Collocation

Hung-Bin, Chen

References:

Foundations of Statistical Natural Language Processing, chap 5

Introduction

- Collocation
- Frequency
- Mean and Variance
- Hypothesis Testing
  - *t*- test
  - Chi-square test
  - Likelihood ratios
- Mutual Information

- corpus
  - The reference corpus consists of four months of the New York Times newswire: 1990/08 ～ 11. 115 Mb of text and 14 million words

Collocation (1/4)

- A **Collocation** is an expression consisting of two or more words that correspond to some conventional way of saying things

- **Collocations** of a given word are statements of the habitual or customary place of that word
  - E.g.,
    - broad daylight
    - strong tea
    - kick the bucket
    - hear it through the grapevine

Collocation (2/4)

- Definition of a collocation
  - (Choueka, 1988)
  - [A collocation is defined as] “a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components."

Collocation (3/4)

- Criteria:
  - non-compositionality
    - The meaning of a collocation is not a straight-forward composition of the meanings of its parts.
    - E.g., *to hear it through the grapevine*
  - non-substitutability
    - We cannot substitute near-synonyms for the components of a collocation.
    - E.g., *strong tea* $\neq$ *powerful tea*

Collocation (4/4)

- Criteria:
  - non-modifiability
    - Many collocations cannot be freely modified with additional lexical material or through grammatical transformations
    - *To get a frog in one ' s throat ≠ get an ugly frog in one ' s throat*
  - non-translatable (word for word)
    - we cannot translate it word by word
    - English:
      - *just a pice of cake*
    - Chinese:
      - *只是一片蛋糕??*

Why study collocations?

- In nature language generator (NLG)
  - a sequence of words should be natural
- In lexicography
  - to automatically identify the important collocations to be listed in a dictionary entry

Frequency (1/5)

- The simplest method for finding collocations in a text corpus is counting
  - If two words occur together a lot, then that is evidence that they have a special function that is not simply explained as the function that results from their combination.

- Method:
  - Select the most frequently occurring bigrams

Frequency (2/5)

- We are not very interesting as is shown in right side table
  - Except for "*New York* ", all bigrams are pairs of function words

- Solution:
  - Pass the candidate phrases through a Part of speech tag patterns for collocation filtering
  - (Justeson & Katz, 1995)

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
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
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Frequency (3/5)

- Frequency + POS filter
  - There are only 3 bigrams that we would not regard as noncompositional phrases: last year, last week, and next year

$C(w^1\ w^2)$	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
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
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N

Frequency (4/5)

- Compare "*strong*" and "*powerful*"
  - The nouns w occurring most often in the patterns "*strong* w" and "*powerful* w"

w	$C(strong, w)$	w	$C(powerful, w)$
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5

Frequency (5/5)

- Conclusion
  - works well for fixed phrases
  - Simple method
  - Requires small linguistic knowledge

  - But many collocations consist of two words that stand in a more flexible relationship to one another
  - E.g.
    - 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*

Mean and Variance (1/7)

- The mean and standard deviation characterize the distribution of distances between two words in a corpus
  - A low variance means that the two words usually occur at about the same distance
  - A low variance --> good candidate for collocation

Mean and Variance (2/7)

- The **mean** is the average offset (signed distance) between two words in a corpus

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

- The **variance** measures how much the individual offsets deviate from the mean

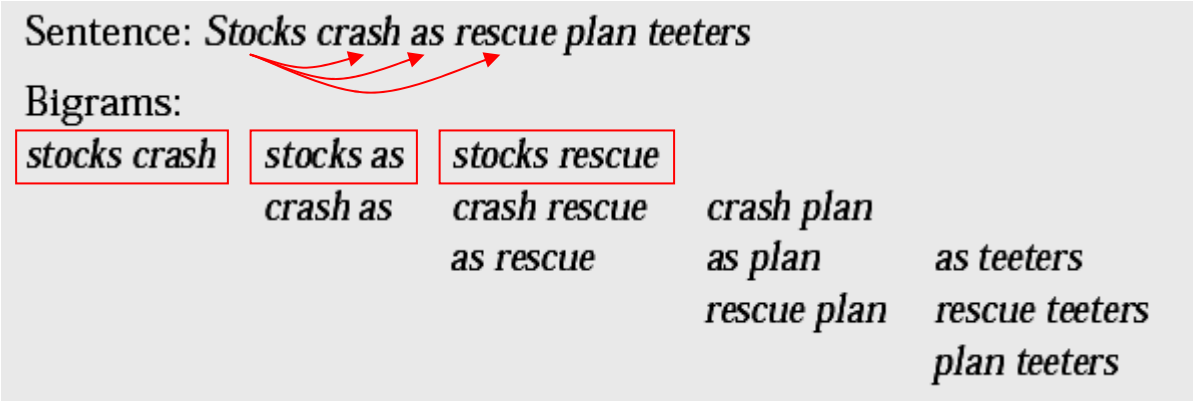
$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

n is the number of times the two words

d_i is the offset of the ith pair of candidates

Mean and Variance (3/7)

- Capture collocations of fixed distance
  - bigram, trigram …
  - E.g.,
    Using a three word **collocational window** to capture bigrams at a distance



Mean and Variance (4/7)

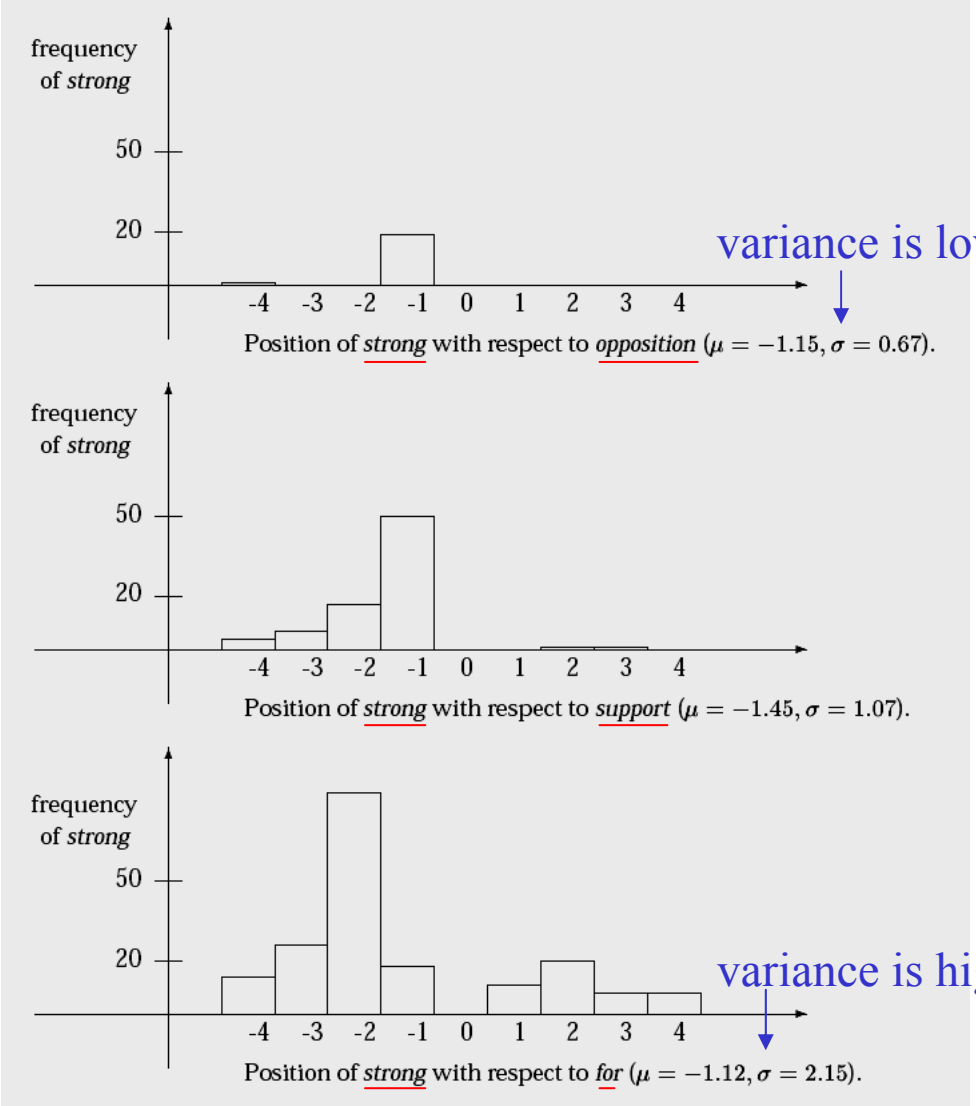
- Capture collocations of variable distances
  - E.g., Discovering the relationship between *knocked* and *door* is to compute the *mean* and *variance* of the offsets
    - she *knocked* on his *door*
    - they *knocked* at the *door*
    - 100 women *knocked* on Donaldson ' s *door*
    - a man *knocked* on the metal front *door*

$$\text{Mean}, \bar{d} = \frac{3+3+5+5}{4} = 4$$

$$\text{Std. deviation}, s = \sqrt{\frac{(3-4)^2 + (3-4)^2 + (5-4)^2 + (5-4)^2}{3}}$$

Mean and Variance (5/7)

- For example



Mean and Variance (6/7)

- std. dev. ~0 & mean offset ~1 --> would be found by frequency method
- std. dev. ~0 & high mean offset --> very interesting, but would not be found by frequency method

σ	μ	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by

Mean and Variance (7/7)

- Criteria:
  - If offsets (d_i) are the nearly in all co-occurrences
    - variance is low
    - definitely a collocation
  - If offsets (d_i) are randomly distributed
    - variance is high
    - not a collocation

Hypothesis Testing (1/2)

- High frequency and low variance can be accidental
  - two words to co-occur a lot just by chance

- We formulate a null hypothesis H_0 that there is noassociation between the words beyond chance occurrenc
  - compute the probability *p* that the event would occur if H_0 were true, and then reject
  - Typically a *significant level* of $p < 0.05, 0.01, 0.01$ or 0.001

Hypothesis Testing (2/2)

- How can we apply the methodology of hypothesis testing
  - We formulate a *null hypothesis* H_0
    - H_0 : no real association (just chance…)
    - if two words w1 and w2 do not form a collocation, then w1 and w2 are independently
    - generated completely independently is simply given by:

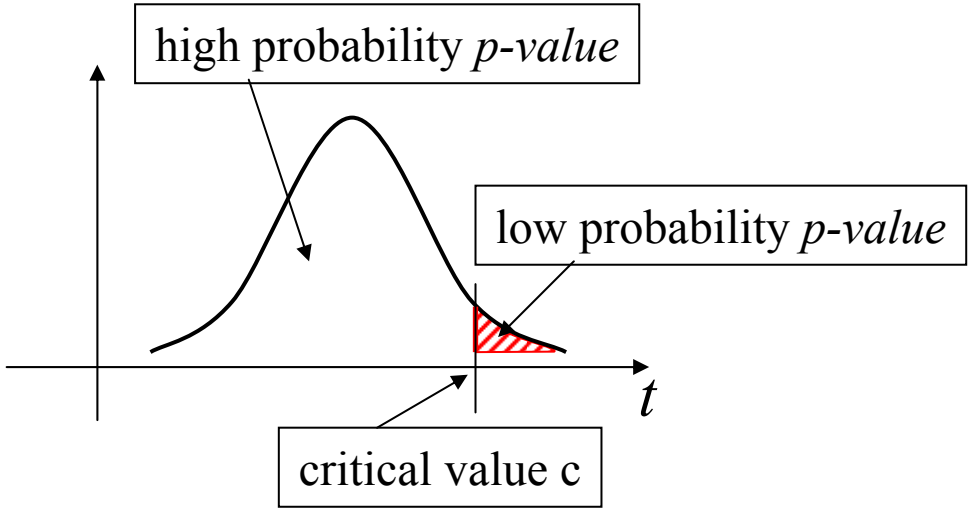
$$P(w^1 w^2) = P(w^1)P(w^2)$$

observed **expected**

The *t* test (1/7)

- The *t* test looks at the mean and variance of a sample of measurements
  - where the null hypothesis is that the sample is drawn from a distribution with mean μ

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



x is the sample mean, $s2$ is the sample variance, N is the sample size, and μ is the mean of the distribution

The *t* test (2/7)

- a simple example:
  - H_0 : Null hypothesis is that the mean height of a population of men is 158cm
  - Data is given a sample of 200 men with x =169 and $s2$ = 2600
  - Test: this sample is from the general population (the null hypothesis) or whether it is from a different population of smaller men

$$t = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$$

Confidence level of
$\alpha = 0.005$, $t_0 = 2.576$

Result: Since the t we got is larger than 2.576, we can reject the null hypothesis with 99.5% confidence. The sample is not drawn from a population with mean 158cm.

The *t* test (3/7)

- Example with collocations
  - ***"new companies"*** Is it a collocation??
  - In a corpus:

w	c(w)
new	15,828
companies	4,675
new companies	8
total words	14,307,668

  - null hypothesis
    - occurrences of new and companies are independent
    - P(*new companies*) = P(*new*) P(*companies*)

The *t* test (4/7)

- P(*new companies*) = P(*new*) P(*companies*)

$$\text{expected mean } \mu = \text{P(new) P(companies)}$$
$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$
$$\text{observed mean } \bar{x} = \text{P(new companies)}$$
$$= \frac{8}{14307668} \approx 5.591 \times 10^{-7}$$

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 0.99932$$

apply binomial distribution: $s^2 = np(1-p)$, when n=1
then the variance $s^2 = p(1-p) \approx p$, for most bigrams p is small

The *t* test (5/7)

- With a confidence level α =0.005, critical value is 2.576

- Since t $\approx$ 0.999932 < 2.576
  - We cannot reject the null hypothesis that new and companies occur independently and do not form a collocation

The *t* test (6/7)

- The t test applied to 10 bigrams that occur with frequency 20

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

For the top five bigrams, we can reject the null hypothesis.

The *t* test (7/7)

- Notes:
  - the *t* test takes into account the frequency of a bigram relative to the frequencies of its component words
    - If a high proportion of the occurrences of both words, then its t value is high
  - The t test and other statistical tests are most useful as a method for *ranking* collocations.

Hypothesis testing of differences (1/2)

- The t test apply to a slightly different collocation discovery problem
  - to find words whose co-occurrence patterns best distinguish between two words
  - the null hypothesis is that the average difference is 0 ($\mu=0$)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ is the sample mean of population 1

$\bar{x}_2$ is the sample mean of population 2

s_1^2 is the sample variance of population 1

s_2^2 is the sample variance of population 2

n_1 is the sample size of population 1

n_1 is the sample size of population 2

Hypothesis testing of differences (2/2)

- Used to see if 2 words (near-synonyms) are used in the same context or not
  - "*strong*" *vs* "*powerful*"

t	$C(w)$	$C(strong\ w)$	$C(powerful\ w)$	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
3.9000	802	18	1	showing

chi-square test (1/9)

- An alternative test for dependence
  - The *t*-test has an assumes is that probabilities are approximately normally distributed, which is not true in general
  - the χ^2 -test does not assume normally distributed probabilities

- chi-square test
  - The essence is to compare the observed frequencies in the table with the frequencies expected for independence

	w1	¬w1
w2	obs(w1,w2) exp(w1,w2)	obs(¬w1,w2) exp(¬w1,w2)
¬w2	obs(w1,¬w2) exp(w1,¬w2)	obs(¬w1,¬w2) exp(¬w1,¬w2)

chi-square test (2/9)

- χ^2 test statistic
  - sums the differences between observed frequencies and expected values for independence

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_n - E_n)^2}{E_n}$$

$$= \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

	w1	¬w1
w2	Obs_1 Exp_1	Obs_2 Exp_2
¬w2	Obs_3 Exp_3	Obs_4 Exp_4

chi-square test (3/9)

- Example with collocations
  - χ^2 test

$$x^2 = \sum_{ij} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

  - Observed frequencies

Observed	w1 = new	w1 ≠ new	TOTAL
w2 = companies	8 (new companies)	4667 (ex: old companies)	4675 c(companies)
w2 ≠ companies	15820 (ex: new machines)	14287173 (ex: old machines)	14302993 c(¬ companies)
TOTAL	15828 c(new)	14291840 c(¬ new)	N = 14307668

chi-square test (4/9)

- Expected frequencies Exp_{ij}
- E.g., expected frequency for cell (1,1) (*new companies*)
  - If "*new*" and "*companies*" occurred completely independent
  - we would expect 5.17 occurrences of "*new companies*" on average

$$Exp_{11} = \frac{8+4667}{N} \times \frac{8+15820}{N} \times N = 5.17$$

Observed	w1 = new	w1 ≠ new
w2 = companies	5.17 c(new) x c(companies) / N 15828 x 4675 / 14307676	4669.83 c(companies) x c(¬ new) / N 4675 x 14291848 / 14307676
w2 ≠ companies	15822.83 c(new) x c(¬ companies) / N 15828 x 14303001 /14307676	14287178.17 c(¬ new) x c(¬ companies) / N 14291848 x 14303001 / 14307676

chi-square test (5/9)

- χ^2 test
  - sums the differences

$$x^2 = \frac{(8-5.17)^2}{5.17} \times \frac{(4667-4669.83)^2}{4669.83} \times \frac{(15820-15822.83)^2}{15822.83} \times \frac{(14287173-14287178.17)^2}{14287178.17} \approx 1.55$$

  - The χ 2 test can be applied to tables of any size, but it has a simpler form for 2-by-2 tables:

$$x^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

$$x^2 = \frac{14307668(8\times 142871818 - 4667\times 15820)^2}{(8+4667)(8+15820)(4667+142871818)(15820+142871818)} \approx 1.55$$

chi-square test (6/9)

- χ^2 test
  - The probability level of α =0.05 the critical value is 3.84
  - Since χ^2 = 1.55 < 3.84:
    - So we cannot reject H0 (that *new* and *companies* occur independently of each other)
    - So *new companies* is not a good candidate for a collocation

chi-square test (7/9)

- χ^2 test for machine translation
  - To identify translation word pairs in aligned corpora
  - E.g., sentence pairs which have "*cow*" in the English sentence and "*vache*" in the French sentence

Observed	cow	¬ cow	TOTAL
vache	56	6	65
¬ vache	8	570934	570942
TOTAL	67	570940	571007

  - $\chi 2 = 456\,400 >> 3.84$ (with $\alpha = 0.05$)
  - So "vache" and "cow" are not independent… and so are translations of each other

chi-square test (8/9)

- χ^2 test for corpus similarity
  - Compute χ^2 for the 2 populations (corpus1 and corpus2)
  - Ho: the 2 corpora have the same word distribution

Observed	Corpus 1	Corpus 2	Ratio
Word 1	60	9	60/9 =6.7
Word 2	500	75	6.6
Word 3	124	20	6.2
…	…	…	…
Word 500	…	…	…

chi-square test (9/9)

- χ^2 test is appropriate for large probabilities
- χ^2 is not appropriate with sparse data (if numbers in the 2 by 2 tables are small)
- Against using χ^2 if the total sample size is smaller than 20 or if it is between 20 and 40 and the expected value in any of the cells is 5 or less

Likelihood ratios (1/6)

- Two Hypothesis used in Likelihood ratios
  - Hypothesis one is a formalization of independence
  - Hypothesis two is a formalization of dependence

$$\text{Hypothesis } 1 : P(w^2 \mid w^1) = p = P(w^2 \mid \neg w^1)$$
$$\text{Hypothesis } 2 : P(w^2 \mid w^1) = p_1 \neq p_2 = P(w^2 \mid \neg w^1)$$

  - We use the usual MLE for p, p_1 and p_2 and write c_1 c_2 and c_{12} for the number of occurrences of w_1, w_2 and $w_1 w_2$ in corpus

$$p = \frac{c_2}{N}, \quad p_1 = \frac{c_{12}}{c_1}, \quad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

Likelihood ratios (2/6)

- Assuming a binomial distribution:

$$b(k;n,x) = \binom{n}{k} x^k (1-x)^{n-k}$$

	H_1	H_2
$P(w^2 \mid w^1)$	$P = c_2 / N$	$P_1 = c_{12} / c_1$
$P(w^2 \mid \neg w^1)$	$P = c_2 / N$	$P_2 = (c_2-c_{12}) / (N-c_1)$
c_{12} out of c_1 bigrams are w^1w^2	$b(c_{12} ; c_1, p)$	$b(c_{12} ; c_1, p_1)$
c_2–c_{12} out of N-c1 bigrams are :w^1w^2	$b(c_2-c_{12} ; N-c_1, p)$	$b(c_2-c_{12} ; N-c_1, p_2)$

$$L(H_1) = b(c_{12} ; c_1, p) \times b(c_2-c_{12} ; N-c_1, p)$$
$$L(H_2) = b(c_{12} ; c_1, p_1) \times b(c_2-c_{12} ; N-c_1, p_2)$$

Likelihood ratios (3/6)

- ratios

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}$$

$$= \log \frac{b(c_{12}\ ;\ c_1,\ p) \times b(c_2\text{-}c_{12}\ ;\ N\text{-}c_1,\ p)}{b(c_{12}\ ;\ c_1,\ p_1) \times b(c_2\text{-}c_{12}\ ;\ N\text{-}c_1,\ p_2)}$$

$$= \log L(c_{12}\ ;\ c_1,\ p) + \log L(c_2\text{-}c_{12}\ ;\ N\text{-}c_1,\ p)$$

$$- \log L(c_{12}\ ;\ c_1,\ p_1) - \log L(c_2\text{-}c_{12}\ ;\ N\text{-}c_1,\ p_2)$$

Where $L(k, n, x) = x^k (1-x)^{n-k}$

Likelihood ratios (4/6)

- ratios

$-2\log\lambda$	$C(w^1)$	$C(w^2)$	$C(w^1w^2)$	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
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip

Easier to interpret *!!*

Likelihood ratios (5/6)

- Likelihood ratios is simply a number that tells us how much more likely one hypothesis is than the other
- –2log λ is asymptotically χ 2 distributed
- Likelihood ratios is easily to interpret the sparse data
- The approximation is usually good, even for small sample sizes

Likelihood ratios (6/6)

- Relative frequency ratios

$$r = \frac{2/14307668}{68/11731564} \approx 0.024116$$

ratio	1990	1989	w^1	w^2
0.0241	2	68	Karim	Obeid
0.0372	2	44	East	Berliners
0.0372	2	44	Miss	Manners
0.0399	2	41	17	earthquake
0.0409	2	40	HUD	officials
0.0482	2	34	EAST	GERMANS
0.0496	2	33	Muslim	cleric
0.0496	2	33	John	Le
0.0512	2	32	Prague	Spring
0.0529	2	31	Among	individual

Ten bigrams that occurred twice in the 1990 New York Times corpus, ranked according to the (inverted) ratio of relative frequencies in 1989 and 1990

Mutual Information (1/6)

- Uses a measure from information-theory
  - Originally defined mutual information between particular events *x* and *y*, in our case the occurrence of two words
  - If two events *x* and *y* are independent, then I(x,y) = 0

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{P(x \mid y)}{P(x)} = \log_2 \frac{P(x \mid y)}{P(y)}$$

Mutual Information (2/6)

- Assume:
  - c(Ayatollah) = 42
  - c(Ruhollah) = 20
  - c(Ayatollah, Ruhollah) = 20
  - N = 143 076 668
- Then:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$\mathrm{I(Ayatollah, Ruhollah)} = \log_2 \left(\frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \right) \approx 18.38$$

Mutual Information (3/6)

- Mutual Information with the same ranking as t-test

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	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			
0.46	14093	14776			
0.29	15019	15629			

mutual information

$$\text{I(Ayatollah, Ruhollah)} = \log_2 \left(\frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \right) \approx 18.38$$

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

t-test

Mutual Information (4/6)

- well measure of independence
  - values close to 0 --> independence

- bad measure of dependence
  - bigrams composed of low-frequency words will receive a higher score than bigrams composed of high-frequency words
  - because score depends on frequency ratios

$$I(x,y) = \log \frac{P(x \mid y_1)}{P(x)} \text{ , } \log \frac{P(x \mid y_2)}{P(x)}$$

$$=> \log \frac{\frac{31950}{31950+4793}}{P(x)} \approx \log \frac{0.87}{P(x)} < \log \frac{\frac{4974}{4974+441}}{P(x)} \approx \log \frac{0.92}{P(x)}$$

Mutual Information (5/6)

- These examples illustrate that a large proportion of bigrams are not well

inaccurate due to sparseness *!!*

I_{1000}	w^1	w^2	w^1w^2	bigram	I_{23000}	w^1	w^2	w^1w^2	bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Mutual Information (6/6)

- because of originally define is a bad measure of dependence on the frequency

- redefined as $C(w_1w_2)I(w_1,w_2)$
  - to compensate the bias of the original definition in favor of low-frequency events