Collocations

L645
Advanced NLP
Fall 2009
**Collocation**

**Collocations** are characteristic co-occurrence patterns of two (or more) lexical items

- Tend to occur with greater than random chance
- The meaning tends to be more than the sum of its parts

These are extremely hard to define by intuition:

- Pro: Corpora have been able to reveal connections previously unseen
- Con: It’s not always clear what the theoretical basis of collocations are
Defining a collocation

“People disagree on collocations”

• Intuition does not seem to be a completely reliable way to figure out what a collocation is

• Many collocations are overlooked: people notice unusual words & structures, but not ordinary ones

But what your collocations are depends on exactly how you calculate them

• There is some notion that they are more than the sum of their parts

So, how can we practically define a collocation? . . .
Collocations

**Collocations** are expressions of two or more words that are in some sense conventionalized as a group

- *strong tea* (cf. *powerful tea*)
- *international best practice*
- *kick the bucket*

In examining collocations, we are placing an importance on the context: “You shall know a word by a company it keeps” (Firth 1957)

- In other words, there are lexical properties that more general syntactic properties do not capture
A **colligation** is a slightly different concept:

- collocation of a node word with a particular class of words (e.g., determiners)

Colligations often create “noise” in a list of collocations

- e.g., *this house* because *this* is so common on its own, and determiners appear before nouns
- Thus, people sometimes use stop words to filter out non-collocations, as we will see
Kinds of Collocations

Collocations come in different guises:

- **Light verbs**: verbs convey very little meaning but must be the right one:
  - *make a decision* vs. *take a decision*, *take a walk* vs. *make a walk*
- **Phrasal verbs**: main verb and particle combination, often translated as a single word:
  - *to tell off*, *to call up*
- **Proper nouns**: slightly different than others, but each refers to a single idea (e.g., *Brooks Brothers*)
- **Terminological expressions**: technical terms that form a unit (e.g., *hydraulic oil filter*)
A note on semantic prosody

**Semantic prosody** = “a form of meaning which is established through the proximity of a consistent series of collocates” (Louw 2000)

- These are typically negative: e.g., *peddle, ripe for, get oneself [VERBED]*

- The idea is that you can tell the semantic prosody of a word by the types of words it frequently co-occurs with

This type of co-occurrence often leads to general semantic preferences

- e.g., *utterly, totally*, etc. typically have a feature of ‘absence or change of state’
Properties of a colocation

Prototypically, collocations meet the following criteria:

- Non-compositional: meaning of *kick the bucket* not composed of meaning of parts
  - More subtly: *red/white hair* sort of composable, but the color of *red/white* here is different than usual
- Non-substitutable: *orange hair* just as accurate as *red hair*, but we don’t (can’t?) say it
- Non-modifiable: often we cannot modify a collocation, even though we normally could modify one of those words: *kick the red bucket*
Compositionality tests

The previous properties are good tests, but hard to verify with corpus data

(At least) two tests we can use with corpora:

• Is the collocation translated word-by-word into another language?
  – e.g., Collocation *make a decision* is not translated literally into French

• Do the two words co-occur more frequently together than we would otherwise expect?
  – e.g., *of the* is frequent, but both words are frequent, so we might expect this

We will discuss this second test in detail
The simplest thing to do to find collocations is to use frequency counts: two words appearing together a lot are a collocation.

The problem is that we get lots of uninteresting pairs of function words (table 5.1)

<table>
<thead>
<tr>
<th>C(w₁, w₂)</th>
<th>w₁</th>
<th>w₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>80871</td>
<td>of</td>
<td>the</td>
</tr>
<tr>
<td>58841</td>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>26430</td>
<td>to</td>
<td>the</td>
</tr>
<tr>
<td>21842</td>
<td>on</td>
<td>the</td>
</tr>
</tbody>
</table>
POS Filtering

To remove frequent pairings which are uninteresting, we can use a POS filter (Justeson and Katz 1995)

• only examine word sequences which fit a particular part-of-speech pattern:
  A N, N N, A A N, A N N, N A N, N N N, N P N

  A N        linear function
  N A N      mean squared error
  N P N      degrees of freedom

• Crucially, all other sequences are removed

  P D        of the
  MV V      has been
POS filtering (2)

Some results after tag filtering (Table 5.3)

<table>
<thead>
<tr>
<th>C(w₁, w₂)</th>
<th>w₁</th>
<th>w₂</th>
<th>Tag Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>11487</td>
<td>New</td>
<td>York</td>
<td>A N</td>
</tr>
<tr>
<td>7261</td>
<td>United</td>
<td>States</td>
<td>A N</td>
</tr>
<tr>
<td>5412</td>
<td>Los</td>
<td>Angeles</td>
<td>N N</td>
</tr>
<tr>
<td>3301</td>
<td>last</td>
<td>year</td>
<td>A N</td>
</tr>
</tbody>
</table>

⇒ Fairly simple, but surprisingly effective

- This would need to be refined to handle verb-particle collocations
- Also, kind of inconvenient to write out patterns you want
Longer Distance Connections

Two words may commonly go together, but they may not be strictly collocational, i.e., they may not be right next to each other, as in *knock* and *door*:

(1) she knocked on his door
(2) they knocked at the door
(3) 100 women knocked on Donaldson’s door
(4) a man knocked on the metal front door

So, how can we tell if they’re related?
Offsets: Mean and Variance

Generally, words that go together appear near each other, so we can examine the offset between the two words, *knocked* and *door*

- Mean, or average offset: \( \bar{x} = \frac{\sum d_i}{n} \), where \( d_i \) is each offset, and \( n \) is the total number of examples

\[
(5) \quad \bar{x} = \frac{3+3+5+5}{4} = 4.0
\]

- Variance measures how far off each offset is from the mean:

\[
(6) \quad s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1} \approx 1.33
\]

- Standard deviation \( s \) is the square root of \( s^2 \) and is about 1.15 in this case

A low deviation means that the mean is a pretty accurate indicator of the distance
It is a lot of calculations to look at the offsets for every possible pair of words

- Can restrict the search to be within a window of a set number of words, e.g., 5

The standard deviation gives us useful information—i.e., the words restrict one another in position

- If we search for “bigrams at a distance,” then we can use all the other techniques we’ll talk about

For now, we'll focus on words next to each other ...
Determining strength of collocation

When we have two words appearing next to each other, we want to compare the likelihood of this being a chance event vs. it being a surprise.

- In other words, do the two words appear next to each other more than we might expect, based on what we know about their individual frequencies?
  - Is this an accidental pairing or not?

- We will look at different techniques which define this differently.

- The more data we have, the more confident we will be in our assessment of a collocation or not.

We’ll look at bigrams, but the techniques will work for words within five words of each other, translation pairs, phrases, etc.
(Pointwise) Mutual Information

One way to see if two words are strongly connected is to compare

- the probability of the two words appearing together if they are independent \( p(w_1)p(w_2) \)
- the actual probability of the two words appearing together \( p(w_1w_2) \)

The pointwise mutual information is a measure to do this:

\[
I(w_1, w_2) = \log \frac{p(w_1w_2)}{p(w_1)p(w_2)}
\]
Pointwise Mutual Information Equation

Our probabilities \((p(w_1w_2), p(w_1), p(w_2))\) are all basically calculated in the same way:

\[
\begin{align*}
8. & \quad p(x) = \frac{C(x)}{N} \\
9. & \quad I(w_1, w_2) = \log \frac{p(w_1w_2)}{p(w_1)p(w_2)} = \log \frac{C(w_1w_2)}{\frac{C(w_1)C(w_2)}{N}} = \log[N \frac{C(w_1w_2)}{C(w_1)C(w_2)}]
\end{align*}
\]

- \(N\) is the number of words in the corpus
- The number of bigrams \(\approx\) the number of unigrams
Mutual Information example

We want to know if *Ayatollah Ruhollah* is a collocation in a data set we have:

- \( C(\text{Ayatollah}) = 42 \)
- \( C(\text{Ruhollah}) = 20 \)
- \( C(\text{AyatollahRuhollah}) = 20 \)
- \( N = 14307668 \)

(10) \[ I(\text{Ayatollah}, \text{Ruhollah}) = \log_2 \frac{20}{\frac{42}{N} \times \frac{20}{N}} = \log_2 N \frac{20}{42 \times 20} \approx 18.38 \]

To see how good a collocation this is, we need to compare it to others
Problems for Mutual Information

The formula we have also has the following equivalencies:

\[
I(w_1, w_2) = \log \frac{p(w_1w_2)}{p(w_1)p(w_2)} = \log \frac{P(w_1|w_2)}{P(w_1)} = \log \frac{P(w_2|w_1)}{P(w_2)}
\]

Mutual information tells us how much more information we have for a word, knowing the other word

- But a decrease in uncertainty isn’t quite right ...

A few problems:

- Sparse data: infrequent bigrams for infrequent words get high scores
- Tends to measure independence (value of 0) better than dependence
- Doesn’t account for how often the words do not appear together (Table 5.15)
Motivating Contingency Tables

What we can instead get at is: which bigrams are likely, out of a range of possibilities?

Looking at the Arthur Conan Doyle story *A Case of Identity*, we find the following possibilities for one particular bigram:

- *sherlock* followed by *holmes*
- *sherlock* followed by some word other than *holmes*
- some word other than *sherlock* preceding *holmes*
- two words: the first not being *sherlock*, the second not being *holmes*

These are all the relevant situations for examining this bigram
Contingency Tables

We can count up these different possibilities and put them into a contingency table (or 2x2 table)

<table>
<thead>
<tr>
<th></th>
<th>$B = \text{holmes}$</th>
<th>$B \neq \text{holmes}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A = \text{sherlock}$</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>$A \neq \text{sherlock}$</td>
<td>39</td>
<td>7059</td>
<td>7098</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>7059</td>
<td>7105</td>
</tr>
</tbody>
</table>

The Total row and Total column are the **marginals**

- The values in this chart are the observed frequencies ($f_o$)
Observed bigram probabilities

Because each cell indicates a bigram, divide each of the cells by the total number of bigrams (7105) to get probabilities:

<table>
<thead>
<tr>
<th></th>
<th>holmes</th>
<th>¬ holmes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sherlock</td>
<td>0.00099</td>
<td>0.0</td>
<td>0.00099</td>
</tr>
<tr>
<td>¬ sherlock</td>
<td>0.00549</td>
<td>0.99353</td>
<td>0.99901</td>
</tr>
<tr>
<td>Total</td>
<td>0.00647</td>
<td>0.99353</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The marginal probabilities indicate the probabilities for a given word, e.g., \( p(\text{sherlock}) = 0.00099 \) and \( p(\text{holmes}) = 0.00647 \).
Expected bigram probabilities

If we assumed that *sherlock* and *holmes* are independent—i.e., the probability of one is unaffected by the probability of the other—we would get the following table:

<table>
<thead>
<tr>
<th></th>
<th>holmes</th>
<th>¬ holmes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sherlock</td>
<td>0.00647 x 0.00099</td>
<td>0.99353 x 0.00099</td>
<td>0.00099</td>
</tr>
<tr>
<td>¬ sherlock</td>
<td>0.00647 x 0.99901</td>
<td>0.99353 x 0.99901</td>
<td>0.99901</td>
</tr>
<tr>
<td>Total</td>
<td>0.00647</td>
<td>0.99353</td>
<td>1.0</td>
</tr>
</tbody>
</table>

* This is simply $p_e({w_1, w_2}) = p(w_1)p(w_2)$
## Expected bigram frequencies

Multiplying by 7105 (the total number of bigrams) gives us the expected number of times we should see each bigram:

<table>
<thead>
<tr>
<th></th>
<th>holmes</th>
<th>¬ holmes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sherlock</td>
<td>0.05</td>
<td>6.95</td>
<td>7</td>
</tr>
<tr>
<td>¬ sherlock</td>
<td>45.5</td>
<td>7052.05</td>
<td>7098</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>7059</td>
<td>7105</td>
</tr>
</tbody>
</table>

- The values in this chart are the expected frequencies ($f_e$)
Pearson’s chi-square test

The chi-square ($\chi^2$) test measures how far the observed values are from the expected values:

(12) \[ \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \]

\[
\chi^2 = \frac{(7-0.05)^2}{0.05} + \frac{(0-6.95)^2}{6.95} + \frac{(39-45.5)^2}{45.5} + \frac{(7059-7052.05)^2}{7052.05}
\]

(13) \[ = 966.05 + 6.95 + 1.048 + 0.006 \]

\[ = 974.05 \]

If you look this up in a table, you’ll see that its unlikely to be chance

NB: The $\chi^2$ test does not work well for rare events, i.e., $f_e < 6$
Table lookup?

Are we saying that to find strong collocations we have to look up their $\chi^2$ values in a table?

- No: the values we get for these different tests are most useful in comparison to values for other bigrams in the corpus

- 974.05 isn’t interesting in itself; the fact that it is higher than almost any other bigram’s value is interesting.
(Other) Hypothesis Testing: $t$ test

What we’re doing with the $\chi^2$ test is **hypothesis testing** = can we accept or reject the *null hypothesis* that the two words are independent?

- Null hypothesis ($H_0$): no association between the words beyond chance
  
  i.e., independent: $p(w_1, w_2) = p(w_1)p(w_2)$

- Like the $\chi^2$ test, we compare the expected probability to the observed one:

\begin{equation}
(t) \quad t = \frac{\bar{x} - \mu}{s^2 \sqrt{\frac{N}{N}}} \approx \frac{p_o - p_e}{\sqrt{\frac{p_o}{N}}}
\end{equation}

High values of $t$ indicate a strong collocation.

- $t$ test can be used to compare co-occurrence patterns (e.g. *strong* vs. *powerful*)
Likelihood ratio

A likelihood ratio calculates the ratio between a hypothesis of independence and a hypothesis of dependence:

- **Hypothesis 1** (independence): \( p(w_2|w_1) = p = p(w_2|\neg w_1) \)

- **Hypothesis 2** (dependence) \( p(w_2|w_1) = p_1 \neq p_2 = p(w_2|\neg w_1) \)

We use maximum likelihood estimates to get \( p, p_1, \) and \( p_2, \) and then calculate the log likelihood ratio as follows:

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

where \( L(H) \) is the likelihood of \( H \) and is calculated by assuming a binominal distribution. ... In other words, we test the hypothesis that the \(|A|\) and \(|\neg A|\) rows in the 2x2 table came from the same binominal distribution.

NB: Better for sparse data than the other tests (but not perfect)
Taking it from here

So, what’s a collocationist to do?

- All methods require the same information—i.e., frequencies of two individual words appearing independently vs. appearing together

- Study up on what test is most appropriate for your particular data.

- Always hand-examine when you’re done to make sure it was indeed appropriate.

A good site to use is: http://www.collocations.de