Corpus Linguistics
Collocations

Defining a collocation

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

1. Firthian definition: combinations of words that co-occur more frequently than by chance
   - “You shall know a word by the company it keeps” (Firth 1957)
2. Phraseological definition: The meaning tends to be more than the sum of its parts
   - “a sequence of two or more consecutive words, . . . whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components” (Choueka 1988)

Some examples by different definitions:
- Firth + phraseology: couch potato
- Firth only: potato peeler
- Phraseology only: broken record

Corpora have been able to reveal connections previously unseen

Thus, people sometimes use stop words to filter out non-collocations

Semantic prosody = “a form of meaning which is established through the proximity of a consistent series of collocates” (Louw 2000)
- Idea: you can tell the semantic prosody of a word by the types of words it frequently co-occurs with
  - These are typically negative: e.g., peddle, ripe for, get oneself verbed
  - 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’

Colligations

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

Related concepts
- Colligations
- Semantic prosody & preference
Defining a collocation
Towards corpus-based metrics

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

Importance of the context: "You shall know a word by a company it keeps" (Firth 1957)
- There are lexical properties that more general syntactic properties do not capture

This slide and the next 3 adapted from Manning and Schütze (1999), *Foundations of Statistical Natural Language Processing*

Prototypical collocations

Prototypically, collocations meet the following criteria:

- Non-compositional: meaning of kick the bucket not composed of meaning of parts
- Non-substitutable: orange hair just as accurate as red hair, but some don’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

Kinds of collocations

Calculations ideally take into account variability:

- Light verbs: verbs convey very little meaning but must be the right one:
  - make take a decision, take 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)
- Syntactically adaptable expressions: bite biting hit the dust, take leave of your senses
- Non-adjacent collocations: faint stale apricot smell

Calculating collocations

Simplest approach: 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 (M&S 1999, table 5.1)

\[
C(w_1, w_2) = \frac{w_1 \cdot w_2}{\text{observed} - \text{expected}}
\]

(from Gries 2009)

<table>
<thead>
<tr>
<th>w_1</th>
<th>w_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>80871</td>
<td>the</td>
</tr>
<tr>
<td>58841</td>
<td>in</td>
</tr>
<tr>
<td>26430</td>
<td>to</td>
</tr>
<tr>
<td>21842</td>
<td>on</td>
</tr>
</tbody>
</table>

(Slides 12–24 are based on Manning & Schütze (M&S) 1999)
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 N N N
  - A N N N
  - A N N N
  - A N N N
  - A N N N
  - A N N N

- Crucially, all other sequences are removed

  - P D of the
  - MV V has been

(Pointwise) Mutual Information

Pointwise mutual information (PMI) compares:

- Observed: the actual probability of the two words appearing together ($p(w_1, w_2)$)
- Expected: the probability of the two words appearing together if they are independent ($p(w_1)p(w_2)$)

The pointwise mutual information is a measure to do this:

1. $I(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$
2. The higher the value, the more surprising it is

Mutual Information example

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

1. $C(\text{Ayatollah}) = 42$
2. $C(\text{Ruhollah}) = 20$
3. $C(\text{Ayatollah, Ruhollah}) = 20$
4. $N = 14,307,668$

4. $I(\text{Ayatollah, Ruhollah}) = \log_2 \frac{20}{\frac{42}{14,307,668}} = \log_{2} N \cdot \frac{20}{\frac{42}{14,307,668}} \approx 18.38$

To see how good a collocation this is, we need to compare it to others

POS filtering (2)

Some results after tag filtering (M&S 1999, table 5.3)

<table>
<thead>
<tr>
<th>C(w_1, w_2)</th>
<th>w_1</th>
<th>w_2</th>
<th>Tag Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>11487</td>
<td>New York</td>
<td>A N</td>
<td></td>
</tr>
<tr>
<td>7261</td>
<td>United States</td>
<td>A N</td>
<td></td>
</tr>
<tr>
<td>5412</td>
<td>Los Angeles</td>
<td>N N</td>
<td></td>
</tr>
<tr>
<td>3301</td>
<td>last year</td>
<td>A N</td>
<td></td>
</tr>
</tbody>
</table>

⇒ Fairly simple, but surprisingly effective

- Needs to be refined to handle verb-particle collocations
- Kind of inconvenient to write out patterns you want

Problem for Mutual Information

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 (M&S 1999, table 5.15)
Motivating Contingency Tables

What can we 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

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 = holmes</th>
<th>B ≠ holmes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = sherlock</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>A ≠ 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

- Values in this chart are the observed frequencies (fo)

Observed bigram probabilities

Each cell indicates a bigram: divide each cell by 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>

Marginal probabilities indicate probabilities for a given word
- e.g., \( p(\text{sherlock}) = 0.00099 \) and \( p(\text{holmes}) = 0.00647 \)

Expected bigram probabilities

Assuming sherlock & holmes are independent results in:

<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>

- Values in this chart are expected frequencies (fe)

Pearson’s chi-square test

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

\[
\begin{align*}
(5) \quad \chi^2 &= \sum \frac{(f_o - f_e)^2}{f_e} \\
(6) \quad \chi^2 &= \frac{(7 - 0.05)^2}{0.05} + \frac{(0 - 6.95)^2}{6.95} + \frac{(39 - 45.5)^2}{45.5} + \frac{(7059 - 7052.05)^2}{7052.05} \\
&= 966.05 + 6.95 + 1.048 + 0.006 \\
&= 974.05
\end{align*}
\]

Looking this up in a table shows it’s unlikely to be chance

- \( \chi^2 \) test does not work well for rare events, i.e., fe ≤ 5
- Other tests can be employed using the same tables
Gries (2009) lists some other points to consider:

- **Fertility**: # of unique types associate with a word
- **Lexical gravity**: window-based approaches that find the most informative contextual slots
- **Multi-word collocations**: breaking down the string into most informative units for expected frequencies
- **Variable n**: bottom-up approaches to defining the size of n for n-gram collocates
- **Discontinuous n-grams**