Lexicography

Corpora for lexicography

- Can extract authentic & typical examples, with frequency information
- With sociolinguistic meta-data, can get an accurate description of usage and, with monitor corpora, its change over time
- Can complement intuitions about meanings

The study of loanwords, for example, can be bolstered by corpus studies

Collocations & colligations

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

What a collocation is

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 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: ??*kicked the red bucket*

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*)

Firth 1957: “You shall know a word by the company it keeps”

- Collocational meaning is a *syntagmatic* type of meaning, not a conceptual one
- e.g., in this framework, one of the meanings of *night* is the fact that it co-occurs with *dark*

Example: *ass* is associated with a particular set of adjectives (think of *goose* if you prefer)

- *silly, obstinate, stupid, awful*
- We can see a *lexical set* associated with this word

Lexical sets & collocations vary across genres, subcorpora, etc.

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

Semantic prosody & preference

- *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’

Notes on a collocation’s definition

We often look for words which are adjacent to make up a collocation, but this is not always true

- e.g., *computers run*, but these 2 words may only be in the same proximity.

We can also speak of upward/downward collocations:

- *downward*: involves a more frequent node word A with a less frequent collocate B
- *upward*: weaker relationship, tending to be more of a grammatical property
Corpus linguistics
Krishnamurthy 2000

Where collocations fit into corpus linguistics:
1. Pattern recognition: recognize lexical and grammatical units
2. Frequency list generation: rank words
3. Concordancing: observe word behavior
4. Collocations: take concordancing a step further ...

Determining strength of collocation

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

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)

\[
C(w_1, w_2) = \frac{p(w_1, w_2)}{p(w_1)p(w_2)}
\]

(Slides 14–30 are based on Manning & Schütze (M&S) 1999)

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

(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 \( (p(w_1, w_2)) \)
The pointwise mutual information is a measure to do this:

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

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

(2) \( p(x) = \frac{C(x)}{N} \)
- \( N \) is the number of words in the corpus
- The number of bigrams \( \approx \) the number of unigrams

(3) \( I(w_1, w_2) = \log \frac{p(w_1,w_2)}{p(w_1)p(w_2)} = \log \frac{C(w_1,w_2)}{C(w_1)C(w_2)} \)

Problems for Mutual Information

The formula we have also has the following equivalencies:

(5) \( I(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} = \log \frac{P(w_1|w_2)}{P(w_1)} = \log \frac{P(w_2|w_1)}{P(w_2)} \)

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 (M&S 1999, table 5.15)

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>7105</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 \( f_{ij} \)

Mutual Information example

We want to know if Ayatollah Ruhollah is a collocation in a data set we have:
- \( C(\text{Ayatollah}) = 42 \)
- \( C(\text{Ruhollah}) = 20 \)
- \( C(\text{AyatollahRuhollah}) = 20 \)
- \( N = 14,307,668 \)

(4) \( I(\text{Ayatollah, Ruhollah}) = \log_2 \frac{42}{\frac{20 \times 20}{7105}} = \log_2 N \approx 18.38 \)

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

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

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 \( p(\text{holmes}) = 0.00647 \)
Corpus Linguistics
Application #2:
Collocations
Defining a collocation
Calculating collocations
Practical work

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 \( p_e(w_1, w_2) = p(w_1)p(w_2) \)

Pearson's chi-square test

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

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

\( (6) \)

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

\( = 966.05 + 6.95 + 1.048 + 0.006 \)

\( = 974.05 \)

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

NB: The \( \chi^2 \) test does not work well for rare events, i.e., \( f_e < 6 \)

Expected bigram probabilities

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 \( (f_e) \)

Working with collocations

The question is:

- What significant collocations are there that start with the word *sweet*?
- Specifically, what nouns tend to co-occur after *sweet*?

What do your intuitions say?

Next time, we will work on how to calculate collocations ...