Chapter 5: **Collocations**

presented by Dustin Boswell

May 3, 2004

*from *Foundations of Statistical Natural Language Processing*
What is a Collocation?

- “An expression consisting of two or more words that correspond to some conventional way of saying things.” - Ch. 5 of FSNLP

- “Collocations of a given word are statements of the habitual or customary places of that word.” - Firth (1957)

- “A phrase that means more than the sum of its parts.” - Dustin

There is no exact definition.

Examples:

‘‘strong tea’’, ‘‘New York’’, ‘‘weapons of mass destruction’’, etc.
What is a Collocation?

Characteristics:

- **non-compositionality**
  Eg: white wine (wine isn’t white...)

- **non-substitutability**
  Eg: white yellow wine (doesn’t work)

- **non-modifiability**
  Eg: I have a (slimy?)frog in my throat

Non-Examples:

- of the
- doctor ... nurse
  (related words are simply co-occurrences)
Why care about Collocations?

- Sentence Parsing:
  Helps identify noun/verb phrases.

- Natural Language Generation & Translation:
  Avoid awkward output like
  ‘‘powerful tea’’ or ‘‘to take a decision’’

- Dictionary Building:
  Identify phrases that essentially act like individual words
How do we find Collocations?

- Counting frequencies of adjacent words
- Mutual Information between words
- Hypothesis Testing
  - t test
  - t test of differences
  - $\chi^2$ test
  - likelihood ratios
Counting Frequencies of Adjacent Words

- Method: Simply choose the most frequent adjacent pairs.
- Difficulty: prepositions are frequent.
- Results:

<table>
<thead>
<tr>
<th>$C(w^1 w^2)$</th>
<th>$w^1$</th>
<th>$w^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>80871</td>
<td>of</td>
<td>the</td>
</tr>
<tr>
<td>58841</td>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>26430</td>
<td>to</td>
<td>the</td>
</tr>
<tr>
<td>21842</td>
<td>on</td>
<td>the</td>
</tr>
<tr>
<td>21839</td>
<td>for</td>
<td>the</td>
</tr>
<tr>
<td>18568</td>
<td>and</td>
<td>the</td>
</tr>
</tbody>
</table>
Counting Frequencies of Adjacent Words

- Fix: only look for phrases with special “part of speech patterns”
- Results:

<table>
<thead>
<tr>
<th>$C(w^1 \ w^2)$</th>
<th>$w^1$</th>
<th>$w^2$</th>
<th>Tag Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>11487</td>
<td>New</td>
<td>York</td>
<td>A N</td>
</tr>
<tr>
<td>7261</td>
<td>United</td>
<td>States</td>
<td>A N</td>
</tr>
<tr>
<td>5412</td>
<td>Los</td>
<td>Angeles</td>
<td>N N</td>
</tr>
<tr>
<td>3301</td>
<td>last</td>
<td>year</td>
<td>A N</td>
</tr>
<tr>
<td>3191</td>
<td>Saudi</td>
<td>Arabia</td>
<td>N N</td>
</tr>
<tr>
<td>2699</td>
<td>last</td>
<td>week</td>
<td>A N</td>
</tr>
<tr>
<td>2514</td>
<td>vice</td>
<td>president</td>
<td>A N</td>
</tr>
<tr>
<td>2378</td>
<td>Persian</td>
<td>Gulf</td>
<td>A N</td>
</tr>
</tbody>
</table>
Counting Frequencies of Adjacent Words

Summary

+ Easy to implement

+ Gets the simple cases right

- Too sensitive to frequent bigrams. (strong man)

- Ignores rare bigrams
Pointwise Mutual Information

\[ PMI(w^1, w^2) = \log_2 \frac{P(w^1, w^2)}{P(w^1)P(w^2)} \]

\( w^1 \) and \( w^2 \) are values of random variables, like word tokens.

Don’t confuse this with the usual Mutual Information:

\[ MI(W^1; W^2) = E[ PMI(w^1, w^2) ] \]
\[ = \sum_{w^1, w^2} P(w^1, w^2) \log_2 \frac{P(w^1, w^2)}{P(w^1)P(w^2)} \]

\( W^1 \) and \( W^2 \) are random variables, like word locations.
Pointwise Mutual Information

- Method: Choose bigrams with highest $PMI = I(w^1, w^2)$
- Results:

<table>
<thead>
<tr>
<th>$I(w^1, w^2)$</th>
<th>$C(w^1)$</th>
<th>$C(w^2)$</th>
<th>$C(w^1 w^2)$</th>
<th>$w^1$</th>
<th>$w^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.38</td>
<td>42</td>
<td>20</td>
<td>20</td>
<td>Ayatollah</td>
<td>Ruhollah</td>
</tr>
<tr>
<td>17.98</td>
<td>41</td>
<td>27</td>
<td>20</td>
<td>Bette</td>
<td>Midler</td>
</tr>
<tr>
<td>16.31</td>
<td>30</td>
<td>117</td>
<td>20</td>
<td>Agatha</td>
<td>Christie</td>
</tr>
<tr>
<td>15.94</td>
<td>77</td>
<td>59</td>
<td>20</td>
<td>videocassette</td>
<td>recorder</td>
</tr>
<tr>
<td>15.19</td>
<td>24</td>
<td>320</td>
<td>20</td>
<td>unsalted</td>
<td>butter</td>
</tr>
<tr>
<td>1.09</td>
<td>14907</td>
<td>9017</td>
<td>20</td>
<td>first</td>
<td>made</td>
</tr>
<tr>
<td>1.01</td>
<td>13484</td>
<td>10570</td>
<td>20</td>
<td>over</td>
<td>many</td>
</tr>
<tr>
<td>0.53</td>
<td>14734</td>
<td>13478</td>
<td>20</td>
<td>into</td>
<td>them</td>
</tr>
<tr>
<td>0.46</td>
<td>14093</td>
<td>14776</td>
<td>20</td>
<td>like</td>
<td>people</td>
</tr>
<tr>
<td>0.29</td>
<td>15019</td>
<td>15629</td>
<td>20</td>
<td>time</td>
<td>last</td>
</tr>
</tbody>
</table>
Pointwise Mutual Information

- Difficulty: PMI is too sensitive to rare bigrams

<table>
<thead>
<tr>
<th>PMI</th>
<th>$w^1$</th>
<th>$w^2$</th>
<th>$w^1w^2$</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.46</td>
<td>106</td>
<td>6</td>
<td>1</td>
<td>Schwartz eschews</td>
</tr>
<tr>
<td>13.06</td>
<td>76</td>
<td>22</td>
<td>1</td>
<td>FIND GARDEN</td>
</tr>
<tr>
<td>11.25</td>
<td>22</td>
<td>267</td>
<td>1</td>
<td>fewest visits</td>
</tr>
<tr>
<td>8.97</td>
<td>43</td>
<td>663</td>
<td>1</td>
<td>Indonesian pieces</td>
</tr>
<tr>
<td>8.04</td>
<td>170</td>
<td>1917</td>
<td>6</td>
<td>marijuana growing</td>
</tr>
<tr>
<td>5.73</td>
<td>15828</td>
<td>51</td>
<td>3</td>
<td>new converts</td>
</tr>
</tbody>
</table>

- The problem is that $\frac{P(w^1, w^2)}{P(w^1)P(w^2)}$ easily becomes large for infrequent individual words.
Pointwise Mutual Information

- Possible Fixes:

  • Ignore bigrams that occur less than (say) 20 times

  • Redefine $PMI(w^1, w^2)' = C(w^1, w^2) \times PMI(w^1, w^2)$
Hypothesis Testing

- We really just want to know if words collocate more often than chance.

- Define a null hypothesis \( H_0 \) that says two words are independent:
  \[
P(w_1w_2) = P(w_1)P(w_2)
  \]

- If \((w^1, w^2)\) is a collocation, the hypothesis should be rejected to some significance level.
Hypothesis Testing: The $t$ test

- $H_0$: we have data coming from a normal distribution with mean $\mu$.

- *Data*: we observe $N$ points with sample mean $\bar{x}$, and sample variance $s^2$

- Compute the $t$ statistic: $t = \frac{\bar{x} - \mu}{\sqrt{s^2/N}}$

- If $t$ is greater than some threshold, reject $H_0$.

- The threshold (lookup in table) is 2.576 for large $N$ and 99.5% confidence.
The $t$ test Applied to Collocations

- Our statistic is the frequency of the bigram.
  - $\mu$ is the frequency assuming the words are independent
  - $\bar{x}$ is the observed frequency
  - $s^2$ is the observed variance (of this 'binomial')
  - The 'frequency' is really just the probability as calculated by simply counting:
    \[ P(w^1w^2) = \frac{\text{Count}(w^1w^2)}{N} \]
Hypothesis Testing: The $t$ test: Example

- We have a corpus with
  - $N = 14$ million words.
  - $C_{\text{new}} = 15,828$
  - $C_{\text{companies}} = 4675$

- $H_0$: ‘‘new companies’’ occurs with probability

\[
\mu = P_{\text{new}} P_{\text{companies}} \\
= \frac{15,828}{14\text{million}} \times \frac{4675}{14\text{million}} \\
\approx 3.6 \times 10^{-7}
\]
Hypothesis Testing: The $t$ test: Example

- **Data**: we observe 8 occurrences of new companies, so
  \[ \bar{x} = \frac{8}{14\text{million}} \approx 5.6 \times 10^{-7}. \]

- For a Bernoulli trial, $s^2 = p(1 - p) \approx p$ (for small $p$).

- Compute the $t$ statistic:
  \[
t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{5.6 \times 10^{-7} - 3.6 \times 10^{-7}}{\sqrt{\frac{5.6 \times 10^{-7}}{14\text{million}}}} \approx 1.00
  \]

- $t$ is not greater than 2.576, so new companies is not a collocation.
Hypothesis Testing: The $t$ test

- Method: Choose bigrams with highest t-statistic

- Results:

<table>
<thead>
<tr>
<th>$t$</th>
<th>$C(w^1)$</th>
<th>$C(w^2)$</th>
<th>$C(w^1 \cdot w^2)$</th>
<th>$w^1$</th>
<th>$w^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4721</td>
<td>42</td>
<td>20</td>
<td>20</td>
<td>Ayatollah</td>
<td>Ruhollah</td>
</tr>
<tr>
<td>4.4721</td>
<td>41</td>
<td>27</td>
<td>20</td>
<td>Bette</td>
<td>Midler</td>
</tr>
<tr>
<td>4.4720</td>
<td>30</td>
<td>117</td>
<td>20</td>
<td>Agatha</td>
<td>Christie</td>
</tr>
<tr>
<td>4.4720</td>
<td>77</td>
<td>59</td>
<td>20</td>
<td>videocassette</td>
<td>recorder</td>
</tr>
<tr>
<td>4.4720</td>
<td>24</td>
<td>320</td>
<td>20</td>
<td>unsalted</td>
<td>butter</td>
</tr>
<tr>
<td>2.3714</td>
<td>14907</td>
<td>9017</td>
<td>20</td>
<td>first</td>
<td>made</td>
</tr>
<tr>
<td>2.2446</td>
<td>13484</td>
<td>10570</td>
<td>20</td>
<td>over</td>
<td>many</td>
</tr>
<tr>
<td>1.3685</td>
<td>14734</td>
<td>13478</td>
<td>20</td>
<td>into</td>
<td>them</td>
</tr>
<tr>
<td>1.2176</td>
<td>14093</td>
<td>14776</td>
<td>20</td>
<td>like</td>
<td>people</td>
</tr>
<tr>
<td>0.8036</td>
<td>15019</td>
<td>15629</td>
<td>20</td>
<td>time</td>
<td>last</td>
</tr>
</tbody>
</table>

Table 5.6  Finding collocations: The $t$ test applied to 10 bigrams that occur with frequency 20.
Hypothesis Testing: The $t$ test of differences

- Consider two words with similar meaning:
  strong, and powerful
- We want to find collocates that best distinguish the usage of the two.
  Ex: powerful computer vs. strong computer
- $H_0$: we expect both pairs to occur just as frequently.
- Compute a similar $t$ statistic:
  $$t = \frac{x_1 - x_2}{\sqrt{s_1^2 + s_2^2}}$$
- Find words (like computer) with highest $t$ score.
Hypothesis Testing: The $t$ test of differences

- Results:

<table>
<thead>
<tr>
<th>$t$</th>
<th>$C(w)$</th>
<th>$C(strong\ \text{w})$</th>
<th>$C(powerful\ \text{w})$</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1622</td>
<td>933</td>
<td>0</td>
<td>10</td>
<td>computers</td>
</tr>
<tr>
<td>2.8284</td>
<td>2337</td>
<td>0</td>
<td>8</td>
<td>computer</td>
</tr>
<tr>
<td>2.4494</td>
<td>289</td>
<td>0</td>
<td>6</td>
<td>symbol</td>
</tr>
<tr>
<td>2.4494</td>
<td>588</td>
<td>0</td>
<td>6</td>
<td>machines</td>
</tr>
<tr>
<td>2.2360</td>
<td>2266</td>
<td>0</td>
<td>5</td>
<td>Germany</td>
</tr>
<tr>
<td>2.2360</td>
<td>3745</td>
<td>0</td>
<td>5</td>
<td>nation</td>
</tr>
<tr>
<td>7.0710</td>
<td>3685</td>
<td>50</td>
<td>0</td>
<td>support</td>
</tr>
<tr>
<td>6.3257</td>
<td>3616</td>
<td>58</td>
<td>7</td>
<td>enough</td>
</tr>
<tr>
<td>4.6904</td>
<td>986</td>
<td>22</td>
<td>0</td>
<td>safety</td>
</tr>
<tr>
<td>4.5825</td>
<td>3741</td>
<td>21</td>
<td>0</td>
<td>sales</td>
</tr>
<tr>
<td>4.0249</td>
<td>1093</td>
<td>19</td>
<td>1</td>
<td>opposition</td>
</tr>
</tbody>
</table>
Hypothesis Testing: Pearson’s chi-square test

- *t*-test has been criticized because it assumes the data is normally distributed.
- Pearson’s $\chi^2$ test also starts by assuming words are independent.
- First, compute a table of observed values:

<table>
<thead>
<tr>
<th>$w_2 = \text{companies}$</th>
<th>$w_1 = \text{new}$</th>
<th>$w_1 \neq \text{new}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2 = \text{companies}$</td>
<td>8</td>
<td>4667</td>
</tr>
<tr>
<td>(new companies)</td>
<td></td>
<td>(e.g., old companies)</td>
</tr>
<tr>
<td>$w_2 \neq \text{companies}$</td>
<td>15820</td>
<td>14287181</td>
</tr>
<tr>
<td>(e.g., new machines)</td>
<td></td>
<td>(e.g., old machines)</td>
</tr>
</tbody>
</table>
Hypothesis Testing: Pearson’s chi-square test

- The $X^2$ statistic is computed as

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$i$ and $j$ are over all rows and columns of the table.

- $O_{ij}$ is the observed value (in the table)

- $E_{ij}$ is the expected value (if words were truly independent). For example, to compute $E_{11}$:

$$E(\text{new companies}) = \frac{C(\text{new})}{N} \times \frac{C(\text{companies})}{N} \times N$$

$$= (3.6 \times 10^{-7}) \times 14\text{million}$$

$$\approx 5.2$$
Hypothesis Testing: Pearson’s chi-square test

• Again, if the $X^2$ statistic is above some threshold we accept the collocation.

• The top 20 collocations are the same for $\chi^2$ and $t$ tests.

• But the $\chi^2$ test is considered more robust, and is more frequently used.
Hypothesis Testing: Pearson’s chi-square test

- Consider another application: Learning word-to-word translations from an aligned corpus.

- Here are observations for how often the French vache was aligned with the English cow:

<table>
<thead>
<tr>
<th></th>
<th>cow</th>
<th>¬ cow</th>
</tr>
</thead>
<tbody>
<tr>
<td>vache</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>¬ vache</td>
<td>8</td>
<td>570934</td>
</tr>
</tbody>
</table>

- $X^2 = 456400$ (very high), so (vache, cow) is a likely translation pair.
Hypothesis Testing: Likelihood ratios

- Another approach to hypothesis testing
- We consider two hypotheses:
  - Hypothesis 1. (Two words are independent.)
    \[ p = P(w^2|w^1) = P(w^2|\neg w^1) = P(w^2) \]
  - Hypothesis 2. (\(w^2\) depends on \(w^1\).)
    \[ p_1 = P(w^2|w^1) \]
    \[ p_2 = P(w^2|\neg w^1) \]
    \[ p_1 \neq p_2 \]
Hypothesis Testing: Likelihood ratios

Quick Notation
- $c_1 = C(w^1)$
- $c_2 = C(w^2)$
- $c_{12} = C(w^1w^2)$

- We will use the binomial model
  \[ b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k} \]
  - A coin is biased to heads with probability $p$.
  - Flip the coin $n$ times.
  - $b(k; n, p)$ is the probability of $k$ heads.
The World According to $H_1$

- What we expect:
  - $P(w^2|w^1) = P(w^2) = p = \frac{c_2}{N}$
  - $P(w^2|\neg w^1) = P(w^2) = p = \frac{c_2}{N}$

- In actuality, $w^1w^2$ occurred $c_{12}$ times - how likely is this?

- We are assuming a binomial distribution:
  - Each time $w^1$ appears, $w^2$ should follow with prob $p$.
  - Each time $\neg w^1$ appears, $w^2$ should follow with prob $p$. 
The Likelihood of our data, according to $H_1$

- Of the $c_1$ times $w^1$ occurred, $w^2$ followed $c_{12}$ times.
- This should happen with probability $b(c_{12}; c_1, p)$.

- Of the $C(\neg w^1) = N - c_1$ times $\neg w^1$ occurred, $w^2$ followed $c_2 - c_{12}$ times.
- This should happen with probability $b(c_2 - c_{12}; N - c_1, p)$.

- The total probability (likelihood) of all the data is simply the product:

$$L(H_1) = \mathbb{P}( \text{ all the times we saw } w^2 ) = b(c_{12}; c_1, p) \times b(c_2 - c_{12}; N - c_1, p)$$
The World According to $H_2$

- What we expect:
  - $P(w^2|w^1) = P(w^2) = p_1 = \frac{c_{12}}{c_1}$
  - $P(w^2|\neg w^1) = P(w^2) = p_2 = \frac{c_2 - c_{12}}{N-c_1}$

- In actuality, $w^1w^2$ occurred $c_{12}$ times - how likely is this?

- Well, we are assuming a binomial distribution:
  - Each time $w^1$ appears, $w^2$ should follow with prob $p_1$.
  - Each time $\neg w^1$ appears, $w^2$ should follow with prob $p_2$. 
The Likelihood of our data, according to $H_2$

- Of the $c_1$ times $w^1$ occurred, $w^2$ followed $c_{12}$ times.
- This should happen with probability $b(c_{12}; c_1, p_1)$.

- Of the $C(\neg w^1) = N - c_1$ times $\neg w^1$ occurred, $w^2$ followed $c_2 - c_{12}$ times.
- This should happen with probability $b(c_2 - c_{12}; N - c_1, p_2)$.

- The total probability (likelihood) of all the data is simply the product:

$$L(H_2) = P(\text{ all the times we saw } w^2 )$$

$$= b(c_{12}; c_1, p_1) \times b(c_2 - c_{12}; N - c_1, p_2)$$
The Likelihood Ratio - general hypothesis testing

- We are given two hypotheses and a some data.
- We have no reason to believe one over the other \( (P(H_1) = P(H_2)) \)
- We pick \( H_2 \) if

\[
\frac{P(H_2|\text{data})}{P(H_1|\text{data})} = \frac{P(\text{data}|H_2)P(H_2)}{P(\text{data}|H_1)P(H_1)} = \frac{P(\text{data}|H_2)}{P(\text{data}|H_1)}
\]

is large.
The Likelihood Ratio

• We want bigrams with large ratio

\[
\frac{L(H_2)}{L(H_1)}
\]

• To make the numbers nice, we equivalently find bigrams with large

\[-2 \log \left( \frac{L(H_1)}{L(H_2)} \right)\]

• It turns out that the value \(-2 \log \left( \frac{L(H_1)}{L(H_2)} \right)\) is “asymptotically \(\chi^2\) distributed” (more so than the \(X^2\) statistic).
### Using The Likelihood Ratio

<table>
<thead>
<tr>
<th>$-2 \log \lambda$</th>
<th>$C(w^1)$</th>
<th>$C(w^2)$</th>
<th>$C(w^1w^2)$</th>
<th>$w^1$</th>
<th>$w^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1291.42</td>
<td>12593</td>
<td>932</td>
<td>150</td>
<td>most</td>
<td>powerful</td>
</tr>
<tr>
<td>99.31</td>
<td>379</td>
<td>932</td>
<td>10</td>
<td>politically</td>
<td>powerful</td>
</tr>
<tr>
<td>82.96</td>
<td>932</td>
<td>934</td>
<td>10</td>
<td>powerful</td>
<td>computers</td>
</tr>
<tr>
<td>80.39</td>
<td>932</td>
<td>3424</td>
<td>13</td>
<td>powerful</td>
<td>force</td>
</tr>
<tr>
<td>57.27</td>
<td>932</td>
<td>291</td>
<td>6</td>
<td>powerful</td>
<td>symbol</td>
</tr>
<tr>
<td>51.66</td>
<td>932</td>
<td>40</td>
<td>4</td>
<td>powerful</td>
<td>lobbies</td>
</tr>
<tr>
<td>51.52</td>
<td>171</td>
<td>932</td>
<td>5</td>
<td>economically</td>
<td>powerful</td>
</tr>
<tr>
<td>51.05</td>
<td>932</td>
<td>43</td>
<td>4</td>
<td>powerful</td>
<td>magnet</td>
</tr>
<tr>
<td>50.83</td>
<td>4458</td>
<td>932</td>
<td>10</td>
<td>less</td>
<td>powerful</td>
</tr>
<tr>
<td>50.75</td>
<td>6252</td>
<td>932</td>
<td>11</td>
<td>very</td>
<td>powerful</td>
</tr>
<tr>
<td>49.36</td>
<td>932</td>
<td>2064</td>
<td>8</td>
<td>powerful</td>
<td>position</td>
</tr>
<tr>
<td>48.78</td>
<td>932</td>
<td>591</td>
<td>6</td>
<td>powerful</td>
<td>machines</td>
</tr>
<tr>
<td>47.42</td>
<td>932</td>
<td>2339</td>
<td>8</td>
<td>powerful</td>
<td>computer</td>
</tr>
</tbody>
</table>
Conclusion

• Counting Frequencies of adjacent words
  - too sensitive to frequent pairs

• Mutual Information between words
  - too sensitive to rare words

• Hypothesis Testing
  – t test - okay
  – $\chi^2$ test - better
  – likelihood ratios - even better
The end

Questions?