# Collocations

## Reading: Chap 5, Manning & Schutze

(note: this chapter is available online from the book's page
http://nlp.stanford.edu/fsnlp/promo)

Instructor: Rada Mihalcea

# Outline

What is a collocation?
Automatic approaches 1: frequency-based methods
Automatic approaches 2: ruling out the null hypothesis, t-test
Automatic approaches 3: chi-square and mutual information

# 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 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**

# Outline

What is a collocation?

**Automatic approaches 1: frequency-based methods**

Automatic approaches 2: ruling out the null hypothesis, t-test

Automatic approaches 3: chi-square and 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 s

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

  - 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 sample mean is:

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

And the sample variance:

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

# 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 |

# Outline

What is a collocation?
Automatic approaches 1: frequency-based methods
**Automatic approaches 2: ruling out the null hypothesis, t-test**
Automatic approaches 3: chi-square and mutual information

# 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 real data mean (observed), $s^2$ is the variance, N is the sample size, and μ is the mean of the distribution (expected).

# *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 $s^2 = p(1-p) \approx p^2$

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

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

# 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 (l=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.

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

# Hypothesis testing of differences

Words that co-occur significantly more frequently with *powerful*, and with *strong*

| t | C(w) | C(strong w) | C(powerful w) | Word |
|---|---|---|---|---|
| 3.16 | 933 | 0 | 10 | computers |
| 2.82 | 2337 | 0 | 8 | computer |
| 2.44 | 289 | 0 | 6 | symbol |
| 2.44 | 588 | 0 | 5 | Germany |
| 2.23 | 3745 | 0 | 5 | nation |
| | | | | |
| 7.07 | 3685 | 50 | 0 | support |
| 6.32 | 3616 | 58 | 7 | enough |
| 4.69 | 986 | 22 | 0 | safety |
| 4.58 | 3741 | 21 | 0 | sales |
| 4.02 | 1093 | 19 | 1 | opposition |

# Outline

What is a collocation?
Automatic approaches 1: frequency-based methods
Automatic approaches 2: ruling out the null hypothesis, t-test
**Automatic approaches 3: chi-square and mutual information**

# Pearson's $\chi^2$ (chi-square) test

- *t*-test assumes that probabilities are approximately normally distributed, which is not true in general. The $\chi^2$ test doesn't make this assumption.
- the essence of the $\chi^2$ 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

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

# $\chi^2$ 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 $\chi^2$ 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.

# $\chi^2$ 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

$\chi^2$ is then determined as 1.55

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

# 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.

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