Words, Collocations, Hypothesis Testing

Arjun Mukherjee†

Course webpage:
http://www.cs.uh.edu/~arjun/courses/nlp

† Contains contents from [Manning et al., 2008] , and various other sources. Referenced in place.

Collocations

- Juxtaposing of words corresponding to a convention in language usage.

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

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

- **Q: How to find them?**

Collocations: Frequency vs. Informative

- **Q: Which out of these collocations are more informative?**

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

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

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

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

- A: Collocations with subject (Noun /Noun Phrases) and modifiers (Adjectives) tend to yield informative patterns [Justeson and Katz, 1995]

Hypothesis Testing

- **Key question:** Does am event occur more often than just by chance?
  - Do two consecutive words (w_1, w_2) form a collocation?
- **Null Hypothesis(H_0):** Assume there is no association between w_1, w_2
- **Test statistic:** Compute the probability, p that the event would occur assuming H_0 holds using a statistical test (t-test, χ^2 test, Fisher's exact test etc.)
- ***p*-value:** Probability, p that the test statistic result (of the occurrence of the event) assuming/given H_0 is true/holds.
- **Interpretation:** if $p<0.05$, we can reject H_0 and accept the alternate hypothesis (i.e., (w_1, w_2) form a collocation) is 95% confidence.

Hypothesis Testing: Significance Levels

- Thumb rule for interpreting levels of significance.

  - $p \leq 0.01$: very strong presumption against null hypothesis
  - $0.01 < p \leq 0.05$: strong presumption against null hypothesis
  - $0.05 < p \leq 0.1$: low presumption against null hypothesis
  - $p > 0.1$: no presumption against the null hypothesis

t-Test

- A common test used for testing collocations
- Looks at the mean and variance of a sample of measurements
- **Null hypothesis (H_0):** Sample is drawn from a Normal distribution with mean μ
- **Test statistic:** Looks at difference between observed and expected means, scaled by sample variance.

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

where $\bar{x}$ is the sample mean, s^2 is the sample variance, N is the sample size, and μ is the mean of the distribution. If the t statistic is large enough we can reject the null hypothesis. We can find out exactly how large it has to be by looking up the table of the t distribution we have compiled in

t-Test

- **Interpretation/Key Question answered:** How likely is one to get a sample of that (observed) mean and variance assuming the sample was drawn from a Normal distribution with mean μ ?

- p-value of t-test can be looked up using a t-table.

Here's an example of applying the t test. Our null hypothesis is that the mean height of a population of men is 158cm. We are given a sample of 200 men with $\bar{x} = 169$ and $s^2 = 2600$ and want to know whether this sample is from the general population (the null hypothesis) or whether it is from a different population of smaller men. This gives us the following t according to the above formula:

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

- A value of t = 3.05 corresponds to p ~ 0.005.
  What does this mean? Can we say the sample was drawn from a population which is distributed $N(\mu = 158cm)$?

Ascertaining collocations using *t*-Test

- Use t-test for finding the likelihood of (w1, w2) being a collocation vs. appearing just by chance ?

To see how to use the *t* test for finding collocations, let us compute the *t* value for *new companies*. What is the sample that we are measuring the mean and variance of? There is a standard way of extending the *t* test for use with proportions or counts. We think of the text corpus as a long sequence of *N* bigrams, and the samples are then indicator random variables that take on the value 1 when the bigram of interest occurs, and are 0 otherwise.

Using maximum likelihood estimates, we can compute the probabilities of *new* and *companies* as follows. In our corpus, *new* occurs 15,828 times, *companies* 4,675 times, and there are 14,307,668 tokens overall.

$$P(new) = \frac{15828}{14307668}$$

$$P(companies) = \frac{4675}{14307668}$$

Ascertaining collocations using *t*-Test

The null hypothesis is that occurrences of *new* and *companies* are independent.

$$H_0 : P(new\ companies) = P(new)P(companies) = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

If the null hypothesis is true, then the process of randomly generating bigrams of words and assigning 1 to the outcome *new companies* and 0 to any other outcome is in effect a Bernoulli trial with $p = 3.615 \times 10^{-7}$ for the probability of *new company* turning up. The mean for this distribution is $\mu = 3.615 \times 10^{-7}$ and the variance is $\sigma^2 = p(1 - p)$ (see section 2.1.9), which is approximately p. The approximation $\sigma^2 = p(1 - p) \approx p$ holds since for most bigrams p is small.

It turns out that there are actually 8 occurrences of *new companies* among the 14,307,668 bigrams in our corpus. So, for the sample, we have that the sample mean is: $\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$. Now we have everything we need to apply the t test:

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

- Assuming sample variance to be same as that of the Bernoulli trial of H_0
- t = 0.99 corresponds to p ~ 0.25.
- What can we say about "new companies" then? Do they form a collocation? Or they occur independently?

χ^2 test

- Drawback of t-test: Assumes samples are normally distributed in both alternate/null hypothesis).

- Chi squared (χ^2) test directly compare the difference between observed vs. expected frequencies.

- For large differences, we reject H_0 where H_0 is the null hypothesis that the no significant difference exist between observed and expected frequencies.

χ^2 test

- A contingency table summarizes the frequency distribution of variables.

	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*)

Table 5.8 A 2-by-2 table showing the dependence of occurrences of *new* and *companies*. There are 8 occurrences of *new companies* in the corpus, 4,667 bigrams where the second word is *companies*, but the first word is not *new*, 15,820 bigrams with the first word *new* and a second word different from *companies*, and 14,287,181 bigrams that contain neither word in the appropriate position.

Table 5.8 shows the distribution of *new* and *companies* in the reference corpus that we introduced earlier. Recall that $C(new) = 15,828$, $C(companies) = 4,675$, $C(new\ companies) = 8$, and that there are 14,307,668 tokens in the corpus. That means that the number of bigrams $w_i w_{i+1}$ with the first token not being *new* and the second token being *companies* is $4667 = 4675 - 8$. The two cells in the bottom row are computed in a similar way.

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

- We can express $E_{i,j}$ as a function of $O_{i,j}$

The expected frequencies E_{ij} are computed from the marginal probabilities, that is, from the totals of the rows and columns converted into proportions. For example, the expected frequency for cell (1,1) (*new companies*) would be the marginal probability of *new* occurring as the first part of a bigram times the marginal probability of *companies* occurring as the second part of a bigram (multiplied by the number of bigrams in the corpus):

$$\frac{8+4667}{N} \times \frac{8+15820}{N} \times N \approx 5.2$$

That is, if *new* and *companies* occurred completely independently of each other we would expect 5.2 occurrences of *new companies* on average for a text of the size of our corpus.

The χ^2 test can be applied to tables of any size, but it has a simpler form for 2-by-2 tables: (see exercise 5.9)

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

This formula gives the following χ^2 value for table 5.8:

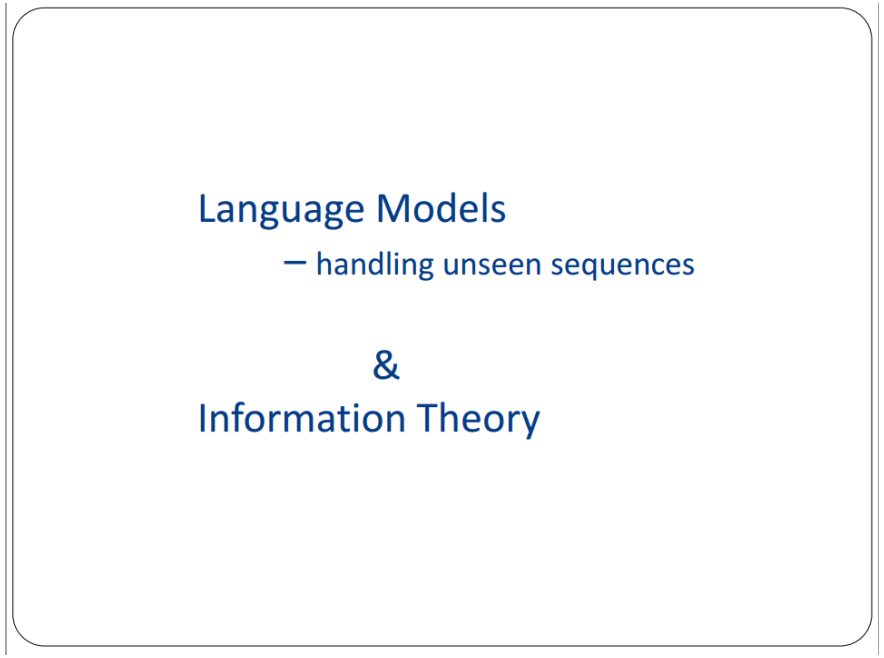
$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8+4667)(8+15820)(4667+14287181)(15820+14287181)} \approx 1.55$$

- $\chi^2 \approx 1.55 \rightarrow p \approx 0.975$ using table look up. Can we say "new companies" is a collocation based on χ^2 test?

N-gram Language models

- We refer to slides by Y. Choi

Part (1)



Part (2)

- Focus on slides (4, 29) in Part (1) and slides (2-10) in Part (2)