# Statistical NLP: Lecture 7

Collocations

(Ch 5)

## Introduction

- Collocations are characterized by limited ***compositionality***.
- Large overlap between the concepts of ***collocations*** and ***terms***, ***technical term*** and ***terminological phrase***.
- Collocations sometimes reflect interesting attitudes (in English) towards different types of substances: ***strong*** cigarettes, tea, coffee versus ***powerful*** drug (e.g., heroin)

## Definition (w.r.t Computational and Statistical Literature)

- [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. [Chouekra, 1988]

## Other Definitions/Notions (w.r.t. Linguistic Literature)

- Collocations are not necessarily adjacent
- Typical criteria for collocations: non-compositionality, non-substitutability, non-modifiability.
- Collocations cannot be translated into other languages.
- Generalization to weaker cases (strong association of words, but not necessarily fixed occurrence.

## Linguistic Subclasses of Collocations

- Light verbs: verbs with little semantic content
- Verb particle constructions or Phrasal Verbs
- Proper Nouns/Names
- Terminological Expressions

## Overview of the Collocation Detecting Techniques Surveyed

- Selection of Collocations by ***Frequency***
- Selection of Collocation based on ***Mean and Variance*** of the distance between focal word and collocating word.
- ***Hypothesis Testing***
- ***Mutual Information***

## Frequency (Justeson & Katz, 1995)

1. Selecting the most frequently occurring bigrams
2. Passing the results through a part-of-speech filter

☑ Simple method that works very well.

## Mean and Variance (I)
## (Smadja et al., 1993)

- Frequency-based search works well for fixed phrases. However, many collocations consist of two words in more flexible relationships.
- The method computes the mean and variance of the offset (signed distance) between the two words in the corpus.
- If the offsets are randomly distributed (i.e., no collocation), then the variance/sample deviation will be high.

## Mean and Variance (II)

- n = number of times two words collocate
- μ = sample mean
- $d_i$ = the value of each sample
- Sample deviation:

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

## Example of flexible collocation

knocked on the door (3)
knocked at the door (3)
knocked on John’s door (5)
knocked on the metal front door (5)

μ = (3+3+5+5)/4 = 4
s = sqrt($(3-4)^2$ + $(3-4)^2$ + $(5-4)^2$ + $(5-4)^2$) / 3) = 1.15

If s is big => no collocation
If μ is not zero => flexible collocation

## Hypothesis Testing: Overview

- High frequency and low variance can be accidental. 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***.
- We formulate a ***null hypothesis*** $\boldsymbol{H_0}$ (no association - only chance) and calculate the probability p that a collocation would occur if $H_0$ were true, and then ***reject*** $\boldsymbol{H_0}$ if p is too low. Otherwise, retain $H_0$ as possible.

## Hypothesis Testing: The *t*-test

- 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 $\mu$.
- 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 $\mu$.
- To apply the ***t***-test to collocations, we think of the text corpus as a long sequence of N bigrams.

## Hypothesis Testing: Formula

N = number of bigrams

$\mu$ = sample mean for $H_0$

$\overline{x}$ = observed sample mean

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

p = probability that the event would occur if $H_0$ were true

Significance level

$p < 0.05$ means 95% confidence

$p < 0.01$ means 99% confidence

## Example
## *new companies* – collocation or not?

| | $w_1$ = new | $w_1 \neq$ new |
|---|---|---|
| $w_2$ = companies | $O_{11} = 8$ | $O_{12} = 4667$ |
| $w_2 \neq$ companies | $O_{21} = 15820$ | $O_{22} = 14287173$ |

P(new) = (15820 + 8) / 14307668

P(companies) = (4667 + 8) / 14307668

$H_0$: P(new companies) = P(new) * P(companies) = 0.0000003615 = $\mu$

$\overline{x}$ = 8 / 14307668 = 0.0000005591

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

$$t = \frac{0.0000005591 - 0.0000003615}{\sqrt{\frac{0.0000005591}{14307668}}} = 0.999932 < 2.576$$ => We cannot reject null hypothesis

# Hypothesis testing of differences (Church & Hanks, 1989)

- We may also want to find words whose co-occurrence patterns best distinguish between two words. This application can be useful for lexicography.
- 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.

# Hypothesis testing of difs. (II)

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

$P(w_1\ v) = C(w_1\ v) / N$

$P(w_2\ v) = C(\ w_2\ v) / N$

Example: *strong tea* vs. *powerful tea*

$$t = \frac{C(w_1 v) - C(w_2 v)}{\sqrt{C(w_1 v) + C(w_2 v)}}$$

## t-test for statistical significance of the difference between two systems

| | System 1 | System 2 |
|---|---|---|
| scores | 71,61,55,60,68,49,<br>42,72,76,55,64 | 42,55,75,45,54,51,<br>55,36,58,55,67 |
| total | 673 | 593 |
| n | 11 | 11 |
| Mean $\bar{X}_i$ | 61.2 | 53.9 |
| $s_i$^2 = sum ($x_{ij}$ - $\bar{x}_i$)^2 | 1081.6 | 1186.9 |
| df | 10 | 10 |

## t-test for differences (continued)

- Pooled $s^2 = (1081.6 + 1186.9) / (10 + 10) = 113.4$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2\,s^2}{n}}} = \frac{61.2 - 53.9}{\sqrt{\frac{2\ 113.4}{11}}} = 1.60$$

- For rejecting the hypothesis that System 1 is better then System 2 with a probability level of $\alpha = 0.05$, the critical value is t=1.725 (from statistics table)
- We cannot conclude the superiority of System 1 because of the large variance in scores

## Chi-Square test (I): Method

- Use of the ***t***-test has been criticized because it assumes that probabilities are approximately normally distributed (not true, generally).
- The Chi-Square test does not make this assumption.
- The essence of the test is to compare observed frequencies with frequencies expected for independence. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

## Chi-Square test (II): Formula

$$X^2 = \sum_{i,j=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$E_{11} = \frac{O_{11} + O_{12}}{N} \times \frac{O_{11} + O_{21}}{N} \times N$$

$$E_{12} = \ldots\ldots.; E_{21} = \ldots\ldots.; E_{22} = \ldots\ldots.$$

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

# Chi-Square test (III): Applications

- One of the early uses of the Chi square test in Statistical NLP was the identification of translation pairs in aligned corpora (Church & Gale, 1991).
- A more recent application is to use Chi square as a metric for corpus similarity (Kilgariff and Rose, 1998)
- Nevertheless, the Chi-Square test should not be used in small corpora.

# Example
## *new companies* – collocation or not?

| | $w_1$ = new | $w_1 \neq$ new |
|---|---|---|
| $w_2$ = companies | $O_{11} = 8$ | $O_{12} = 4667$ |
| $w_2$ = companies | $O_{21} = 15820$ | $O_{22} = 14287173$ |

$E_{ij}$ = marginal probabilities = totals of row i and column j converted into proportions = expected values for independence

$X^2 = 1.55 < 3.841$ needed for $p < 0.05$, one degree of freedom for 2x2 table

## Likelihood Ratios I: Within a single corpus (Dunning, 1993)

- 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.
- In applying the likelihood ratio test to collocation discovery, we examine the following two alternative explanations for the occurrence frequency of a bigram w1 w2:
  - The occurrence of w2 is independent of the previous occurrence of w1
  - The occurrence of w2 is dependent of the previous occurrence of w1

## Log likelihood

$$H_1 : P(w_2 \mid w_1) = P(w_2 \mid \neg w_1) = p$$

$$H_2 : P(w_2 \mid w_1) = p_1 \neq p_2 = P(w_2 \mid \neg w_1)$$

$$p = \frac{c_2}{N}; p_1 = \frac{c_{12}}{c_1}; p_2 = \frac{c_2 - c_{12}}{N - c_1}; c_1 = C(w_1); c_2 = C(w_2); c_{12} = C(w_1 w_2);$$

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

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

## Likelihood Ratios II: Between two or more corpora (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.
- This approach is most useful for the discovery of subject-specific collocations.

## Mutual Information (I)

- An information-theoretic measure for discovering collocations is ***pointwise mutual information*** (Church et al., 89, 91)
- Pointwise Mutual Information is roughly a measure of how much one word tells us about the other.
- Pointwise mutual information does not work well with sparse data.

## Mutual Information (II)

$$MI(x,y) = P(X,Y)\log\frac{P(x,y)}{P(x)P(y)}$$

$$PMI(x,y) = \log\frac{P(x,y)}{P(x)P(y)}$$

$$PMI(x,y) = \log\frac{C(x,y)N}{C(x)C(y)}$$

PMI = E (MI)

## Example

PMI(new, companies) =
= log ((8 * 14307668) / (4675 * 15828)) = 1.546

PMI(house, commons) = 4.2
PMI(videocasette, recorder) = 15.94