Words and their meaning

Some upcoming lectures:
- Word disambiguation
  - one word, multiple meanings
- Word clustering
  - multiple words, “same” meaning
- Collocations - this lecture
  - multiple words together, different meaning than than the sum of its parts
  - Simple measures on text, yielding interesting, insights into language, meaning, culture.

Today’s Main Points
- What is collocation?
- Why do people care?
- Three ways of finding them automatically.

Collocations
- An expression consisting of two or more words that correspond to some conventional way of saying things.
- Characterized by limited compositionality.
  - compositional: meaning of expression can be predicted by meaning of its parts.
  - “strong tea”, “rich in calcium”
  - “weapons of mass destruction”
  - “kick the bucket”, “hear it through the grapevine”

Collocations important for...
- Terminology extraction
  - Finding special phrases in technical domains
- Natural language generation
  - To make natural output
- Computational lexicography
  - To automatically identify phrases to be listed in a dictionary
- Parsing
  - To give preference to parses with natural collocations
- Study of social phenomena
  - Like the reinforcement of cultural stereotypes through language (Stubbs 1996)

Contextual Theory of Meaning
- In contrast with “structural linguistics”, which emphasizes abstractions, properties of sentences
- Contextual Theory of Meaning emphasizes the importance of context
  - context of the social setting (not idealized speaker)
  - context of discourse (not sentence in isolation)
  - context of surrounding words
  - Firth: “a word is characterized by the company it keeps”
- Example [Halliday]
  - “strong tea”, coffee, cigarettes
  - “powerful drugs”, heroin, cocaine
  - Important for idiomatically correct English, but also social implications of language use
Method #1
Frequency

80871 of the
58841 in the
26430 to the
21842 on the
21839 for the
18658 and the
16121 that the
15630 at the
15494 to be
13899 in a
13689 of a
13361 by the
13183 with the
12622 from the
11428 New York
10007 he said

Method #1
Frequency with POS Filter
AN, NN, AAN, ANN, NAN, NNN, NPN

11487 New York A N
7281 United States A N
5412 Los Angeles A N
3301 last year N N
3191 Saudi Arabia N N
2699 last week A N
2514 vice president A N
2378 Persian Gulf A N
2161 San Francisco N N
2106 President Bush N N
2001 Middle East A N
1942 Saddam Hussein N N
1867 Soviet Union A N
1850 White House A N
1633 United Nations A N
1328 oil prices N N
1210 next year A N
1074 chief executive A N
1073 real estate A N

Method #2
Mean and Variance

• Some collocations are not of adjacent words, but words in more flexible distance relationship
  – she knocked on his door
  – they knocked at the door
  – 100 women knocked on Donaldson’s door
  – a man knocked on the metal front door
• Not a constant distance relationship
• But enough evidence that “knock” is better than “hit”, “punch”, etc.

Method #2
Mean and Variance

Sentence: Stocks crash as rescue plan teeters.
Time-shifted bigrams:

1 2 3
stocks crash crash as rescue
... stocks rescue crash plan as teeters

• To ask about relationship between “stocks” and “crash”, gather many such pairs, and calculate the mean and variance of their offset.

\[
\text{mean} = \bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i \\
\text{variance} = \sigma^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n - 1}
\]

Method #2
Mean and Variance

Position of “strong” versus “opposition” (mean=-1.15, deviation=0.67)

Position of “strong” versus “support” (mean=-1.45, deviation=1.07)
Method #2
Mean and Variance

Position of "strong" versus "for" (mean=1.12, deviation=2.15)

Method #3
Likelihood Ratios

- Determine which of two probabilistic models is more appropriate for the data.
  - H1 = hypothesis of model 1
  - H2 = hypothesis of model 2

likelihood ratio = \( \log \left( \frac{L(H_1)}{L(H_2)} \right) \)

- Hypothesis 1: \( p(w_2|w_1) = p = p(w_2|\neg w_1) \)
- Hypothesis 2: \( p(w_2|w_2) = p_1 \neq p_2 = p(w_2|\neg w_1) \)

Data
- \( N = \) total count of all words
- \( c_1 = \) count of word 1
- \( c_2 = \) count of word 2
- \( c_{12} = \) count of bigram \( w_1w_2 \)

Method #3
Likelihood Ratios

- Determine which of two probabilistic models is more appropriate for the data.

\[
\log \left( \frac{L(H_1)}{L(H_2)} \right) = \log \left( \frac{b(c_{12}; c_1, p_1)}{b(c_{12}; c_1, p)} \right) \cdot \log \left( \frac{b(c_{2-c_{12}}; N-c_1, p_2)}{b(c_{2-c_{12}}; N-c_1, p)} \right)
\]

Method #3
Likelihood Ratio example data

<table>
<thead>
<tr>
<th>-2log 1</th>
<th>c1</th>
<th>c2</th>
<th>c12</th>
<th>w1</th>
<th>w2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1291</td>
<td>12593</td>
<td>932</td>
<td>150</td>
<td>most</td>
<td>powerful</td>
</tr>
<tr>
<td>89</td>
<td>379</td>
<td>932</td>
<td>10</td>
<td>politically</td>
<td>powerful</td>
</tr>
<tr>
<td>82</td>
<td>932</td>
<td>934</td>
<td>10</td>
<td>powerful</td>
<td>computers</td>
</tr>
<tr>
<td>80</td>
<td>932</td>
<td>3424</td>
<td>13</td>
<td>powerful</td>
<td>force</td>
</tr>
<tr>
<td>57</td>
<td>932</td>
<td>291</td>
<td>6</td>
<td>powerful</td>
<td>symbol</td>
</tr>
<tr>
<td>51</td>
<td>932</td>
<td>40</td>
<td>4</td>
<td>powerful</td>
<td>lobbies</td>
</tr>
<tr>
<td>51</td>
<td>171</td>
<td>932</td>
<td>5</td>
<td>economically</td>
<td>powerful</td>
</tr>
<tr>
<td>51</td>
<td>932</td>
<td>43</td>
<td>4</td>
<td>powerful</td>
<td>magnet</td>
</tr>
<tr>
<td>50</td>
<td>4458</td>
<td>932</td>
<td>10</td>
<td>less</td>
<td>powerful</td>
</tr>
<tr>
<td>50</td>
<td>6252</td>
<td>932</td>
<td>11</td>
<td>very</td>
<td>powerful</td>
</tr>
<tr>
<td>49</td>
<td>932</td>
<td>2064</td>
<td>8</td>
<td>powerful</td>
<td>position</td>
</tr>
<tr>
<td>48</td>
<td>932</td>
<td>591</td>
<td>6</td>
<td>powerful</td>
<td>machines</td>
</tr>
<tr>
<td>47</td>
<td>932</td>
<td>2359</td>
<td>8</td>
<td>powerful</td>
<td>computer</td>
</tr>
<tr>
<td>45</td>
<td>932</td>
<td>396</td>
<td>5</td>
<td>powerful</td>
<td>magnets</td>
</tr>
</tbody>
</table>

Collocation studies helping lexicography

- Want to help dictionary-writers bring out differences between "strong" and "powerful"
  - Understand meaning of a word by the company it keeps.
- Church and Hanks (1989) through statistical analysis concluded that it is a matter of intrinsic vs extrinsic quality
- "Strong" support from a demographic group, means committed, but may not have capability.
- "Powerful" supporter is one who actually has capability to change things.
- But also additional subtleties, helps us analyze cultural attitudes
  - "Strong tea" versus "powerful drugs"
Method #1
“strong” versus “powerful”

<table>
<thead>
<tr>
<th>w</th>
<th>C(strong,w)</th>
<th>w</th>
<th>C(powerful,w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>support</td>
<td>50</td>
<td>force</td>
<td>18</td>
</tr>
<tr>
<td>safely</td>
<td>22</td>
<td>computers</td>
<td>10</td>
</tr>
<tr>
<td>sales</td>
<td>21</td>
<td>position</td>
<td>8</td>
</tr>
<tr>
<td>opposition</td>
<td>19</td>
<td>man</td>
<td>8</td>
</tr>
<tr>
<td>shaking</td>
<td>18</td>
<td>computer</td>
<td>8</td>
</tr>
<tr>
<td>sense</td>
<td>18</td>
<td>man</td>
<td>7</td>
</tr>
<tr>
<td>message</td>
<td>15</td>
<td>symbol</td>
<td>6</td>
</tr>
<tr>
<td>defense</td>
<td>14</td>
<td>military</td>
<td>6</td>
</tr>
<tr>
<td>gains</td>
<td>13</td>
<td>country</td>
<td>6</td>
</tr>
<tr>
<td>criticism</td>
<td>13</td>
<td>weapons</td>
<td>5</td>
</tr>
<tr>
<td>possibility</td>
<td>11</td>
<td>post</td>
<td>5</td>
</tr>
<tr>
<td>feelings</td>
<td>11</td>
<td>people</td>
<td>5</td>
</tr>
<tr>
<td>demand</td>
<td>11</td>
<td>forces</td>
<td>5</td>
</tr>
<tr>
<td>challenges</td>
<td>11</td>
<td>chip</td>
<td>5</td>
</tr>
<tr>
<td>challenge</td>
<td>11</td>
<td>nation</td>
<td>5</td>
</tr>
<tr>
<td>case</td>
<td>10</td>
<td>Germany</td>
<td>5</td>
</tr>
<tr>
<td>supporter</td>
<td>10</td>
<td>senator’s</td>
<td>4</td>
</tr>
<tr>
<td>signal</td>
<td>9</td>
<td>neighbor</td>
<td>4</td>
</tr>
</tbody>
</table>

Likelihood Ratios across different corpora from different times

- Model1 = model for NYTimes 1989
- Model2 = model for NYTimes 1990

<table>
<thead>
<tr>
<th>Ratio</th>
<th>w1</th>
<th>w2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.024</td>
<td>Karim</td>
<td>Obaid</td>
</tr>
<tr>
<td>0.037</td>
<td>East</td>
<td>Berliners</td>
</tr>
<tr>
<td>0.039</td>
<td>Miss</td>
<td>Manners</td>
</tr>
<tr>
<td>0.041</td>
<td>HUD</td>
<td>officials</td>
</tr>
<tr>
<td>0.048</td>
<td>East</td>
<td>Germans</td>
</tr>
<tr>
<td>0.051</td>
<td>Prague</td>
<td>Spring</td>
</tr>
</tbody>
</table>

1989: Muslim cleric Sheik Abdul Krim Obeid abducted, disintegration of communist Eastern Europe, scandal in HUD, October 17 earthquake in San Francisco, Miss Manners no longer carried by NYTimes in 1990.