning the Weights $\mathbf{w}$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w})$$

s.t. $R(\mathbf{w}) \leq \tau$

"Tikhonov" regularization

"Ivanov" regularization
Lasso vs. Group Lasso

\[ R(w) = |w_1| + |w_2| + |w_3| \]

\[ \hat{w} = \arg \min_w \sum_{n=1}^{N} L(f(d_n), y_n; w) \quad \text{s.t.} \quad R(w) \leq \tau \]

Martins et al., EACL 2014 tutorial on structured sparsity in NLP
Lasso vs. Group Lasso

\[ R(\mathbf{w}) = |w_1| + |w_2| + |w_3| \]

\[ R(\mathbf{w}) = \|\langle w_1, w_2 \rangle\|_2 + |w_3| \]

Martins et al., EACL 2014 tutorial on structured sparsity in NLP
Whence Groups?

Back to NLP ...
Sentence Regularizer

\[ R(w) = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_{n,s} \|w_{n,s}\|_2 \]

- Every sentence \( s \) in every document \( n \) gets a group.
- If \( w_{n,s} \) can be driven to zero, that means the sentence is irrelevant to the task.
- Many overlapping groups!

Yogatama and Smith (ICML 2014)
this film is one big joke: you have all
the basics elements
of romance (love at first sight, great
passion, etc.) and gangster flicks
(brutality, dangerous machinations, the
mysterious don, etc.),
but it is all done with the crudest
humor.
it's the kind of thing you either like
viserally and immediately "get" or you
don't.
that is a matter of taste and
expectations.
i enjoyed it and it took me back to the
mid80s, when nicolson and turner were in
their primes.
the acting is very good, if a bit
obviously tongue-in-cheek.
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dangerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t.

that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue - in - cheek.
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th>words/features</th>
<th>groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>actors</td>
<td>✓  ✓</td>
</tr>
<tr>
<td>are</td>
<td>✓  ✓  ✓</td>
</tr>
<tr>
<td>fantastic</td>
<td>✓  ✓  ✓</td>
</tr>
<tr>
<td>.</td>
<td>✓  ✓  ✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>actors</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>are</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>fantastic</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th>words/features</th>
<th>the</th>
<th>actors</th>
<th>are</th>
<th>fantastic</th>
<th>groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Each of 5,000 hierarchical Brown clusters
More Linguistic Structure Regularizers

- Parse tree regularizer

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>actors</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>are</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>fantastic</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Each of 5,000 hierarchical Brown clusters
- Top ten words in each of 1,000 LDA topics
Sparse Group Lasso

$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \| \mathbf{w} \|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; \mathbf{w})$$
Optimization

\[
\min_w R(w) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w)
\]
Optimization

$$\min_{w,v} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w)$$

s.t. $v = Mw$

$\left\{ \begin{align*}
\min_w R(w) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) \\
\text{separate } w \text{ from "copies" } v, \\
\text{constraint forces agreement}
\end{align*} \right\}$
Optimization

\[
\min_{w,v} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) \quad \text{separate } w \text{ from "copies" } v, \\
\text{s.t. } v = Mw \quad \text{constraint forces agreement}
\]
"Augmented Lagrangian"

\[
\frac{1}{2} \| \mathbf{W} - \boldsymbol{\lambda} \| \frac{\mathbf{Z}}{d} + (\mathbf{W} - \lambda) \cdot \mathbf{n} + (\mathbf{W}, (\mathbf{u} \mathbf{p}) \mathbf{j}) \sum_{n=1}^{u} + \frac{1}{\mathbf{m}} \| \gamma + (\lambda) \mathbf{H} \| \mathbf{W} = \boldsymbol{\lambda} \text{ s.t.} \\
\begin{align*}
\text{Separate } \mathbf{W} \text{ from } \text{copies } \mathbf{v}, \\
\text{constraint forces agreement}
\end{align*}
\]

Optimization
Optimization

\[
\min_{w, v} R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) \quad \text{s.t. } v = Mw
\]

\[
\min_{w, v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
\]

**ADMM:** Alternating Directions Method of Multipliers

Alternating, blockwise updates of \(w\) and \(v\)

A “faster” version of dual ascent for solving the augmented Lagrangian (Hestenes ’69; Powell ’69)

(Glowinski & Marrocco ’75; Gabay & Mercier ’76)
“Blockwise” Updates

\[ \min_{w,v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2 \]

\text{w update} \approx \text{loss minimization with elastic net regularization (Zou & Hastie '05)}
“Blockwise” Updates

\[
\min_{w,v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^N L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
\]

v updates: proximal operator for each group:

\[
z_{n,s} = M_{d,s}w - \frac{u_{d,s}}{\rho}
\]

\[
v_{n,s} = \begin{cases}
0 & \text{if } \|z_{n,s}\|_2 \leq \tau \\
\frac{\|z_{n,s}\|_2 - \tau}{\|z_{n,s}\|_2}z_{n,s} & \text{otherwise}
\end{cases}
\]
“Blockwise” Updates

$$
\min_{w,v} \max_u R(v) + \lambda \|w\|_1 + \sum_{n=1}^{N} L(f(d_n), y_n; w) + u \cdot (v - Mw) + \frac{\rho}{2} \|v - Mw\|_2^2
$$

simple dual update $u$
Implications

• Group sparsity and strong sparsity

• Model class is still a (fast) bag of words ... but somehow “informed” by structure

• Learning is more expensive ... but still convex

• A new kind of interpretability ...
this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dagerous machinations, the mysterious don, etc.), but it is all done with the cruelest humor.

it’s the kind of thing you either like viserally and immediately ”get” or you don’t.

that is a matter of taste and expectations.

i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.

the acting is very good, if a bit obviously tongue-in-cheek.
Classification Experiments

- \( L \): Bag of words logistic regression
- Baselines: m.f.c., lasso, ridge, elastic
- Eight datasets
Sentiment

Movies (Socher et al., 2013)  
Votes (Thomas et al., 2006)

- m.f.c.
- lasso
- ridge
- elastic

baselines

sentence
parse
Brown
LDA
Forecasting

![Bar chart showing forecasting results for Science and Bills datasets. The chart compares different baselines including m.f.c., lasso, ridge, elastic, sentence, parse, Brown, and LDA.](image-url)
20 Newsgroups Binary Tasks

![Bar chart showing performance metrics for various tasks in the 20 Newsgroups dataset. The tasks include Science (med/space), Sports (baseball/hockey), Religion (atheism/christian), and Computer (pc/mac). The chart visualizes performance scores using different markers for different methods: m.f.c., lasso, ridge, and elastic.](image-url)
Brown as features or regularizer?

![Graph showing performance metrics for different categories (Science, Sports, Religion, Computer) with various models: best baseline, lasso + Brown, ridge + Brown, elastic + Brown, and our Brown regularizer.](image)
LDA as features or regularizer?
Summary

• Words of a feather (should) flock together

• Idea: use linguistic structure to define *feathers* (flocks) instead of features

• Math: sparse group lasso regularization

• Results: text classification (topics, sentiment, forecasting)

Acknowledgments: Google, IARPA, Pittsburgh Supercomputing Center
Processing Text for HealthCare

Presentations by
Hoifung Poon and Lucy Vanderwende,
Microsoft Research
Processing Text for HealthCare

• Introduction

• The Structure of Free Text in Clinical Records
  • Lucy Vanderwende

• Machine Reading for Cancer Panomics
  • Hoifung Poon
Introduction

Leveraging PubMed and shared task data:

Literome: PubMed-Scale Genomic Knowledge Base in the Cloud
Distant Supervision for Cancer Pathway Extraction from Text
Joint Inference for Knowledge Extraction from Biomedical Literature
Big Mechanisms DARPA Program

Needs patient data:
Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study
Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile
On-time clinical phenotype prediction based on narrative reports

Quantifying the uncertainty in heritability
An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data
Correction for hidden confounders in the genetic analysis of gene expression
The Structure of Free Text in Clinical Records

Lucy Vanderwende

Affiliate Associate Professor,
Biomedical and Health Informatics, UW Medicine
University of Washington

also:
Senior Researcher,
Microsoft Research
Electronic medical records

• Structured data: problem lists, lab results, pharmacy orders, discharge diagnoses, ...

• Unstructured data (Free-form text): radiology reports, operative notes, discharge summaries, ...

• More than 80% of data is in the form of free-text
An example of discharge summary

HISTORY OF PRESENT ILLNESS:
The patient is a 68 year old with acute leukemia.
The patient was in her usual state of health until about three weeks prior to admission when she began to notice increased weakness and bruising.
She presented to a Wood Emergency Department six days prior to admission.
Platelets were 9,000, hemoglobin 9.5, temperature was 100.4.
The patient had a smear there consistent with ALL.
The patient was transferred to Norri Hospital.

REVIEW OF SYSTEMS:
No headache, no nausea, vomiting or diarrhea.
Some shortness of breath with allergies, particularly cats.
No chest pain.
The patient had been doing aerobics three times a week until a couple of weeks before admission.

PAST MEDICAL HISTORY:
The patient's past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestosis.

ALLERGIES:
The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.
The patient does not recollect what her reaction to penicillin was.
The patient also had a history of platelet reaction.

FAMILY HISTORY:
The patient's family history was significant for a brother with colon cancer.
discharge summary: keywords

- acute leukemia
- weakness
- bruising
- Platelets
- hemoglobin
- temperature
- smear
- ALL
- headache
- nausea
- vomiting
- diarrhea
- shortness of breath
- allergies
- chest pain
- admission
- allergies
- depression
- anxiety
- pleural thickening/
- asbestosis
- platelet reaction
- penicillin reactions
- penicillin
- Ampicillin
- colon cancer
HISTORY OF PRESENT ILLNESS:
acute leukemia
weakness bruising
Platelets hemoglobin temperature
smear ALL

REVIEW OF SYSTEMS:
headache, nausea, vomiting diarrhea
shortness of breath allergies
chest pain

PAST MEDICAL HISTORY:
asbestosis allergies depression anxiety pleural thickening /

ALLERGIES:
allergies penicillin reactions Ampicillin
penicillin
platelet reaction

FAMILY HISTORY:
colon cancer
HISTORY OF PRESENT ILLNESS:

- acute leukemia.
- increased weakness and bruising.
- Platelets
- hemoglobin
- temperature
- a smear consistent with ALL.

HISTORY OF PRESENT ILLNESS (continued):

- until about three weeks prior to admission
- six days prior to admission.

REVIEW OF SYSTEMS:

- No headache, no nausea, vomiting or diarrhea.
- Some shortness of breath with allergies, particularly cats.
- No chest pain.

PAST MEDICAL HISTORY:

- past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestos.

ALLERGIES:

- The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

- a history of platelet reaction.

FAMILY HISTORY:

- The patient's family history was significant for a brother with colon cancer.
HISTORY OF PRESENT ILLNESS:

- acute leukemia.
- increased weakness and bruising.
- six days prior to admission.
- hemoglobin 9.5
- temperature 100.4
- Platelets 9,000
- a smear consistent with ALL.

REVIEW OF SYSTEMS:

- No headache, no nausea, vomiting or diarrhea.
- Some shortness of breath with allergies, particularly cats.
- No chest pain.
- until a couple of weeks before admission.

PAST MEDICAL HISTORY:

- past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestos.

ALLERGIES:

- The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.
- a history of platelet reaction.

FAMILY HISTORY:

- The patient's family history was significant for a brother with colon cancer.
**HISTORY OF PRESENT ILLNESS:**

acute leukemia.  

increased weakness and bruising.  

Platelets, hemoglobin, temperature.  

a smear consistent with ALL.

**REVIEW OF SYSTEMS:**

No headache, no nausea, vomiting or diarrhea.  

Some shortness of breath with allergies, particularly cats.  

No chest pain.

**PAST MEDICAL HISTORY:**

past medical history is significant for allergies, depression and anxiety, pleural thickening /  

asbestosis.  

**ALLERGIES:**

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

had a history of platelet reaction.

**FAMILY HISTORY:**

The patient's family history was significant for a brother with colon cancer.

Temporal information, building a timeline
HISTORY OF PRESENT ILLNESS:

- acute leukemia.
- increased weakness and bruising. until about three weeks prior to admission
- six days prior to admission.
- Platelets 9,000, hemoglobin 9.5, temperature 100.4.
- A smear consistent with ALL.

REVIEW OF SYSTEMS:

- No headache, no nausea, vomiting or diarrhea.
- Some shortness of breath with allergies, particularly cats.
- No chest pain. until a couple of weeks before admission.

PAST MEDICAL HISTORY:

- Past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestosis.

ALLERGIES:

- The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.
- A history of platelet reaction.

FAMILY HISTORY:

- The patient's family history was significant for a brother with colon cancer.

Uncertainty: questionable penicillin reactions
HISTORY OF PRESENT ILLNESS:

acute leukemia.

increased weakness and bruising.

until about three weeks prior to admission

six days prior to admission.

Platelets hemoglobin temperature

a smear consistent with ALL.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

past medical history is significant for allergies, depression and anxiety, pleural thickening /

asbestosis.

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.
HISTORY OF PRESENT ILLNESS:

acute leukemia. until about three weeks prior to admission
increased weakness and bruising. six days prior to admission.
Platelets were 9,000, hemoglobin 9.5, temperature was 100.4.
The patient had a smear there consistent with ALL.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.
Some shortness of breath with allergies, particularly cats.
No chest pain.

PAST MEDICAL HISTORY:
past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestosis.

ALLERGIES:
The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.
The patient does not recollect what her reaction to penicillin was.

FAMILY HISTORY:
The patient's family history was significant for a brother with colon cancer.
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. Aerosub.
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 ...
Tool #1 - Section Chunking


Section chunking improves accuracy tagging UMLS concepts

Baseline: MetaMap identifies medical concepts in discharge summaries when checking concepts for comorbidities

<table>
<thead>
<tr>
<th>Comorbidity</th>
<th>System</th>
<th>Prec/Rec/F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>Baseline</td>
<td>82.8/ 84.1/ 83.5</td>
</tr>
<tr>
<td></td>
<td>with Sections</td>
<td>89.5/ 81.0/ 85.0</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Baseline</td>
<td>88.8/ 75.8/ 81.8</td>
</tr>
<tr>
<td></td>
<td>with Sections</td>
<td>92.2/ 75.8/ 83.2</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)

*enabling: 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran’s Affairs (VA) shared-task challenge
Tool #2 - **Assertion analysis**


<table>
<thead>
<tr>
<th>System configuration</th>
<th>Absent</th>
<th>Not patient</th>
<th>Conditional</th>
<th>Hypothetical</th>
<th>Possible</th>
<th>Present</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>macroF</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>+Section</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>+Category specific</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>+Assertion 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>
</tbody>
</table>

18
Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

<table>
<thead>
<tr>
<th>filename</th>
<th>sectionType</th>
<th>CUI</th>
<th>conceptName</th>
<th>semanticTypes</th>
<th>assertValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08693519</td>
<td>history of disease</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08158566</td>
<td>degenerative disc</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08011164</td>
<td>Degenerative</td>
<td>patf</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08735282</td>
<td>Low Back Pain</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08024031</td>
<td>Radiculopathy, grade 1</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08078594</td>
<td>Spondylosis, grade 1</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08064109</td>
<td>Spondylosis</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08080814</td>
<td>Patf</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08097637</td>
<td>Stenosis</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C1261287</td>
<td>Stenosis</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C08038455</td>
<td>CVA, NOS</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C0801847</td>
<td>Diabetes</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C0801894</td>
<td>Diabetes</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C1457887</td>
<td>Symptoms</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C1444648</td>
<td>Microbleeds</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>3_History_of_Present_Illness</td>
<td>C1576875</td>
<td>Write</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>4_Discharge_Diagnoses</td>
<td>C0838438</td>
<td>Spondylosis, lumbar region</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Consultation</td>
<td>C1363945</td>
<td>Treatment</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C2285142</td>
<td>Result</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0801900</td>
<td>Diagnostic</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C08083954</td>
<td>study results</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0808566</td>
<td>Complication</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0858859</td>
<td>change in position</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0827853</td>
<td>Acromegaly</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0801649</td>
<td>Possible</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0859120</td>
<td>Follow-up</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C2285142</td>
<td>Result</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C2285142</td>
<td>Result</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Diagnostic_Studies_Results</td>
<td>C0848954</td>
<td>Hospital course</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Hospital_Course</td>
<td>C0443190</td>
<td>WAS</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Hospital_Course</td>
<td>C1589143</td>
<td>Hyperesthesia</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Hospital_Course</td>
<td>C0802634</td>
<td>condition</td>
<td>dsyn</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Hospital_Course</td>
<td>C0311392</td>
<td>SIGNS</td>
<td>fnfdg</td>
<td>present</td>
</tr>
<tr>
<td>100.txt</td>
<td>6_Hospital_Course</td>
<td>C0838193</td>
<td>Pain</td>
<td>dsyn</td>
<td>present</td>
</tr>
</tbody>
</table>

35 rows in set (0.08 sec)
The patient is a 56-year-old female who was found to have grade 1 L4-L5 spondylolisthesis and lateral recess stenosis at L4-L5 and L5-S1.
Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

“Complication” - absent
... without evidence of **complication** or significant change in position or alignment.
Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

"dermoid cyst" - possible
This is incompletely characterized on the current study and may represent a dermoid cyst.
Extracting Structured Information from Free Text


Next tool being built:
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 #3: There has been **gradual improvement** of diffuse lung opacities consistent with pulmonary edema.
Example annotations

Mineral patchy atelectasis in the right lung is seen, mildly improved since the prior study.

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

Persistent moderate edema and patchy atelectasis/pneumonia.

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


Conclusion

There is rich structure in EMRs far beyond keywords and/or UMLS concepts
We can leverage NLP to makes accessible vast amounts of clinical data available in electronic medical records (EMR)
We extract rich information with the goal of improving clinical research and patient care
Machine Reading for Cancer Panomics

Hoifung Poon
Panomics

... ATTCGGATATTTAAGGC ...

Genome  Transcriptome  Epigenome
Genotype $\rightarrow$ Phenotype

High-Throughput Data

Disease Genes
Drug Targets
Precision Medicine
Vemurafenib on BRAF-V600 Melanoma

Before Treatment

15 Weeks
Vemurafenib on BRAF-V600 Melanoma

Before Treatment

15 Weeks

23 Weeks
Cancer Panomics

High-Throughput Experiments

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning
Example: Tumor Molecular Board
Example: Tumor Molecular Board

- 10-20 highly trained specialists
- Tens of hours on each patient
- Problem: Hard to scale

U.S. 2014: 1.6 million new cases, 585K deaths

- Wanted: Decision support for clinical genomics
Decision Support for Clinical Genomics

Raw Reads → Variant Call RNA-Seq → Clinical Observation → Decision Support → Knowledge Graph → Literature

Clinicians
Decision Support for Clinical Genomics

- Raw Reads
- Variant Call RNA-Seq
- Clinical Observation
- Decision Support
- Knowledge Graph
- Literature
- NLP
- Clinicians
Pathway Knowledge

Genes work synergistically in pathways
Why Hard to Identify Drivers?

Complex diseases ← Perturb multiple pathways

Hanahan & Weinberg [Cell 2011]
Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

EphA2 → Ovarian Cancer → EphB2
Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

EphA2  EphB2

Ovarian Cancer
Cancer Systems Modeling

Transcription: DNA → mRNA
Translation: mRNA → Protein
Activation: Protein → Protein Active

Functional activity
Mutation effect
Drug Target

... ATTCGGATATTTAAGGC ...

[Image of DNA and mRNA structures]
Knowledge → Model

Gene A
DNA → mRNA → Protein → Protein Active

Gene B
DNA → mRNA → Protein → Protein Active

Gene C
DNA → mRNA → Protein → Protein Active

Transcription Factor

Protein Kinase
PubMed

- 24 millions abstracts
- Two new abstracts every minute
- Adds over one million every year
VDR+ binds to SMAD3 to form...

JUN expression is induced by SMAD3/4...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins. BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...
Bottleneck: Annotated Examples

- GENIA (BioNLP Shared Task 2009-2013)
  - 1999 abstracts
  - MeSH: human, blood cell, transcription factor
- Challenge for “supervised” machine learning
- Can we breach this bottleneck?
Free Lunch: Existing KBs

- Many KBs available
  - Gene/Protein: GeneBank, UniProt, …
  - Pathways: NCI, Reactome, KEGG, BioCarta, …
- Indirect supervision
Relation Extraction

**NCI-PID Pathway KB**

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

**TP53 inhibits BCL2.**

*Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.*

*BCL2 transcription is suppressed by P53 expression.*

*The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …

……*
Relation Extraction

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
</tbody>
</table>

**NCI-PID Pathway KB**

*TP53 inhibits BCL2.*

Tumor suppressor P53 downregulates the activity of BCL2 proteins. BCL2 transcription is suppressed by P53 expression. The inhibition of B-cell CLL/Lymphoma 2 expression by TP53...

**Distant Supervision**
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Grounded Semantic Parsing

- Generalize distant supervision to extracting nested events
- Prior: Favor semantic parse grounded in KB
- Outperformed 19 out of 24 participants in GENIA Shared Task [Kim et al. 2009]


http://literome.azurewebsites.net
PubMed-Scale Extraction

- Preliminary pass:
  - 1.5 million instances
  - 13,000 genes, 838,000 unique regulations

- Applications:
  - UCSC Genome Browser, MSR Interactions Track
  - Expression profile modeling
  - Validate de novo pathway prediction
  - Etc.

Personalized medicine approach to treating AML

The Leukemia & Lymphoma Society (LLS) and the Knight Cancer Institute at Oregon Health & Science University are leading a pioneering collaboration to develop a personalized medicine approach to improve outcomes for patients with acute myeloid leukemia (AML), a particularly devastating cancer of the blood and bone marrow. LLS provided $8.2 million to fund Beat AML, and here is how the collaboration will work:

1. In coordination with the Knight Cancer Institute, Stanford University, UT Southwestern Medical Center and Huntsman Cancer Institute will collect data from 900 AML patient samples within 3 years.

2. Illumina will perform genetic sequencing to identify mutations in the patient samples collected.

3. Drug and biotech companies will work with the collaboration to test drug compounds that target mutations suspected of driving disease progression. Array BioPharma will be first to test a therapeutic.

4. Intel will work with Knight Cancer’s bioinformatics team to apply its technology to accelerate computational analysis of the mutation data collected.
Future: VA MVP

Panomics

EHR

Precision Medicine

Literature
Collaborators

- **Chicago**: Andrey Rzhetsky, Kevin White
- **OHSU**: Brian Drucker, Jeff Tyner
- **Berkeley AMP Lab**: David Patterson
- **Wisconsin**: Mark Craven, Anthony Gitter
- **Microsoft Research**: Chris Quirk, Kristina Toutanova, David Heckerman, Scott Yih, Lucy Vanderwende, Bill Bolosky, Ravi Pandya
Summary

... ATTCGGATATTTAAGGC ...
... ATTCGGGTATTTAAGGCC ...

High-Throughput Data

KB

Disease Genes
Drug Targets

......