Collocation

Reporter : 陳燦輝

Reference

- Reference1

- *Foundations of Statistical Natural Language Processing*, Chapter 3

- Pearce, D. "Synonymy in collocation extraction." In Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Pittsburgh, PA.

Outline

- What is collocation
- Why study collocations?
- Approaches to finding collocations
- Summary and Conclusions

What is collocation

- is an expression of 2 or more words that correspond to a conventional way of saying things.
  - *broad daylight*
  - Why not? ?*bright daylight* or ?*narrow darkness*
  - *Big mistake* but not ?*large mistake*
- overlap with the concepts of:
  - *terms*, *technical terms* & *terminological phrases*
    - Collocations extracted form technical domains
      - Ex: *hydraulic oil filter, file transfer protocol*

What is collocation (cont)

- More example :
  - *strong tea*
  - *to check in*
  - *heard it through the grapevine*
  - *he knocked at the door*
  - *...*

What is collocation (cont)

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."

What is collocation (cont)

Criteria:

- non-compositionality
- non-substitutability
- non-modifiability
- non-translatable word for word

What is collocation (cont)

Non-Compositionality

- A phrase is compositional if its meaning can be predicted from the meaning of its parts
  - Ex : *a young man*
- Collocations have limited compositionality
  - there is usually an element of meaning added to the combination
  - Ex: *strong tea*
- Idioms are the most extreme examples of non-compositionality
  - *Ex: to hear it through the grapevine*

What is collocation (cont)

Non-Substitutability

- We cannot substitute near-synonyms for the components of a collocation.
  - *Strong* is a near-synonym of *powerful*
    - *strong tea* ?*powerful tea*
  - *yellow* is as good a description of the color of white wines
    - *white wine* ?*yellow wine*

What is collocation (cont)

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*

What is collocation (cont)

Non-translatable (word for word)

- English:
  - make a decision
- French:
  - ?faire une décision

- to test whether a group of words is a collocation:
  - translate it into another language
  - if we cannot translate it word by word
  - then it probably is a collocation

What is collocation (cont)

Linguistic Subclasses of Collocations

- Phrases with light verbs:
  - Verbs with little semantic content in the collocation
  - *have, do…*
- Proper nouns *(proper names)*
  - *John Smith*
- Terminological expressions
  - concepts and objects in technical domains
  - *hydraulic oil filter*

Why study collocations?

- In nature language generator (NLG)
  - The output should be natural
    - *make a decision* *?take a decision*
- In lexicography
  - Identify collocations to list them in a dictionary
  - To distinguish the usage of synonyms or near-synonyms

Why study collocations (cont)

- In parsing
  - To give preference to most natural attachments
    - *plastic (can opener)*   *? (plastic can) opener*
- In corpus linguistics and psycholinguists
  - Ex: To study social attitudes towards different types of substances
    - *strong cigarettes/tea/coffee*
    - *powerful drug*

Approaches to finding collocations

- Frequency
- Mean and Variance
- Hypothesis Testing
  - t-test
  - χ^2-test (Chi-Square test)
  - Likelihood ratio test
- Mutual Information
- Synonymy in collocation extraction

Approaches to finding collocations (cont)

Frequency (cont)

- (Justeson & Katz, 1995)
- Hypothesis:
  - if 2 words occur together very often, they must be interesting candidates for a collocation
- Method:
  - Select the most frequently occurring bigrams (sequence of 2 adjacent words)

Approaches to finding collocations (cont)

Frequency (cont)

- Not very interesting…
- Except for "*New York*", all bigrams are pairs of function words

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

So, let's pass the results through a part-of-speech filter

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficient*
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*

Approaches to finding collocations (cont)

Frequency + POS filter

Simple method that works very well!

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

Approaches to finding collocations (cont)

Frequency: Conclusion

- Advantages:
  - works well for fixed phrases
  - Simple method & accurate result
  - Requires small linguistic knowledge

- But: many collocations consist of two words in more flexible relationships
  - *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*

Approaches to finding collocations (cont)

Mean and Variance

- (Smadja et al., 1993)
- Looks at the distribution of distances between two words in a corpus
- looking for pairs of words with low variance
  - A low variance means that the two words usually occur at about the same distance
  - A low variance --> good candidate for collocation
- Need a **Collocational Window** to capture collocations of variable distances

	knock			door

	knock					door

Approaches to finding collocations (cont)

Mean and Variance (cont)

- *This is an example of a three word window.*
- Sentence : stocks crash as rescue plan teeters
  - Bigram :

stocks crash | *stocks as* | *stocks rescue*

crash as | *crash rescue* | *crash plan*

Approaches to finding collocations (cont)

Mean and Variance (cont)

- The **mean** is the average offset (signed distance) between two words in a corpus
- The **variance** measures how much the individual offsets deviate from the mean

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

  - n is the number of times the two words (two candidates) co-occur
  - d_i is the offset of the i^{th} pair of candidates
  - $\bar{d}$ is the mean offset of all pairs of candidates

Approaches to finding collocations (cont)

Mean and Variance (cont)

- If offsets (d_i) are the same in all co-occurrences
  - --> variance is zero
  - --> definitely a collocation
- If offsets (d_i) are randomly distributed
  - --> variance is high
  - --> not a collocation

Approaches to finding collocations (cont)

Mean and Variance (cont)

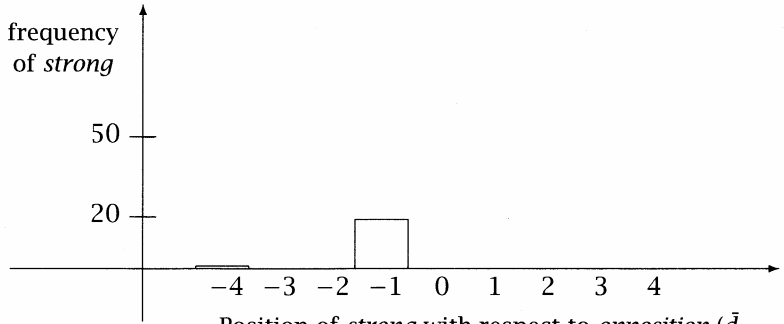
- An Example
  - *she knocked on his door*
  - *they knocked at the door*
  - *100 women knocked on Donaldson's door*
  - *a man knocked on the metal front door*

- Mean d = $\frac{(3+3+5+5)}{4} = 4.0$

- Std. deviation s = $\sqrt{\frac{(3-4.0)^2+(3-4.0)^2+(5-4.0)^2+(5-4.0)^2}{3}} \approx 1.15$

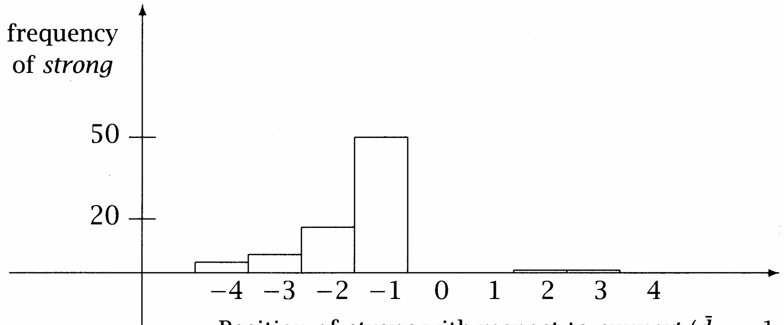
Approaches to finding collocations (cont)

- *"strong…opposition"*
  - variance is low
  - --> interesting collocation



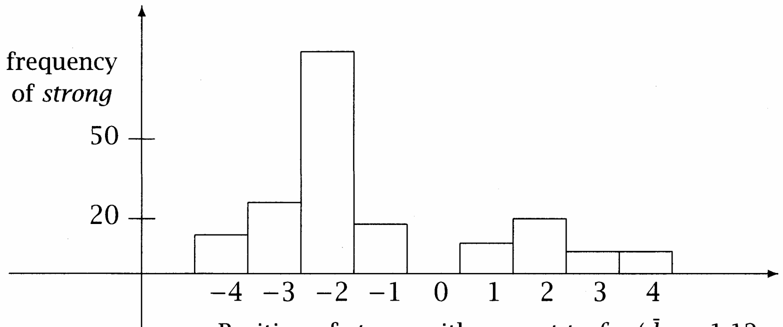
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

- *"strong…support"*



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

- *"strong…for"*
  - variance is high
  - --> not interesting collocation



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Approaches to finding collocations (cont)

- Mean and variance versus Frequency

std. dev. ~0 & mean offset ~1 --> would be found by frequency method

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

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

high deviation --> not interesting

Approaches to finding collocations (cont)

Mean & Variance: Conclusion

- looser relationship between words

- intervening material and relative position

Approaches to finding collocations (cont)

Hypothesis Testing

- If 2 words are frequent… they will frequently occur together…
- Frequent bigrams and low variance can be accidental (two words can co-occur by chance)
- We want to determine whether the co-occurrence is random or whether it occurs more often than chance
- This is a classical problem in statistics called *Hypothesis Testing*
  - When two words co-occur, Hypothesis Testing measures how confident we have that this was due to chance or not

Approaches to finding collocations (cont)

Hypothesis Testing (cont)

- We formulate a *null hypothesis* H_0
  - H_0 : no real association (just chance…)
  - H_0 states what should be true if two words do not form a collocation
  - if 2 words w_1 and w_2 do not form a collocation, then w_1 and w_2 are independently of each other.

Approaches to finding collocations (cont)

Hypothesis Testing: t-test

- or Student's t-test
- H_0 states that: $P(w_1, w_2) = P(w_1)P(w_2)$
- We calculate the probability p-value that H_0 was true
- If p-value is too low, we reject H_0 , Otherwise, retain H_0 as possible
  - Typically if under a *significant level* of *p* < *0.05, 0.01,* or 0.001
- Assume the sample is drawn from a normal distribution

Approaches to finding collocations (cont)

Hypothesis Testing: t-test (cont)

- t-test compares:
  - the sample mean (computed from observed values)
  - to a expected mean
- determines the likelihood (p-value) that the difference between the 2 means occurs by chance.
  - a p-value close to 1 --> it is very likely that the expected and sample means are the same
  - a small p-value (ex: 0.01) --> it is unlikely (only a 1 in 100 chance) that such a difference would occur by chance

Approaches to finding collocations (cont)

Hypothesis Testing: t-test (cont)

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

Difference between the observed mean and the expected mean

$\bar{x}$ is the sample mean
μ is the expected mean of the distribution
s^2 is the sample variance
N is the sample size

the higher the value of t, the greater the confidence that:

- there is a significant difference
- it's not due to chance
- the 2 words are not independent

Approaches to finding collocations (cont)

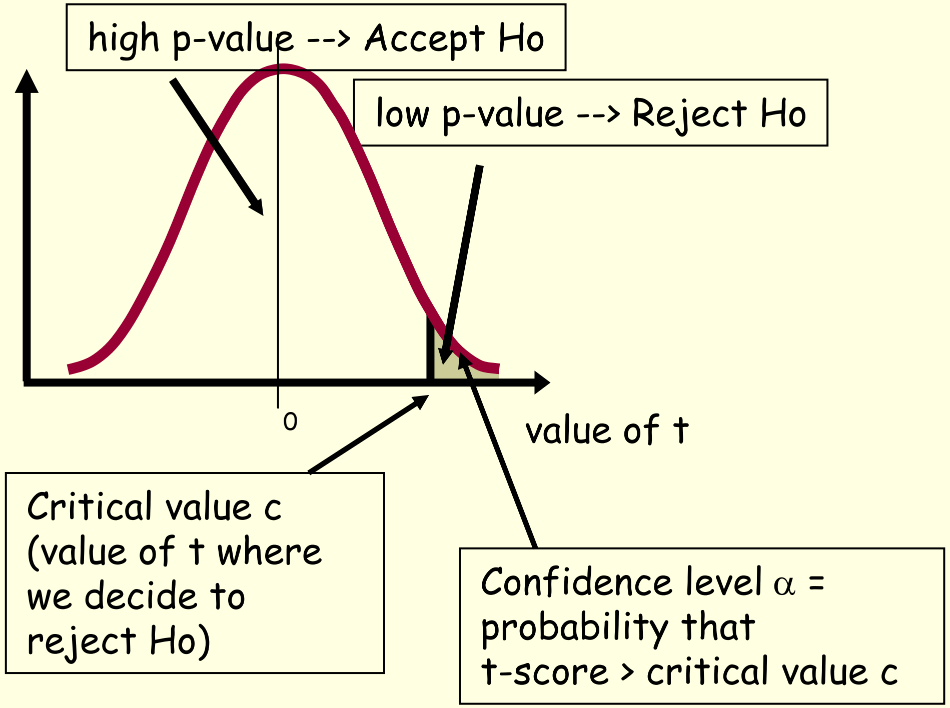
Hypothesis Testing: t-test (cont)

T-distribution

$$f_r(t) = \frac{\Gamma[\frac{1}{2}(r+1)]}{\sqrt{r\pi}\,\Gamma(\frac{1}{2}r)(1+\frac{t^2}{r})^{(r+1)}/2}$$

r : degree of freedom (df)

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$$



Approaches to finding collocations (cont)

Hypothesis Testing: t-test (cont)

- We think of a corpus of N words as a long sequence of N bigrams
- the samples are seen as random variables that:
  - take the value 1 when the bigram of interest occurs
  - take the value 0 otherwise

Approaches to finding collocations (cont)

t-Test: a simple example :

- Null hypothesis is that the mean height of a population of men is 158cm
- We are given a sample of 200 men with x =169 and s^2 = 2600

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

Confidence level of α = 0.005, we fine 2.576
Since the t we got is larger than 2.576, we can reject the null hypothesis with 99.5% confidence. So we can say that the sample is not drawn from a population with mean 158cm, and our probability of error is less than 0.5%

Approaches to finding collocations (cont)

t-Test: Example with collocations

- In a corpus:
  - ***new*** occurs 15,828 times
  - ***companies*** occurs 4,675 times
  - ***new companies*** occurs 8 times
  - there are 14,307,668 tokens overall
- Is ***new companies*** a collocation?
- Null hypothesis:
  - Independence assumption
  - P(*new companies*) = P(*new*) P(*companies*)

$$= \frac{15\,828}{14\,307\,668} \times \frac{4\,675}{14\,307\,668} \approx 3.615 \times 10^{-7}$$

Approaches to finding collocations (cont)

t-Test: Example with collocations (cont)

- If the null hypothesis is true, then:
  - if we randomly generate bigrams of words
  - assign 1 to the outcome *new companies*
  - *assign* 0 to any other outcome
  - …in effect a Bernoulli trial
  - then the probability of having *new companies* is expected to be 3.615×10^{-7}
  - So the **expected** mean is $\mu = 3.615 \times 10^{-7}$
  - The variance $s^2 = p(1-p) \approx p$ since for most bigrams p is small
    - in binomial distribution: $s^2 = np(1-p)$ … but here, n=1

Approaches to finding collocations (cont)

t-Test: Example with collocations (cont)

- But we counted 8 occurrences of the bigram *new companies*
- So the **observed** mean is $\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$
- By applying the t-test, we have: $t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 1$
- With a confidence level α=0.005, critical value is 2.576

	p C	0.05 90%	0.025 95%	0.01 98%	**0.005** **99%**	0.001 99.8%	0.0005 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

- Since t=1 < 2.576
  - we cannot reject the H_o
  - so we cannot claim that *new* and *companies* form a collocation

Approaches to finding collocations (cont)

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

pass the t-test ($t > 2.756$) so: we can reject the null hypothesis so they form collocation

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
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- fail the t-test ($t < 2.756$) so:
- we cannot reject the null hypothesis
- so they do not form a collocation

- Notes:
  - Frequency-based method could not have seen the difference in these bigrams, because they all have the same frequency
  - 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 occurs in the bigram, then its *t* is high.
  - The *t* test is mostly used to rank collocations

Approaches to finding collocations (cont)

t-Test: Hypothesis testing of differences

- Used to see if 2 words (near-synonyms) are used in the same context or not
  - *"strong"* vs *"powerful"*
- can be useful in lexicography
- we want to test:
  - if there is a difference in 2 populations
    - Ex: height of woman / height of man
  - the null hypothesis is that there is no difference
  - i.e. the average difference is 0 (μ =0)

Approaches to finding collocations (cont)

t-Test: Hypothesis testing of differences (cont)

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

$\bar{x}_1$ 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_2 is the sample size of population 2

t	C(w)	C(strong w)	C(powerful w)	Word
3.1622	933	0	10	computers
2.8284	2377	0	8	computer
2.4494	289	0	6	symbol
2.2360	2266	0	5	Germany
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales

Approaches to finding collocations (cont)

χ^2-test

- problem with the t test is that it assumes that probabilities are approximately normally distributed…
- the χ^2-test does not make this assumption
- The essence of the χ^2-test is the same as the t-test
  - Compare observed frequencies and expected frequencies for independence
  - if the difference is large
  - then we can reject the null hypothesis of independence

Approaches to finding collocations (cont)

χ^2-test (cont)

$$\chi^2 = \frac{(O_1 - E_1)^2}{\sigma_1^{\ 2}} + \frac{(O_2 - E_2)^2}{\sigma_2^{\ 2}} + \cdots + \frac{(O_k - E_k)^2}{\sigma_k^{\ 2}}$$

$$= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{\sigma_i^{\ 2}}$$

$$= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \text{(assumption counts are distributed according to the Poisson distribution)}$$

$$= \mathrm{X}^2$$

- sums the differences between observed frequencies
- and expected values for independence
- scaled by the magnitude of the expected values

Approaches to finding collocations (cont)

χ^2-test (cont)

- In the table :

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

- Observed frequencies Obs_{ij}

Observed	w^1 = new	w^1 ≠ new	TOTAL
w^2 = companies	8 (*new companies*)	4 667 (ex: *old companies*)	4 675 c(*companies*)
w^2 ≠ companies	15 820 (ex: *new machines*)	14 287 181 (ex: *old machines*)	14 303 001 c(~*companies*)
TOTAL	15 828 c(*new*)	14 291 848 c(~*new*)	14 307 676 N = 4 675 + 14 303 001 = 15 828 +14 291 848

Approaches to finding collocations (cont)

χ^2-test (cont)

- Expected frequencies Exp_{ij}
  - If independence
  - Computed from the marginal probabilities (the totals of the rows and columns converted into proportions)

Expected	w^1 = new	w^1 ≠ new
w^2 = 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
w^2 ≠ companies	15 822.83 c(new) x c(~companies) / N 15828 x 14303001 /14307676	14 287 178.17 c(~new) x c(~companies) / N 14291848 x 14303001 / 14307676

  - Ex: expected frequency for cell (1,1) (*new companies*)
    - marginal probability of *new* occurring as the first part of a bigram times marginal probability of *companies* occurring as the second part of bigram:

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

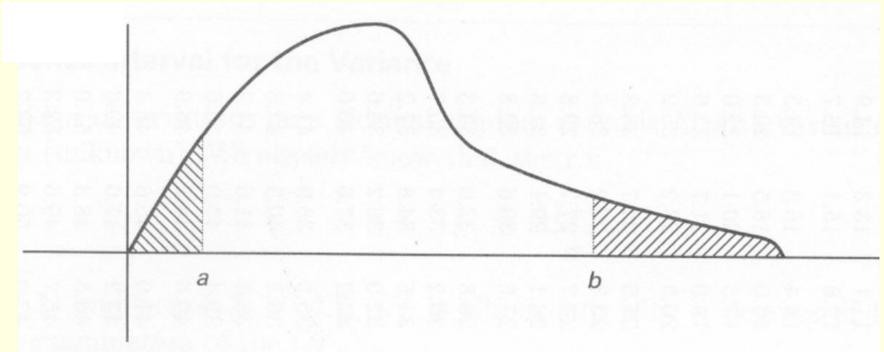
  - If "*new*" and "*companies*" occurred completely independent of each other
  - we would expect 5.17 occurrences of "*new companies*" on average

Approaches to finding collocations (cont)

χ^2-test (cont)

- But is the difference significant?

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(46\,667-46\,669.83)^2}{46\,669} + \frac{(15\,820-15\,822.83)^2}{15\,823} + \frac{(14\,287\,181-14\,287\,178.17)^2}{14\,287\,186} \approx 1.55$$



df in an table = (n-1)(c-1) = (2-1)(2-1) =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

Approaches to finding collocations (cont)

χ^2-test (cont)

- The probability level of α=0.05 the critical value is 3.84
- Since 1.55 < 3.84:
  - So we cannot reject H_0 (that *new* and *companies* occur independently of each other)
  - So *new companies* is not a good candidate for a collocation

Approaches to finding collocations (cont)

χ^2-test for machine translation

- (Church & Gale, 1991)
- To identify translation word pairs in aligned corpora
- Ex:

Nb of aligned **sentence pairs** containing “cow” in English and “vache” in French

Observed frequency	“cow”	~“cow”	TOTAL
“vache”	59	6	65
~“vache”	8	570 934	570 942
TOTAL	67	570 940	571 007

- $\chi^2 = 456\ 400 \gg 3.84$ (with $\alpha = 0.05$)
- So “vache” and “cow” are not independent… and so are translations of each other

Approaches to finding collocations (cont)

χ^2-test for corpus similarity

- (Kilgarriff & Rose, 1998)
- Ex:

Observed frequency	Corpus 1	Corpus 2	Ratio
Word1	60	9	60/9 =6.7
Word2	500	76	6.6
Word3	124	20	6.2
…	…	…	…
Word500	…	…	…

- Compute χ^2 for the 2 populations (corpus1 and corpus2)
- H_o: the 2 corpora have the same word distribution

Approaches to finding collocations (cont)

χ^2-test: Conclusion

- Differences between the *t* statistic and χ^2 statistic do not seem to be large
- But:
  - the χ^2 test is appropriate for large probabilities
    - where t test fails because of the normality assumption
  - the 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/

Approaches to finding collocations (cont)

Likelihood ratios

- It is simply a number that tells us how much more likely one hypothesis is than the other.
- Likelihood ratios are more appropriate for sparse data than the Chi-Square test. In addition, they are easier to interpret than the Chi-Square statistic.

Approaches to finding collocations (cont)

Likelihood ratios (cont)

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

- Hypothesis 1 is a formalization of independence, hypothesis 2 is a formalization of dependence which is good evidence for an interesting collocation

- 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^1w^2 in corpus

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

Approaches to finding collocations (cont)

Likelihood ratios (cont)

- Assuming a binomial distribution: $b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)}$

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

Table 5.11 How to compute Dunning's likelihood ratio test. For example, the likelihood of hypothesis H_2 is the product of the last two lines in the rightmost column.

$$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)$$

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

Approaches to finding collocations (cont)

Likelihood ratios (cont)

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

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

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

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

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

Approaches to finding collocations (cont)

Likelihood ratios (cont)

$-2\log\lambda$	$C(w^1)$	$C(w^2)$	$C(w^1w^2)$	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
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.

H1 is $e^{0.5\times 82.96} \approx 1.3\times 10^{18}$ times more likely than H2

Easier to interpret

Approaches to finding collocations (cont)

Likelihood ratios (cont)

- $-2\log \lambda$ is asymptotically χ^2 distributed (Mood et al. 1974:440)

- The approximation is usually good, even for small sample sizes.

Approaches to finding collocations (cont)

Pointwise Mutual Information

- Uses a measure from information-theory

- Pointwise mutual information between 2 events x and y (in our case the occurrence of 2 words) is roughly:
  - a measure of how much one event (word) tells us about the other
  - or a measure of the independence of 2 events (or 2 words)
    - If 2 events x and y are independent, then I(x,y) = 0

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

Approaches to finding collocations (cont)

Pointwise Mutual Information (cont)

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

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

$$I(Ayatollah, Ruhollah) = \log_2 \left(\frac{\frac{20}{14\,307\,668}}{\frac{42}{14\,307\,668} \times \frac{20}{14\,307\,668}} \right) \approx 18.38$$

- So? The occurrence of “Ayatollah” at position i increases by 18.38bits if “Ruhollah” occurs at position i+1
- works particularly badly with sparse data(favors low frequency events).

Approaches to finding collocations (cont)

Pointwise Mutual Information (cont)

- With pointwise mutual information:

$I(w_1,w_2)$	$C(w_1)$	$C(w_2)$	$C(w_1\ w_2)$	w_1	w_2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

- With t-test (see p.37 of slides)

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
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- Same ranking as t-test

Approaches to finding collocations (cont)

Pointwise Mutual Information (cont)

- good measure of independence
  - values close to 0 --> independence
- bad measure of dependence
  - because score depends on frequency
  - all things being equal, bigrams of low frequency words will receive a higher score than bigrams of high frequency words
  - so sometimes we take $C(w_1\ w_2)\ I(w_1, w_2)$

Approaches to finding collocations (cont)

Pointwise Mutual Information (cont)

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

These examples illustrate that a large proportion of bigrams are not well characterized and that mutual information is particularly sensitive to estimates that are inaccurate due to sparseness.

Approaches to finding collocations (cont)

Synonymy in collocation extraction

- Different between baggage and luggage?

- A new definition of collocation : a pair of words is considered a collocation if one of words significantly prefer a particular lexical realization of the concept the other the represents.

Approaches to finding collocations (cont)

Synonymy in collocation extraction (cont)

- Formalization :
  - A sequence of pairs of words, $p^1 \cdots p^N$
  - The occurrence count of a particular pair of words $< w_a, w_b >$ is defined by $c(w_a, w_b) = \sum_{i=1}^{N} \delta(p^i =< w_a, w_b >)$
  - Where $\delta(x)$ is 1 if x is true and 0 if x is false
  - WordNet is defined as a set of synsets, W, where

$$W = \{S_1, S_2, \cdots\}$$

  - WordNet : http://wordnet.princeton.edu/

Approaches to finding collocations (cont)

Synonymy in collocation extraction (cont)

- Formalization (cont)
  - Each synset consists of a set of words which realize the same concept
  - The co-occurrence set , cs_w of a word, w is defined as : $cs_w = \{w_v : c(w, w_v) > 0\}$
  - Synsets are filtered with respect to w to obtain its Candidate Collocation SynSets CCS_w , is defined as : $CCS_w = \{S \in W : |S \cap cs_w| > 1\}$
  - Thus, each CSS consists of at least two elements whose co-occurrence count with w is non-zero.

Approaches to finding collocations (cont)

Synonymy in collocation extraction (cont)

- Formalization (cont)

$$w' = \arg\max_{w \in S} c(w, w_v)$$

$$f' = \max_{w \in S} c(w, w_v)$$

$$f'' = \max_{w \in S'} c(w, w_v) \text{ where } S' = S - w'$$

$$s = \frac{f' - f''}{f'}$$

A value of $s \approx 1$ indicates high collocation strength and $s \approx 0$ Indicates low

Approaches to finding collocations (cont)

Synonymy in collocation extraction : conclusion

- A assumption that any given synset has one and only one element that forms a collocation with a particular target word.
- Using the Non-Substitutability criterion of collocation.