Statistical Text Analysis for Social Science

Brendan O’Connor
Machine Learning Department
Carnegie Mellon University

Thesis defense presentation, Aug. 19, 2014
Computational Social Science

David Lazer,1 Alex Pentland,2 Lada Adamic,3 Sinan Aral,4 Albert-László Barabási,5 Devon Brewer,6 Nicholas Christakis,1 Noshir Contractor,7 James Fowler,8 Myron Gutmann,3 Tony Jebara,9 Gary King,1 Michael Macy,10 Deb Roy,2 Marshall Van Alstyne2,11

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven “computational social science” has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the
Computational Social Science

Official social data

Data collection

100 BCE

1829

1890

Data analysis

Data computation

Semi-structured social data

Digitized behavior
Billions of users, messages/day

Digitized news
Thousands of articles/day

Digitized archives
Millions of books/century
Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran’s nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran’s program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.

Antigovernment Protesters Try to Shut Down Bangkok

BANGKOK — Antigovernment protesters seeking to block next month’s elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. “We have to shut down Bangkok,” said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. “This is our last resort.” By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.
Text as “Data”?
TextGenerator(SocialAttributes) → Text

1. Infer: attributes of society (language for measurement)
   \[ P(SocAttr \mid Text, Generator) \]
   - e.g. opinion, communities, events...

2. Infer: social determinants of language use
   \[ P(Generator \mid Text, SocialAttributes) \]
   - e.g. bias, influence...

Language generation as social process
\[ P(TextGen \mid Text, SocAttr) \]

Language for social measurement
\[ P(SocAttr \mid Text, TextGen) \]
TextGenerator(SocialAttributes) → Text

Society (SocialAttributes) → Writing (TextGenerator) → Text Data (Text)

Public opinion → Social media usage

Language for social measurement

P(SocAttr | Text, TextGen)

[O’Connor et al., 2010]
TextGenerator(SocialAttributes) → Text

Society (SocialAttributes) → Writing (TextGenerator) → Text Data (Text)

Data generation process

Real-world political events

News media process

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)

[O’Connor, Stewart, Smith 2013]

Israeli–Palestinian Diplomacy
TextGenerator(SocialAttributes) → Text
TextGenerator(SocialAttributes) → Text

Data generation process

Society (SocialAttributes) → Writing (TextGenerator) → Text Data (Text)

Geography of authors

Social media usage

Language generation as social process

P(TextGen | Text, SocAttr)

Challenges for direct autoregressive models

The simplest modeling approach would be an autoregressive model that operates directly on the word counts or frequencies (Wei, 1994). A major challenge for such models is that Twitter offers only a sample of all public messages, and the sampling rate can change in unclear ways (Morstatter et al., 2013). For example, for much of the timespan of our data, Twitter’s documentation

---

Language generation as social process

P(TextGen | Text, SocAttr)

[Eisenstein et al. 2010, O’Connor et al. 2010, Eisenstein et al. 2012]
TextGenerator(SocialAttributes) → Text

- Social media and polls
- Geography and language
- Social determinants of lexical diffusion
- Events in international relations
- Text exploration on document covariates
Data generation process

TextGenerator(SocialAttributes) → Text

Society (SocialAttributes) → Writing (TextGenerator) → Text Data (Text)

Descriptive analysis:
Statistical associations

• Text exploration on document covariates

[O’Connor, 2014]
MiTextExplorer: a Mutual Information Text Explorer using Linked Brushing with Document Covariates

[O’Connor, 2014]

http://brenocon.com/mte
How are X and Y related?  (Anscombe 1973)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.46</td>
</tr>
<tr>
<td>8</td>
<td>6.77</td>
</tr>
<tr>
<td>13</td>
<td>12.74</td>
</tr>
<tr>
<td>9</td>
<td>7.11</td>
</tr>
<tr>
<td>11</td>
<td>7.81</td>
</tr>
<tr>
<td>14</td>
<td>8.84</td>
</tr>
<tr>
<td>6</td>
<td>6.08</td>
</tr>
<tr>
<td>4</td>
<td>5.39</td>
</tr>
<tr>
<td>12</td>
<td>8.15</td>
</tr>
<tr>
<td>7</td>
<td>6.42</td>
</tr>
<tr>
<td>5</td>
<td>5.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>19</td>
<td>12.50</td>
</tr>
<tr>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>
How are X and Y related?  (Anscombe 1973)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>0.82</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>0.82</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>0.82</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>0.82</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>0.82</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
<td>0.82</td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
<td>0.82</td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
<td>0.82</td>
</tr>
<tr>
<td>14</td>
<td>8.10</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
<td>0.82</td>
</tr>
<tr>
<td>4</td>
<td>3.10</td>
<td>0.82</td>
</tr>
<tr>
<td>12</td>
<td>9.13</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.46</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>6.77</td>
<td>0.82</td>
</tr>
<tr>
<td>13</td>
<td>12.74</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>7.11</td>
<td>0.82</td>
</tr>
<tr>
<td>11</td>
<td>7.81</td>
<td>0.82</td>
</tr>
<tr>
<td>14</td>
<td>8.84</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>6.08</td>
<td>0.82</td>
</tr>
<tr>
<td>4</td>
<td>5.39</td>
<td>0.82</td>
</tr>
<tr>
<td>12</td>
<td>8.15</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>6.42</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>5.73</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6.58</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>5.76</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>7.71</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>8.47</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>7.04</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>5.25</td>
<td>0.82</td>
</tr>
<tr>
<td>19</td>
<td>12.50</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>5.56</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>7.91</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>6.89</td>
<td>0.82</td>
</tr>
</tbody>
</table>
How are X and Y related?  
(Anscombe 1973)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
</tr>
</tbody>
</table>

$r = 0.82$

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

$r = 0.82$

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.46</td>
</tr>
<tr>
<td>8</td>
<td>6.77</td>
</tr>
<tr>
<td>13</td>
<td>12.74</td>
</tr>
<tr>
<td>9</td>
<td>7.11</td>
</tr>
<tr>
<td>11</td>
<td>7.81</td>
</tr>
<tr>
<td>14</td>
<td>8.84</td>
</tr>
<tr>
<td>6</td>
<td>6.08</td>
</tr>
<tr>
<td>4</td>
<td>5.39</td>
</tr>
<tr>
<td>12</td>
<td>8.15</td>
</tr>
<tr>
<td>7</td>
<td>6.42</td>
</tr>
<tr>
<td>5</td>
<td>5.73</td>
</tr>
</tbody>
</table>

$r = 0.82$

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>19</td>
<td>12.50</td>
</tr>
<tr>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>
How are $X$ and $Y$ related? (Anscombe 1973)

**Pearson correlation**

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

assumes $(x, y) \sim N(\mu, \Sigma)$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

Is there an analogue to the scatterplot, when text is a variable?

**Scatterplot:**

$x =$ horizontal position

$y =$ vertical position

**Simple Non-parametric(?)**
Linking and brushing

GGobi software
(Cook and Swayne 2007, Buja et. al 1996, etc.)

Is there an analogue to linking/brushing, when text is a variable?
Text and document covariates

- **X**: Text
  - Discrete, high-dimensional (e.g. bag of words)
- **Y**: Document covariates (metadata)
  - Time, author attributes, social context, geography, community membership...
  - Discrete or continuous
  - Lower dimensional
- Goal is **exploratory data analysis**: first-cut insight into \( \text{relationship}(X,Y) \)
- Requirement: speed for interactivity
Linked views of the data

(A) Covariate display
(C) Covariate-word associations
(E) Keyword-in-context text display
[A] \rightarrow [C]: \text{words related to covariate query } Q
Q \text{ selection: “brushing”}

\begin{align*}
\text{rank}_w \frac{p(w|Q)}{p(w)} \\
\text{where } p(w|Q) \geq \text{TermProbThresh} \\
\text{count}_Q(w) \geq \text{TermCountThresh}
\end{align*}

(Exponentiated) Pointwise Mutual Information (a.k.a. lift)
$[C] \rightarrow [D]:$ word-word associations

(Exponentiated) Pointwise Mutual Information (a.k.a. lift)

$$\text{rank}_v \frac{p(v|w \in \text{doc})}{p(v)}$$
KWIC (keyword-in-context)
KWIC reveals word senses

user1110

guess i'm going to the jungle (la) @killa_kimbo its totally true :o " (@seanygrey i will be in la by morning :) that's a fuckin

user29006

@per . @gastelo12 did u bust ? la ! la laa laa la la laa . goodmorning . @gastelo12 did u bust ? la ! la laa laa la la laa . goodmorning

two did u bust ? la ! la laa laa la la laa . goodmorning my little sis did u bust ? la ! la laa laa la la laa . goodmorning my little sis

user31473

me @cherylsatjipto ;) balik dr la kpn ? bb is a distraction , it k

user34771

my twin sister is going to be in la for my bros middleschool gr:

user47627

san francisco is way better that la trust me . :) @teammahone y

user5149

king you a nuisance . i'll be in la this weekend hobnobbing with myself right now . just drove to la from sf and back alone for th

user5239

co @jorge_cortesc en pipolos la comida esta super grasosa #t

@s me voy a dormir aca ya es la 1am supongo q alla las 3am
Covariate -- word analysis

direct PMI -vs- topic model bottleneck

- Feature selection
- Monroe et al. (2008)

- \( p(\text{text} | \text{covariates}) \): Dirichlet-Multinomial Regression, Author-Topic Model, Labeled LDA, Structural Topic Model ...

- \( p(\text{text}, \text{covariates}) \): Supervised LDA, MedLDA, GeoTM ...
Related work: Text Exploration

- Voyant/Voyeur (Rockwell et al. 2010)
- WordSeer (Shrikumar 2013)
- Jigsaw (Görg et al. 2013)
- Topical Guide (Gardner et al. 2010)
- etc...
• Other uses

• Figure out NLP models and parameters (what should be a stopword?)

• Select documents to read in an intelligent way (by covariates)

• What variables to use in a model?

• Identify coding errors in the data

• Extensions

• Structure from NLP tools

• Interactive labeling and keyword query building [King et al 2014]

Prototype available:  http://brenocon.com/mte
Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran’s nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran’s program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.

Antigovernment Protesters Try to Shut Down Bangkok

BANGKOK — Antigovernment protesters seeking to block next month’s elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. “We have to shut down Bangkok,” said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. “This is our last resort.” By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.
Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran’s nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran’s program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.

Antigovernment Protesters Try to Shut Down Bangkok

BANGKOK — Antigovernment protesters seeking to block next month’s elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. “We have to shut down Bangkok,” said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. “This is our last resort.” By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.
Event data through knowledge engineering

[Schrodt 1994, Leetaru and Schrodt 2013]

Event classes
(~200)

Dictionary:
Verb patterns per event class
(~15000)

Extract events from news text

03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

191 - Impose blockade, restrict movement
not allow to enter ; mj 02 aug 2006
barred travel
block traffic from ; ab 17 nov 2005
block road ; hux 1/7/98

Issue: Hard to maintain and adapt to new domains
Our inference process

Data: twenty years of news articles

Natural Language Processing

Event phrases of actor interactions

Probabilistic Graphical Model
Purely from textual data, jointly learns both

(1) Event class dictionaries

“diplomacy”
arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to,

“verbal conflict”
accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say—ccomp come from, say—ccomp, suspect, slam, accuse government—poss,

“material conflict”
kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about—pobj troops in, kill, have troops

[O’Connor, Stewart, and Smith, 2013]
Event phrases

“ISR meet with PSE”

\[ P(w = \text{“meet with”} \mid t, s = \text{ISR}, r = \text{PSE}) \]

Too sparse for human interpretability
Do word semantics cluster on social context?

$s=$ISR, $r=$PSE

$t=$Jul 15-21, 2002
say $\leftarrow$ccomp be to
release to
take control of
occupy
wound in
scuffle with
be $\leftarrow$xcomp meet
meet with
meet with
arrest

$t=$Jul 3-9, 2006
commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

$s=$USA, $r=$FRA

$t=$Feb 2-8, 1998
travel $\leftarrow$xcomp meet with
consider
meet with
meet with
meet with

$t=$Dec 22-28, 2003
release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict
Do word semantics cluster on social context?

Clustering approach: Mixed-membership models ("topic models," "admixtures")

Contextual event class probabilities

\[ \theta_{s,r,t} = \begin{cases} 1 & \text{ISR, r=PSE} \\ 2 & \text{USA, r=FRA} \end{cases} \]

Event class dictionaries

\[ \phi_1 \quad \phi_2 \]

- agree with, arrest, be \(<xcomp> meet\), carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say \(<xcomp> be to\), scuffle with, shell, start around, strike, take control of, travel \(<xcomp> meet with\), welcome, welcome by, win, win from, wound in
Model

Event prior models

M1: independent contexts
M2: temporal smoothing

[Blei and Lafferty 2006, Quinn and Martin 2002]

\( \beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2) \)

\( \eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \text{Diag}[\sigma_1^2 \ldots \sigma_K^2]) \)

\( (\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k}) \)

\( z \sim \text{Mult}(\theta_{s,r,t}) \)

\( w \sim \text{Mult}(\Phi \theta_{s,r,t}) \)

\( \phi_k \sim \text{Dir}(b) \)

K=100 \rightarrow 80 \text{ million parameters}
Learning: blocked Gibbs sampling

\[ p(\beta, (\eta, \theta), \sigma_1^2 \ldots \sigma_K^2, z, \phi, b \mid w) \]

\[ \beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I} \tau^2) \]
\[ \eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \text{Diag}[\sigma_1^2 \ldots \sigma_K^2]) \]
\[ (\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k}) \]
\[ z \sim \text{Mult}(\theta_{s,r,t}) \]
\[ w \sim \text{Mult}(\phi_z) \]
\[ \phi_k \sim \text{Dir}(b) \]
Learning: blocked Gibbs sampling

\[ p(\beta, (\eta, \theta), \sigma_1^2 \ldots \sigma_K^2, z, \phi, b \mid w) \]

**Linear dynamical system**
Forward filter backward sampler (FFBS)
[Carter and Kohn 1994, West and Harrison 1997]

**Logistic normal**
Metropolis-within-Gibbs, Laplace approximation proposal
[Hoff 2003]

**Dirichlet-multinomial**
Collapsed sampling
[Griffiths and Steyvers 2005]

\[ \beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I} \tau^2) \]
\[ \eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \text{Diag}[\sigma_1^2 \ldots \sigma_K^2]) \]
\[ (\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k}) \]
\[ z \sim \text{Mult}(\theta_{s,r,t}) \]
\[ w \sim \text{Mult}(\phi_z) \]
\[ \phi_k \sim \text{Dir}(b) \]

Conjugate normal
Slice sampling
[Neal 2003]
Event classes: word posteriors

Most probable phrases in $\phi_k$

arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say $\leftarrow$ ccomp come from, say $\leftarrow$ ccomp, suspect, slam, accuse government $\leftarrow$ poss, accuse agency $\leftarrow$ poss, criticize, identify

kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about $\leftarrow$ pobj troops in, kill, have troops $\leftarrow$ partmod station in, station in, injure in, invade, shoot in
**Event classes: word posteriors**

- **“diplomacy”**
  - arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

- **“verbal conflict”**
  - accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say ← ccomp come from, say ← ccomp, suspect, slam, accuse government ← poss, accuse agency ← poss, criticize, identify

- **“material conflict”**
  - kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ← pobj troops in, kill, have troops ← partmod station in, station in, injure in, invade, shoot in

Most probable phrases in $\phi_k$
Case study

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with

**Israeli–Palestinian Diplomacy**

- **A**: Israel-Jordan Peace Treaty
- **B**: Hebron Protocol
- **C**: U.S. Calls for West Bank Withdrawal
- **D**: Deadlines for Wye River Peace Accord
- **E**: Negotiations in Mecca
- **F**: Annapolis Conference

![Graph showing diplomatic activities from 1994 to 2007](image-url)
Evaluations

Data generation process

Society (Social Attributes)

Writing (Text Generator)

Text Data (Text)

Inference & collection

Event datasets from political science (Previous work)

Evaluation: do these correlate?

Inference

Event datasets

Israeli–Palestinian Diplomacy


Evaluation: do these correlate?

Inference
Evaluations

Lexicon / Ontology reconstruction

Real-world conflict reconstruction
Geographic lexical variation in Twitter

[Geographic topic model]

$r \sim \pi$

$(\text{lat, lon}) \sim N(\mu_r, \Sigma_r)$

$\theta \sim \text{Dir}(\alpha)$

$z \sim \theta$

$w \sim \exp(\eta_{zr})$

$\phi_k \sim N(\bar{a}, b^2 \text{I})$

$\eta_{kj} \sim N(\bar{\phi}_k, s_k^2 \text{I})$

User’s locations from DPMM Gaussian mixture

User’s topics

have regional variants

---

**“basketball”**

PISTONS KOBE
LAKERS game
DUKE NBA
CAVS STUCKEY
JETS KNICKS

**“popular music”**

album music
beats artist video
#LAKERS
ITUNES tour
produced vol

**“daily life”**

tonight shop
weekend getting
ready discount
waiting jam

**“emoticons”**

:) haha :d :(:) :p
xd :/: hahaha
hahah

**“chit chat”**

lol smh jk yea
wyd coo ima
wassup somethin

---

**Boston**

CELTICS victory
BOSTON
CHARLOTTE

playing daughter
PEARL alive war comp

BOSTON

;p gna loveee

es e exam suttin
sippin

**N. California**

THUNDER
KINGS GIANTS
pimp trees clap

SIMON dl
mountain seee

6am OAKLAND

pues hella koo
SAN fckn

hella flirt hut
iono OAKLAND

---

[Geographic topic model image]

[Map of Twitter user locations and topics]

[Eisenstein, O’Connor, Smith, Xing 2010]
Social determinants of language change

[Eisenstein, O’Connor, Smith, Xing 2012 and in review]

Test sociolinguistic theories of how linguistic innovations diffuse
U.S. Census data
200 regions, 2600 words, 165 timesteps = 85M parameters

\[
\begin{align*}
n_{w,r,t} &\sim \text{Binom}(N_{r,t}, \sigma(\nu_w + \tau_{r,t} + \eta_{w,*}, t + \eta_{w,r}, t)) \\
\eta_{w,t} &\sim \text{Normal}(A\eta_{w,t-1}, \Gamma) \\
A &\text{ autoregressive coefficients (size } R \times R)\end{align*}
\]
Social Media NLP
Part-of-speech tagger for Twitter

Example

```
 ikr smh he asked fir yo last name
```

HMM word cluster (features for CRF tagger)

```
yeah yea nah naw yeahh nooo yeh noo noooo yeaa ikr nvm yeahhh nahh nooooo yh yeaaa yeaah yupp naa yeahhhh yeaaahiknow werd noes nahhh naww yeeaaaah yeahhhhh naaa naah nawl nawww yehh ino yeaaaaa yeeah yeeeah wordd yeaaahh nahhhh naaah yeahhhhhh yeaaaaah naaaa yeeeeeah nall yeaaaaaa
```

http://www.ark.cs.cmu.edu/TweetNLP/

[Gimpel, Schneider, O’Connor, Das, Mills, Eisenstein, Heilman, Yogatama, Smith, 2011]
[Owoputi, O’Connor, Dyer, Gimpel, Schneider, Smith, 2013]
Text Analysis for Social Science

- Tools for discovery and measurement
- Social, spatial, temporal context
- Probabilistic models
- A little bit of NLP can go a long way

Future work
- Text visualization / exploration tools
- Semantics: belief structures from text
- Incorporate a-priori knowledge
- Causal inference