Word collocations

What is a Collocation?

- A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things.
- The words together can mean more than their sum of parts (*The Times of India, disk drive*)
  - Previous examples: hot dog, mother in law

- Examples of collocations
  - noun phrases like *strong tea* and *weapons of mass destruction*
  - phrasal verbs like *to make up*, and other phrases like *the rich and powerful*.
- Valid or invalid?
  - *a stiff breeze* but not *a stiff wind* (while either *a strong breeze* or *a strong wind* is okay).
  - *broad daylight* (but not *bright daylight* or *narrow darkness*).

Criteria for Collocations

- Typical criteria for collocations:
  - non-compositionality
  - non-substitutability
  - non-modifiability.

- Collocations usually cannot be translated into other languages word by word.

- A phrase can be a collocation even if it is not consecutive (as in the example *knock . . . door*).

Non-Compositionality

- A phrase is compositional if the meaning can be predicted from the meaning of the parts.
  - E.g. new companies
- A phrase is non-compositional if the meaning cannot be predicted from the meaning of the parts
  - E.g. hot dog

- Collocations are not necessarily fully compositional in that there is usually an element of meaning added to the combination. Eg. *strong tea.*

- Idioms are the most extreme examples of non-compositionality. Eg. *to hear it through the grapevine.*

Non-Substitutability

- We cannot substitute near-synonyms for the components of a collocation.
- For example
  - We can't say *yellow wine* instead of *white wine* even though *yellow* is as good a description of the color of white wine as *white* is (it is kind of a yellowish white).

- Many collocations cannot be freely modified with additional lexical material or through grammatical transformations (**Non-modifiability**).
  - E.g. white wine, but not whiter wine
  - mother in law, but not mother in laws

Linguistic Subclasses of Collocations

- Light verbs:
  - Verbs with little semantic content like *make*, *take* and *do.*
  - *E.g. make lunch*, *take easy,*
- Verb particle constructions
  - *E.g. to go down*
- Proper nouns
  - *E.g. Bill Clinton*
- Terminological expressions refer to concepts and objects in technical domains.
  - *E.g. Hydraulic oil filter*

Principal Approaches to Finding Collocations

- How to automatically identify collocations in text?
- Simplest method: Selection of collocations by **frequency**
- Selection based on **mean and variance** of the distance between focal word and collocating word
- **Hypothesis testing**
- **Mutual information**

Frequency

- Find collocations by counting the number of occurrences.
- Need also to define a maximum size window

- Usually results in a lot of function word pairs that need to be filtered out.
- Fix: pass the candidate phrases through a part of-speech filter which only lets through those patterns that are likely to be “phrases”. (Justesen and Katz, 1995)

$C(w^1\ w^2)$	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

Most frequent bigrams in an Example Corpus

Except for *New York*, all the bigrams are pairs of function words.

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Part of speech tag patterns for collocation filtering (Justesen and Katz).

$C(w^1\ w^2)$	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
1074	chief	executive	A N
1073	real	estate	A N

The most highly ranked phrases after applying the filter on the same corpus as before.

Collocational Window

- Many collocations occur at variable distances. A collocational window needs to be defined to locate these. Frequency based approach can’t be used.
  - 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

- The mean μ is the *average offset* between two words in the corpus.
- The variance:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

  - where *n* is the number of times the two words co-occur, d_i is the offset for co-occurrence *i*, and μ is the mean.

- Mean and variance characterize the distribution of distances between two words in a corpus.
  - High variance means that co-occurrence is mostly by chance
  - Low variance means that the two words usually occur at about the same distance.

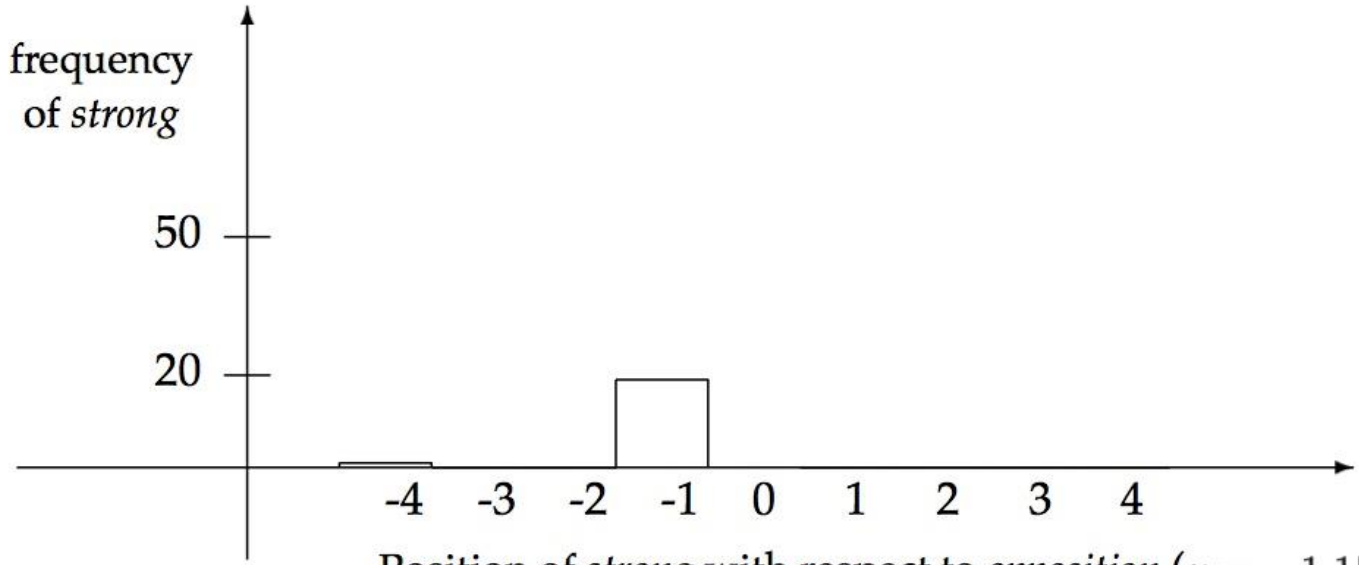
Mean and Variance: An Example

- For the *knock, door* example sentences the mean is:

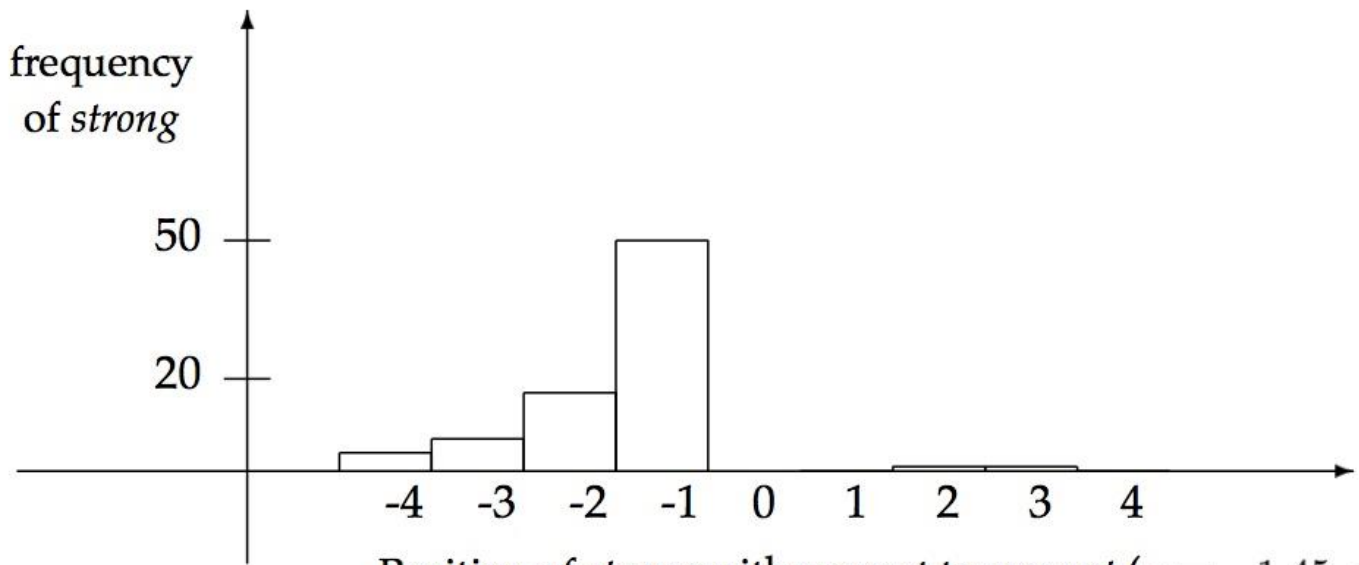
$$\frac{1}{4}(3+3+5+5) = 4.0$$

- And the sample deviation:

$$\sigma = \sqrt{\frac{1}{3}((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)} \approx 1.15$$



Position of *strong* with respect to *opposition* ($\mu = -1.15, \sigma = 0.67$).



Position of *strong* with respect to *support* ($\mu = -1.45, \sigma = 1.07$).

Finding collocations based on mean and variance

s	$\bar{d}$	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
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Ruling out Chance

- Two words can co-occur by chance.
  - High frequency and low variance can be accidental

- **Hypothesis Testing** measures the confidence that this co-occurrence was really due to association, and not just due to chance.

- Formulate a *null hypothesis* H_0 that there is no association between the words beyond chance occurrences.
- The null hypothesis states what should be true if two words do not form a collocation.
- If the null hypothesis can be rejected, then the two words do not co-occur by chance, and they form a collocation
- Compute the probability p that the event would occur if H_0 were true, and then reject H_0 if p is too low (typically if beneath a ***significance level*** of $p < 0.05$, 0.01, 0.005, or 0.001) and retain H_0 as possible otherwise.

The *t*-Test

- *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.
- The test looks at the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely one is to get a sample of that mean and variance, assuming that the sample is drawn from a normal distribution with mean μ.

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

Where x is the sample mean, s^2 is the sample variance, N is the sample size, and μ is the mean of the distribution.

t-Test for finding collocations

- Think of the text corpus as a long sequence of *N* bigrams, and the samples are then indicator random variables with:
  - value 1 when the bigram of interest occurs,
  - 0 otherwise.
- The *t*-test and other statistical tests are useful as methods for ***ranking*** collocations.
- Step 1: Determine the expected mean
- Step 2: Measure the observed mean
- Step 3: Run the t-test

t-Test: Example

- In our corpus, *new* occurs 15,828 times, *companies* 4,675 times, and there are 14,307,668 tokens overall.
- *new companies* occurs 8 times among the 14,307,668 bigrams
- H_0 : P(*new companies*) =P(*new*)P(*companies*)

$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

t-Test example

- For this distribution $\mu = 3.615 \times 10^{-7}$ and $\sigma^2 = p(1-p) \approx p$

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

- t value of 0.999932 is not larger than 2.576, the critical value for a=0.005. So we cannot reject the null hypothesis that *new* and *companies* occur independently and do not form a collocation.

(Student s) t test critical values. A t distribution with d.f. degrees of freedom has percentage C of the area under the curve between $-t^*$ and t^* (two-tailed), and proportion p of the area under the curve between t^* and ∞ (one tailed). The values with infinite degrees of freedom are the same as critical values for the z test.

	P	0.05	0.025	0.01	0.005	0.001	0.0005
	C	90%	95%	98%	99%	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(z)	∞	1.645	1.960	2.326	2.576	3.091	3.291

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 of differences (Church and Hanks, 1989)

- To find words whose co-occurrence patterns best distinguish between two words.
- For example, in computational lexicography we may want to find the words that best differentiate the meanings of *strong* and *powerful*.
- The *t*-test is extended to the comparison of the means of two normal populations.
- Here the null hypothesis is that the average difference is 0 ($\mu = 0$).
- In the denominator we add the variances of the two populations since the variance of the difference of two random variables is the sum of their individual variances.

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

- If *w* is the collocate of interest (e.g., *computers* or *symbol*) and v^1 and v^2 are the words we are comparing (e.g., *powerful* and *strong*), then:

$$\bar{x}_1 = s_1^2 = P(v^1 w), \bar{x}_2 = s_2^2 = P(v^2 w)$$

$$t \approx \frac{P(v^1 w) - P(v^2 w)}{\sqrt{\frac{P(v^1 w) + P(v^2 w)}{N}}}$$

We can simplify this as follows.

$$t \approx \frac{\frac{C(v^1 w)}{N} - \frac{C(v^2 w)}{N}}{\sqrt{\frac{C(v^1 w) + C(v^2 w)}{N^2}}}$$

$$= \frac{C(v^1 w) - C(v^2 w)}{\sqrt{C(v^1 w) + C(v^2 w)}}$$

t	$C(w)$	$C(\textit{strong}\ w)$	$C(\textit{powerful}\ w)$	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
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
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
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

Table 5.7 Words that occur significantly more often with *powerful* (the first ten words) and *strong* (the last ten words).

Pearson's chi-square test

- *t*-test assumes that probabilities are approximately normally distributed, which is not true in general. The χ^2 test doesn't make this assumption.
- the essence of the χ^2 test is to compare the observed frequencies with the frequencies expected for independence
  - if the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.
- Relies on co-occurrence table, and computes

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

χ^2 Test: Example

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

The χ^2 statistic sums the differences between observed and expected values in all squares of the table, scaled by the magnitude of the expected values, as follows:

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

where *i* ranges over rows of the table, *j* ranges over columns, O_{ij} is the observed value for cell *(i, j)* and E_{ij} is the expected value.

χ^2 Test: Example

- Observed values O are given in the table
  - E.g. O(1,1) = 8
- Expected values E are determined from marginal probabilities:
  - E.g. E value for cell (1,1) = *new companies* is expected frequency for this bigram, determined by multiplying:
    - probability of *new* on first position of a bigram
    - probability of *companies* on second position of a bigram
    - total number of bigrams
  - E(1,1) = (8+15820)/N * (8+4667)/N * N =~ 5.2
- χ^2 is then determined as 1.55:

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

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

- Look up significance table:
  - χ^2 = 3.8 for probability level of α = 0.05
  - 1.55 < 3.8
  - we cannot reject null hypothesis → *new companies* is not a collocation

χ^2 critical values. A table entry is the point χ^{2*} with proportion p of the area under the curve being in the right-hand tail from χ^{2*} to ∞ of a χ^2 curve with d.f. degrees of freedom. (When using an r x c table, there are $(r-1)(c-1)$ degrees of freedom.)

	P	0.99	0.95	0.10	0.05	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	3.84	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

χ^2 Test: Applications

- Identification of translation pairs in aligned corpora (Church and Gale, 1991).
  - Determine chi-square for translation pairs

	cow	!cow
vache	59	6
!vache	8	570934

- Corpus similarity (Kilgarriff and Rose, 1998)

Likelihood Ratios

- It is simply a number that tells us how much more likely one hypothesis is than the other.
  - More appropriate for sparse data than the χ^2 test.
  - A *likelihood ratio*, is more interpretable than the χ^2 or *t* statistic.

- For collocation discovery, we examine the following two alternative explanations for the occurrence frequency of a bigram w^1w^2:
  - **Hypothesis 1:** The occurrence of w^2 is independent of the previous occurrence of w^1.
  - **Hypothesis 2:** The occurrence of w^2 is dependent on the previous occurrence of w^1.

- The log likelihood ratio is then:

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

Likelihood ratios

- **Hypothesis 1.** $P(w^2|w^1) = p = P(w^2|\neg w^1)$
- **Hypothesis 2.** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

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

- Likelihood of getting the counts that we actually observed, using a binomial distribution: $\mathrm{b}(k;\ n, x) = \binom{n}{k} x^k (1-x)^{(n-k)}$

	H_1	H_2
c_{12} out of c_1 bigrams are $w^1 w^2$	$\mathrm{b}(c_{12};\ c_1, p)$	$\mathrm{b}(c_{12};\ c_1, p_1)$
$c_2 - c_{12}$ out of $N - c_1$ bigrams are $\neg w^1 w^2$	$\mathrm{b}(c_2 - c_{12};\ N - c_1, p)$	$\mathrm{b}(c_2 - c_{12};\ N - c_1, p_2)$

$$L(H_1) = \mathrm{b}(c_{12};\ c_1, p)\mathrm{b}(c_2 - c_{12};\ N - c_1, p) \text{ for Hypothesis 1}$$

$$L(H_2) = \mathrm{b}(c_{12};\ c_1, p_1)\mathrm{b}(c_2 - c_{12};\ N - c_1, p_2) \text{ for Hypothesis 2}$$

Likelihood ratios (cont.)

The log of the likelihood ratio λ is then as follows:

$$\begin{aligned}
\log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\
&= \log \frac{b(c_{12}, c_1, p) b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1) b(c_2 - c_{12}, N - c_1, p_2)} \\
&= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\
&\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)
\end{aligned}$$

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

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	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
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

Relative Frequency Ratios (Damerau, 1993)

- ratios of relative frequencies between two or more different corpora can be used to discover collocations that are characteristic of a corpus when compared to other corpora.
- useful for the discovery of subject-specific collocations. The application proposed by Damerau is to compare a general text with a subject-specific text. Those words and phrases that on a relative basis occur most often in the subject-specific text are likely to be part of the vocabulary that is specific to the domain

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

Pointwise Mutual Information

- An information-theoretically motivated measure for discovering interesting collocations is *pointwise mutual information* (Church et al. 1989, 1991; Hindle 1990).
- It is roughly a measure of how much one word tells us about the other.

$$
\begin{aligned}
I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\
&= \log_2 \frac{P(x'|y')}{P(x')} \\
&= \log_2 \frac{P(y'|x')}{P(y')}
\end{aligned}
$$

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	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

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

Problems with using Mutual Information

- Decrease in uncertainty is not always a good measure of an interesting correspondence between two events.
- It is a bad measure of dependence.
- Particularly bad with sparse data.

Normalized Pointwise Mutual Information (Bouma, 2009)

$$i_{\mathrm{n}}(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)}\right) \Big/ -\ln p(x,y). \tag{7}$$

Some orientation values of NPMI are as follows: When two words only occur together, $i_{\mathrm{n}}(x,y) = 1$; when they are distributed as expected under independence, $i_{\mathrm{n}}(x,y) = 0$ as the numerator is 0; finally, when two words occur separately but not together, we define $i_{\mathrm{n}}(x,y)$ to be -1, as it approaches this value when $p(x,y)$ approaches 0 and $p(x), p(y)$ are fixed. For comparison, these orientation values for PMI are respectively $-\ln p(x,y)$, 0 and $-\infty$.[6]

Credits

- This slide set has been adapted from the NLP course of Paul Tarau (based on Rada Mihalcea's original slides):

http://www.cse.unt.edu/~tarau/teaching/NLP/