

# 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:

$C(w^1 w^2)$	$w^1$	$w^2$
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the

## Counting Frequencies of Adjacent Words

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

$C(w^1 w^2)$	$w^1$	$w^2$	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N

# 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:

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

$W^1$  and  $W^2$  are random variables, like word locations.

## Pointwise Mutual Information

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

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

## Pointwise Mutual Information

- Difficulty: PMI is too sensitive to rare bigrams

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

- 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  $PMI(w^1, w^2)' = C(w^1, w^2) * PMI(w^1, w^2)$

## 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:  $t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$
- 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

$$\begin{aligned}\mu &= P(\text{new})P(\text{companies}) \\ &= \frac{15,828}{14\text{million}} * \frac{4675}{14\text{million}} \\ &\approx 3.6 \times 10^{-7}\end{aligned}$$

## 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:

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

**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:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2 + s_2^2}{N}}}$$

- Find words (like computer) with highest  $t$  score.

## Hypothesis Testing: The $t$ test of **differences**

- Results:

$t$	$C(w)$	$C(\text{strong } w)$	$C(\text{powerful } w)$	Word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition

## 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:

	$w_1 = \textit{new}$	$w_1 \neq \textit{new}$
$w_2 = \textit{companies}$	8 ( <i>new companies</i> )	4667 ( <i>e.g., old companies</i> )
$w_2 \neq \textit{companies}$	15820 ( <i>e.g., new machines</i> )	14287181 ( <i>e.g., old machines</i> )

## 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  $E_{11}$ :

$$\begin{aligned} E(\text{new companies}) &= \frac{C(\text{new})}{N} \times \frac{C(\text{companies})}{N} \times N \\ &= (3.6 \times 10^{-7}) \times 14\text{million} \\ &\approx 5.2 \end{aligned}$$

## Hypothesis Testing: Pearson's chi-square test

- Again, if the  $\chi^2$  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*:

	<i>cow</i>	$\neg$ <i>cow</i>
<i>vache</i>	59	6
$\neg$ <i>vache</i>	8	570934

- $X^2 = 456400$  (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

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

$$\begin{aligned} L(H_1) &= P(\text{ all the times we saw } w^2 ) \\ &= b(c_{12}; c_1, p) * b(c_2 - c_{12}; N - c_1, p) \end{aligned}$$

## 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:

$$\begin{aligned} L(H_2) &= P(\text{all the times we saw } w^2) \\ &= b(c_{12}; c_1, p_1) * b(c_2 - c_{12}; N - c_1, p_2) \end{aligned}$$

## 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

$$\begin{aligned}\frac{P(H_2|data)}{P(H_1|data)} &= \frac{P(data|H_2)P(H_2)}{P(data|H_1)P(H_1)} \\ &= \frac{P(data|H_2)}{P(data|H_1)}\end{aligned}$$

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

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

## 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?