Course basics

- Ling 795 Seminar in Linguistics
- T 7:00–9:40 AH-2131
- Instructor: Rob Malouf
- Email: rmalouf@mail.sdsu.edu
- Office hrs: BA 310A, Tues 4:30–5:30, Thurs 1:00–2:00, or by appointment
- Website: http://rohan.sdsu.edu/~malouf/ling795.html

Course basics

- Analyzing web texts
- Computer-Mediated Communication (CMC)
  - USENET
  - Email lists
  - MUDs/MOOs
  - Chat rooms
  - Discussion forums
  - Blogs

Course basics

- What is CMC like? How does it differ from other types of communication? Why?
- What challenges do web texts present for natural language processing?
- What can we do and what can't we do with web texts?

Requirements

- The final grade will be based on
  - Project proposal (<1 page) Feb 28 10%
  - Annotated bibliography March 21 10%
  - Data set April 4 10%
  - Final project May 18 50%
  - Class participation — 20%
Schedule

- Computer-mediated communication
- Genre, discourse analysis
- Collecting data
  - Spiders and web scraping, language identification, building ad hoc corpora
- Document processing
  - Lexical associations, classification and clustering, text mining
- Applications
  - Education, politics, how-to, marketing

Document processing

- Moving to deeper analysis of texts
- Information retrieval: find documents which are relevant to a query
- Information extraction: given a relevant document, pull out a relevant segment
- Question answering: given relevant segments, formulate the answer to a question
- Text mining: discover something new about the world which is not explicitly stated in any one text (or even known to the authors)

Document processing

- Edited (semi-)formal text
- Newspaper and magazines
- Academic papers
- Broadcast transcripts
- Personal and hobby sites
- Informal, personal, interactive texts
- Weblogs
- Chat rooms
- Discussion forums

Web texts

EDITORIAL: Minimum Wage

The Lufkin Daily News
Tuesday, January 03, 2006

It’s a given that it is virtually impossible to support a family on minimum wage, which has been set at $5.15 an hour since 1997. In fact, a couple, each earning minimum wage, doesn’t even approach the federal poverty level. Support appears to be growing for increasing the federal minimum wage, which in fact has been supplanted in 17 states and the District of Columbia by state laws establishing higher minimum wages than that set by Congress. And a number of states that haven’t set higher minimum wages plan to debate the issue this year.
Web texts

• LOL. If inflation won’t happen with minimum wage increases, then why not put minimum wage to a 100$ an hour? Don’t worry it won’t affect prices....the extra money for the wages will come out of the sky.
• What’s the matter rockhead? You don’t want to help the poor needy people? You rich bastard koolaid drinker!
• that was pretty stupid man honestly...grow up !
• I don’t have a kid brother

Web texts

• Informal web texts offer a view into on-going and past interactions
• Many linguistic and non-linguistic applications
• Marketing
  • Alternatives to conventional market research methods like poll, surveys, and focus groups
  • Viral marketing
  • Politics, pharmaceuticals

Web texts

• Generally, CMC falls somewhere between spoken and written language
• Orthographic representations of spoken language features
  Al, hahahahahahahahahahahahahahahahahahahahahahahahahahahhaahaaa *sniff* waaaaaaaaaaaaaaaaaaaaaaaaahhhhh
  I laughed, i cried. . . . that post was GREAT! :-) Amusedly,
  -Mirth-
Web texts

- Verbal ‘play’

  sssssssssssss *passes joint to kang*
  ..............
  thanx dude *puff* *hold* ..............
  >:-)
  ............
  kang exhale.. you will die :-)
  *exhale* 
  ;)

Web texts

- Non-verbal play

  :|
  :\
  hehehe
  ...
  that was great
  :/
  :
  hehehehehe
  *exhale*
  :0
  :|  :|  :\ssss :)
  hheeeheee
  :-Q  :|  :|  :\ssss :)

Web texts

- CMC has features that are unlike either spoken or written language

- Telegraphic style

  Les1: as it stands now, meeting on weds?
  Les2: instead of tues
  Brian1: idiot Hess seemed to think you were there tues morning
  Brian2: thot that mtg from 9 to 10 would solve
  Brian3: if you not in ny I’m going to have mtg changed to wedne.

Web texts

- Turn taking and coherence in synchronous CMC

  [1] <ashna> hi jatt
  [2] *** Signoff: puja (EOF From client)
  [3] <Dave-G> kally i was only joking around
  [5] <kally> dave-g it was funny
  [6] <ashna> how are u jatt
  [7] <LUCKMAN> ssa all12
  [8] <Dave-G> kally you da woman!
  [9] <Jatt> ashna: do we know eachother?. I’m ok how are you
  [10] *** LUCKMAN has left channel #PUNJAB
  [11] *** LUCKMAN has joined channel #punjab
  [12] <kally> dave-g good stuff:
  [14] <ashna> jatt no we don’t know each other, i fine
  [15] <Jatt> ashna: where r ya from?
Web texts

• Turn taking and coherence

[1] <ashna> hi jatt
[2] *** Signoff: puja (EOF From client)
[3] <Dave-G> kally i was only joking around
[5] <kally> dave-g it was funny
[6] <ashna> how are u jatt
[7] <LUCKMAN> ssa all12
[8] <Dave-G> kally you da woman!
[9] <Jatt> ashna: do we know eachother?. i'm ok how are you
[10] *** LUCKMAN has left channel #PUNJAB
[11] *** LUCKMAN has joined channel #punjab
[12] <kally> dave-g good stuff:)
[14] <ashna> jatt no we don't know each other, i fine
[15] <Jatt> ashna: where r ya from?

Web texts

• Coherence in asynchronous CMC

On 14 Jan 2006 11:55:06 -0800, "Marc Adler" <marc.adler@gmail.com> wrote:
> http://www.volkskrant.nl/buitenland/1137218431379.html
> "Ook is niet duidelijk of de premier al volledig uit het kunstmatige coma is gehaald."
> Shouldn't that be "Ook is het niet duidelijk..."?

Both are possible.

Miguel Carrasquer Vidal
mcv@wxs.nl

Web texts

• Turn taking and coherence

[1] <ashna> hi jatt
[2] *** Signoff: puja (EOF From client)
[3] <Dave-G> kally i was only joking around
[5] <kally> dave-g it was funny
[6] <ashna> how are u jatt
[7] <LUCKMAN> ssa all12
[8] <Dave-G> kally you da woman!
[9] <Jatt> ashna: do we know eachother?. i'm ok how are you
[10] *** LUCKMAN has left channel #PUNJAB
[11] *** LUCKMAN has joined channel #punjab
[12] <kally> dave-g good stuff:)
[14] <ashna> jatt no we don't know each other, i fine
[15] <Jatt> ashna: where r ya from?

Web texts

• Coherence in asynchronous CMC

On 14 Jan 2006 11:55:06 -0800, "Marc Adler" <marc.adler@gmail.com> wrote:
> http://www.volkskrant.nl/buitenland/1137218431379.html
> "Ook is niet duidelijk of de premier al volledig uit het kunstmatige coma is gehaald."
> Shouldn't that be "Ook is het niet duidelijk..."?

Both are possible.

Miguel Carrasquer Vidal
mcv@wxs.nl
Readings

- Week 1: Introduction
  Danet 1998, Herring 2003
- Week 2: Qualitative genre studies
- Week 3: Quantitative genre studies
  Collot and Bellmore 1996, Yates 1996
- Week 4: Virtual communities
  Ellis et al. 2005, Burnett 2000, Burnett and Buerkle 2004

Homework

- Do readings
- Register on class blog
- Find a virtual community and post a description

User categorization

- Self-selected population can make interpreting results difficult
- Guessing demographic categories of authors can be important
  - Age
  - Gender
  - Education
  - Disease status
  - Political orientation
User categorization

- Given the posts a user has made on a political forum, can we guess their political orientation?
- U.S. Politics (Mullen and Malouf 2006)
  - politics.com
  - 5,246,230 words in 77,854 posts
  - 408 distinct users
  - 77 users (19%) with a single post
  - second most prolific had 3,801 posts
  - most prolific poster had 6,885 posts

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td>88</td>
</tr>
<tr>
<td>Republican</td>
<td>53</td>
</tr>
<tr>
<td>Conservative</td>
<td>30</td>
</tr>
<tr>
<td>R-fringe</td>
<td>5</td>
</tr>
<tr>
<td>Center</td>
<td>40</td>
</tr>
<tr>
<td>Centrist</td>
<td>7</td>
</tr>
<tr>
<td>Independent</td>
<td>33</td>
</tr>
<tr>
<td>Left</td>
<td>107</td>
</tr>
<tr>
<td>Democrat</td>
<td>62</td>
</tr>
<tr>
<td>Liberal</td>
<td>28</td>
</tr>
<tr>
<td>L-fringe</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>33</td>
</tr>
<tr>
<td>Libertarian</td>
<td>22</td>
</tr>
<tr>
<td>Green</td>
<td>11</td>
</tr>
<tr>
<td>Unknown</td>
<td>151</td>
</tr>
</tbody>
</table>

Sentiment analysis

- One dimensional summary (thumbs up or down?) of a corpus of relevant documents
- Turney (2002) uses a simple “PMI-IR” method
- Tag the review and extract noun phrases (by a simple regular expression)
- Identify the semantic orientation (SO) of each noun phrase by PMI-IR
  - Find documents on the web which contain the target phrase near either the word excellent or the word poor (IR)
- Compare the pointwise mutual information (PMI) of each phrase with excellent and poor
### Sentiment analysis

- **Pointwise mutual information** (PMI) is an information theoretic measure of how much two events tend to occur together.

- For two events $x$ and $y$:

  $$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

- The semantic orientation (SO) of a term is:

  $$SO(w) = PMI(w, \text{excellent}) - PMI(w, \text{poor})$$

### Sentiment analysis

- The SO of a document is the average of the SOs of the phrases.

- For example, in bank reviews:

<table>
<thead>
<tr>
<th>Word</th>
<th>SO</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct deposit</td>
<td>1.288</td>
<td>positive</td>
</tr>
<tr>
<td>small part</td>
<td>0.053</td>
<td>neutral</td>
</tr>
<tr>
<td>unethical practices</td>
<td>-8.484</td>
<td>negative</td>
</tr>
</tbody>
</table>

- The system was evaluated using reviews collected from epinions.com.

- Overall accuracy was 74.39%.

### Sentiment analysis

- In principle, the same method could be used for any one-dimensional classification (liberal vs. conservative, cheap vs. expensive, etc.).

- We measure political SO using PMI with the terms *liberal* and *conservative*:

  $$SO(w) = PMI(w, \text{liberal}) - PMI(w, \text{conservative})$$

- PMI scores for words were computed based on counts from the 180 million word Reuters news corpus (RCV1).
Sentiment analysis

- We have text for 183 users who can be classed as either LEFT (Democrat, liberal, L-fringe) or RIGHT (Republican, conservative, R-fringe)
- A simple strategy of assigning the most frequent label (LEFT) to all users gives us 51.91% accuracy
- Turney-style sentiment analysis gives 48.09% accuracy (worse than guessing!)
- Possible problems
  - RCV1 as a data resource (much smaller than WWW, edited text, collected in 1996 and 1997)
  - Applicability of the algorithm

Classifying users

- We can also approach this as a text classification problem
- Typical text classification systems use machine learning to construct a model of a set of labeled texts
- Used for automatic indexing, information retrieval tasks, and message routing
- ‘Naive Bayes’ classifiers use a simple, robust method for learning classification rules from labeled data

Classifying users

- Naive Bayes classifier (Maron 1961)
  - Language users randomly produce words with fixed probabilities determined by category \( P(w|c) \)
  - Probability of a text is the product of the word probabilities:
    \[
    P(\text{text}|c) = \prod P(w_i|c)
    \]
  - Work backwards from a text to find the most likely category:
    \[
    \hat{c} = \arg\max_c P(c|\text{text})
    = \arg\max_c P(c) \prod P(w_i|c)
    \]
Classifying users

- Skewed distribution of posting frequency
- Repeat experiments with users who posted 20 or more times (50 each from LEFT and RIGHT)
- Baseline is now 50.00%
- NB using 50 randomly selected users gives 52.00% accuracy
- NB using 50 frequent posters gives 61.38% accuracy
- Spelling correction increases this to 64.48%, though this difference is not significant

Classifying users

- Accuracy very dependent on sample size
- A larger sample would increase the number of well represented users
- But it would also increase the number of poorly represented users — there’s no escaping the long tail!
- Sometimes we are mainly interested in frequent posters, so this isn’t always bad news

Classifying users

- Vocabulary is often tied to an issue rather than a position (Argawal, et al. 2003)
  - minimum wage, gun control, abortion
- We can use a modified form of PMI to find words which are most closely associated with particular orientations

\[ PMI(x, y) = \frac{P(x, y)}{P(x)P(y)} \log \frac{P(x, y)}{P(x)P(y)} \]

Classifying users

- LEFT
  - bush, iraq, administration, american, president, security, war, pm, report, united, dixie, bush’s, iraqi, george, social, date, staff, intelligence, toby, sergeant, americans, house, national, told, military, . . .
- RIGHT
  - nbsp, sr, amp, weapons, ha, saddam, desh, liberals, destruction, mass, maineman, liberal, democrat, bil, clinton, democrats, understand, author, nation, threat, point, sen, mbl, hussein, date, . . .
Citations

- Efron (2004) predicted the orientation of political blogs using a modified version of Turney's method that was sensitive to co-citation patterns
- Bloggers tend overwhelmingly to link to like-minded sites (Adamic and Glance 2005)
- Discussion group posts don't (usually) have citations to other web sites, but they do quote other posts
- Maintaining coherence

Citations

- Out of 41,605 posts by LEFT or RIGHT users, 4,583 cite another post by a different LEFT or RIGHT user
- For LEFT users, 71.4% of cited posts are by RIGHT users
- For RIGHT users, 82.3% of cited posts are by LEFT users
- This suggests a simple algorithm
  - For each user, count the number of LEFT and RIGHT users that they quote or are quoted by
  - Assign the user the orientation which they are least often linked to

Citations

- Assume the orientations of all other users is known
- This rule assigns the correct orientation to 78.43% of users with at least one linked post
- But, the overall accuracy is only 66.12%
- For frequent posters (20 or more posts), the accuracy for users with a linked post is 83.53%, and 79.38% overall