Stream Characterization from Content

Allen Gorin
Human Language Technology Research
U.S. DoD, Fort Meade MD
a.gorin@ieee.org
Collaborators

Carey Priebe (JHU)
John Grothendieck (BBN)

Nash Borges
John Conroy
Glen Coppersmith
Rich Cox
Mike Decerbo

Dave Marchette
Alan McCree
Youngser Park
Alison Stevens
Jerry Wright
Outline

• Motivation
• HLT Research Issues
• Joint model of content in context
• Experiments on speech using Switchboard
• Experiments on text using Enron
Environmental Awareness

Focus of Attention

Peripheral glances
Environmental Awareness: 
Focus of Attention *plus* Peripheral ‘Vision’

Lower resolution and lossy compression

Enables change and anomaly detection
Coping with Information Overload

- Mature: External Metadata
- Emerging: Metacontent

Pick out the good stuff
Filter and Select
Boil it down
Stream Characterization

- language
- speaker
- topic
Analytic Questions

• Is the information environment stable?
  – describe environment
  – lossy compression

• Did something change?
  – Where? What?
Outline

• Motivation
• HLT Research Issues
• Joint model of content in context
• Experiments on speech using Switchboard
• Experiments on text using Enron
HLT Research Issues

• **Focus on stream statistics**
  – Rather than on individual documents
  – E.g. *Language Characterization* (McCree)
  – Classifier output is *biased* and noisy (Grothendieck)
  – Piece-wise stationary segments (Wright)

• **Content has associated meta-data**
  – Better living through *content in context*
  – Theory, simulations and experiments
  – with Priebe, Grothendieck, et al
Experimental Corpora

• Enron corpus of emails
  – 500K emails over 189 weeks from DoJ/CMU
  – 184 communicants
  – 32 topics as defined by LDC

• Switchboard corpus of spoken dialogs
  – 2500 topical dialogs
  – between pairs of 500 speakers
  – speaker demographics
Outline

• Motivation
• HLT Research Issues
• Joint model of content in context
• Experiments on speech using Switchboard
• Experiments on text using Enron
Joint model of content in context

• Consider a set of communication events

\[ M = \{ z_i = (u_i, v_i, t_i, x_i) \} \in \mathcal{M} \]

• An event in \( M \) is \( z_i \in V \times V \times R^+ \times \Xi \)
  — representing \((to, from, time, content)\)

• A time window defines a graph with content-attributed edges

• Attribution functions \( h_V \) and \( h_E \) to further color vertices and edges
Examples from Enron Corpus
(high-dimensional and heterogeneous features)

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Sender</th>
<th>Receiver</th>
<th>Sender’s Rank</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-01-02</td>
<td>04:15:00</td>
<td>steven.k</td>
<td>jeff.d</td>
<td>Vice President</td>
<td>(1) California Analysis</td>
</tr>
<tr>
<td>2001-02-09</td>
<td>13:49:09</td>
<td>louise.k</td>
<td>andy.z</td>
<td>President</td>
<td>(9) Daily Business</td>
</tr>
<tr>
<td>2001-02-16</td>
<td>21:06:00</td>
<td>drew.f</td>
<td>jeff.d</td>
<td>Vice President</td>
<td>(5) California Enron</td>
</tr>
<tr>
<td>2001-02-26</td>
<td>22:30:00</td>
<td>james.s</td>
<td>john.l</td>
<td>Vice President</td>
<td>(14) Energy Newsfeed</td>
</tr>
<tr>
<td>2001-03-01</td>
<td>07:54:00</td>
<td>diana.s</td>
<td>kate.s</td>
<td>Trader</td>
<td>(5) California Enron</td>
</tr>
<tr>
<td>2001-04-06</td>
<td>05:15:00</td>
<td>mike.g</td>
<td>john.l</td>
<td>Manager</td>
<td>(7) Newsfeed California</td>
</tr>
<tr>
<td>2001-04-16</td>
<td>06:12:00</td>
<td>richard.s</td>
<td>steven.k</td>
<td>Vice President</td>
<td>(9) Daily Business</td>
</tr>
<tr>
<td>2001-05-11</td>
<td>16:02:00</td>
<td>andy.z</td>
<td>john.l</td>
<td>Vice President</td>
<td>(11) Enron Online</td>
</tr>
<tr>
<td>2001-06-27</td>
<td>17:44:24</td>
<td>s..s</td>
<td>geoff.s</td>
<td>Vice President</td>
<td>(9) Daily Business</td>
</tr>
<tr>
<td>2001-09-05</td>
<td>14:36:53</td>
<td>geoff.s</td>
<td>louise.k</td>
<td>Director</td>
<td>(12) Enrononline Daily</td>
</tr>
<tr>
<td>2001-09-15</td>
<td>20:51:20</td>
<td>m..p</td>
<td>louise.k</td>
<td>Vice President</td>
<td>(12) Enrononline Daily</td>
</tr>
<tr>
<td>2001-10-04</td>
<td>14:19:16</td>
<td>john.l</td>
<td>louise.k</td>
<td>CEO</td>
<td>(11) Enron Online</td>
</tr>
<tr>
<td>2001-10-05</td>
<td>18:49:05</td>
<td>j..k</td>
<td>richard.s</td>
<td>Vice President</td>
<td>(9) Daily Business</td>
</tr>
<tr>
<td>2001-10-08</td>
<td>17:50:19</td>
<td>shelley.c</td>
<td>darrell.s</td>
<td>Vice President</td>
<td>(1) California Analysis</td>
</tr>
</tbody>
</table>
SwitchBoard Communications Graph

Vertex ~ speakers
Edges ~ dialogs
Joint Model of Content and Context via Attributed Graphs

• **Edge attributes**
  – Content-derived meta-data (a.k.a. *meta-content*)
  – E.g. topic id, ASR, turn-taking behavior

• **Vertex attributes**
  – *External meta-data* about speaker
  – E.g. demographics such as age, gender, education, ...
  – *Graph-derived* meta-data
  – E.g. vertex degree ~ willingness to communicate
Outline

• Motivation
• HLT Research Issues
• Joint model of content in context
• Experiments on speech using Switchboard
• Experiments on text using Enron
Joint Model of Content and Context

• Random Attributed Graph
  – Provides a joint model of content and context

• In Switchboard
  – Content is an attribute of an edge (dialog)
  – Consider turn-taking behavior in the dialog
  – Context is an attribute of the vertices (speakers)
  – Consider age, education, gender of speakers

• Joint model enables inference of
  – Unobserved demographic distribution
  – From observed turn-taking behavior
Models of Turn-Taking Behavior

• Turn-taking behavior has predictive power
  – for speaker ID (Jones)
  – for speaker traits in meeting room data (Lakowski)
  – for social roles and networks (Pentland)

• Joint model of vertex, edge attributes and graph
  – social correlates of turn-taking behavior
  – Grothendieck and Borges
  – experiment to exploit joint distribution
  – observed meta-content (turn-taking)
  – estimate unseen demographic distributions
Turn-taking Behavior Model
derived from SAD

<table>
<thead>
<tr>
<th>Side 1: $S_1(t)$</th>
<th>I</th>
<th>A</th>
<th>I</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Side 2: $S_2(t)$</td>
<td>A</td>
<td>I</td>
<td>A</td>
<td>I</td>
</tr>
<tr>
<td>Dialog State: $S(t)$</td>
<td>IA</td>
<td>II</td>
<td>AI</td>
<td>AA</td>
</tr>
</tbody>
</table>

A = active
I = inactive
Semi-Markov Model of Turn-Taking Behavior
Latent Classes of Turn-Taking Behavior

- Train turn-taking model from Switchboard corpus
- First-order partition via *divisive clustering*
  - E.g., *Style 0* has more and longer II (both silent)
  - E.g., *Style 1* has more and longer AA (both active)
- Classify each dialog as style 0 or 1
  - Edge attribute (meta-content)
- Classify each speaker as having style 0 or 1
  - Vertex attribute induced from edge attributes
Enriching vertex attributes with edge meta-content and graph meta-data

- $X = \text{external meta-data on speaker } v$
- $Y = \text{conversation turn-taking style}$
- $T(Y) = \text{turn-taking style of speaker } v$
- $\#V = \text{number of conversations including speaker } v$

Diagram:

- Vertex $V$ with attributes $X_1, X_2, X_3, \ldots, T(Y), \#V$
- Edges $Y_1, Y_2, Y_3$ connecting vertices.
Experimental Evaluation

• E.g., overall ratio of male:female is 1:1
  – speakers with *TT style 0 have ratio 2:1*

• Have joint distribution of content and context
  – exploit *observed content* (turn-taking behavior)
  – to *estimate unobserved context* (demographic mix)

• *Experiment*: create speaker sets with mixture proportion \( v \) of style 0, for \( v \) in \([0,1]\)

• Result: across all mixtures \( v \) of styles,
  – predict proportions of age, education, gender, ...
  – yields RMS error \( \sim 0.1 \)
Classic Problems in DSP

• Estimate characteristic parameters
  — Oppenheim (1975)

• To detect a signal in background noise
  — Van Trees (1968)

• Motivates initial focus on change/anomaly detection
Better Living through Content in Context

• *Information Exploitation* = statistical inference

• *Better* = more powerful statistical test
  – *for* change/anomaly detection

• Some results to date
  – Theorem that joint *can* be more powerful
  – Simulation experiments
  – Proof-of-concept experiment on Enron Corpus
Outline

- Motivation
- HLT Research Issues
- Joint model of content in context
- Experiments on speech using Switchboard
- Experiments on text using Enron
Time Series of Attributed Graphs

Generated from observations of some random attributed graph?
Change detection in a time series of Graphs

Homogeneous Chatter Group

Anomalous Chatter Group
Detecting ‘Signal’ in ‘Noise’ - models and theory

\[ G_N(t) \rightarrow G \rightarrow G_S(t) + G_N(t) \]

\( G \) is a probability distribution over attributed graphs

\[ G_S(t) \]
Random Attributed Graphs

• Let’s work through an example with a very simple model of content and context
• Existence of an edge between two vertices is IID Bernoulli with probability $p$
• Content topic (on each edge) is IID Bernoulli with probability $\theta$
• Change detection via testing candidate anomaly (alternative) versus history (null)
Null Hypothesis (noise):
an attributed Erdos-Renyi Graph

Random Graph $\text{ERC}(N, p, \Theta)$

$N = \# \text{ vertices in the graph}$

$p = \text{ probability of an edge}$

Each edge labeled

- with topic 0 or 1
- with $\Theta = \text{ probability of topic 1}$
Alternative Hypothesis (noise + signal): an ERC subgraph with different parameters

Random Graph

$K(N, p, \Theta, M, q, \Theta')$

$N = \#$ vertices in whole graph
$p = \text{prob(edge)}$ in kidney
$\Theta = \text{topic parameter in kidney}$
$M = \#$ vertices in egg
$q = \text{prob(edge)}$ in egg
$\Theta' = \text{topic parameter in egg}$
A statistical test based on fusion of externals and content can be more powerful than a test based on externals alone or content alone.

(Grothendieck and Priebe)
Proof by Construction

- \( T_G = \) # of graph edges
- \( T_C = \) # of graph edges attributed with topic 1
- \( T = 0.5 \ T_G + 0.5 \ T_C \)
- Test for change from homogeneous null graph:
  - Power of test based upon \( T_G \) is \( \beta_G \)
  - Power of test based upon \( T_C \) is \( \beta_C \)
  - Power of test based upon \( T \) is \( \beta \)
- For tests with false alarm rate \( \alpha = 0.05 \),
  - gray-scale plot of power difference \( \Delta = \beta - \max(\beta_G, \beta_C) \)
Power Difference: $\Delta = \beta - \max(\beta_C, \beta_G)$

$\Delta(\Theta', q)$ depends on the parameters of the anomalous chatter group

$p = 0.5$

$\Theta = 0.5$

$q = \text{subgraph connectivity}$

$\Theta' = \text{subgraph topic}$

$\text{Grayscale} = \Delta (\Theta', q)$
Detecting ‘Signal’ in Empirical ‘Noise’

\[ G_N(t) \rightarrow \bigoplus \rightarrow G_S(t) + G_N(t) \]

*Enron Data*

*Model*
Enron Experiment

- Select a stationary region of test statistics for Enron
- Estimate empirical null $G_N(t)$ from that region
- Add ‘signal’ via model $G_S(t)$ which injects egg
- Similar experimental results on power difference!
Conclusions

• Better living through content in context
  – modeled via random attributed graphs
• Better = more powerful statistical inference
• Joint model of content and context can be more powerful for many inference tasks
• Theorem for change/anomaly detection
• Proof of Concept Experiments
  – Inference of demographics from turn-taking behavior
  – Change/Anomaly detection
  – On Switchboard and Enron corpora
Acknowledgements

• Charles Wayne for
  – insights into communication graphs

• Deb Roy for
  – insights into content in context

• Sandy Pentland for
  – insights into social networks and communications
Some References


• Statistical Inference on Random Graphs: Fusion of Graph Features and Content, Grothendieck, Priebe, and Gorin, Computational Statistics and Data Analysis (2010)

• Statistical Inference on random attributed Graphs: Fusion of Graph Features and Content: An Experiment on Timeseries of Enron Graphs, Priebe et al, Computational Statistics and Data Analysis (2010).

• Social Correlates of Turn-taking Behavior, Grothendieck, Gorin, and Borges, [ICASSP 2009] , [full paper submitted]

• Towards Link Characterization from Content: Recovering Distributions from Classifier Output, Grothendieck and Gorin, IEEE Transactions on Speech and Audio, May 2008

• CoCITe – Coordinating Changes in Text, Wright and Grothendieck, to appear, IEEE Trans. on Knowledge and Data Engineering