Collocations

Seminário de Mestrado, 5 Dezembro

Michel Généreux

Some definitions

Multiword expression (MWE)

A multiword expression is a combination of two or more words whose semantic, syntactic, … properties cannot fully be predicted from those of its components, and which therefore has to be listed in a lexicon.

Collocation √

A sequence of words or terms that co-occur more often than would be expected by chance. Can be extracted statistically.

Salience √

How representative an expression is within a specialized corpus. Can be extracted statistically.

Examples

MWE

- kick the bucket (idiom)
- throw <somebody> to the lions

Collocations

- strong tea
- the rich and famous

Salient expression

- a quality player

Collocation vs MWE

MWE

- non-compositionality: semantically (semi-)opaque
- non-modifiability: syntactically rigid
- non-substitutability: lexically determined

Collocation

- the constraints applied to MWE can be relaxed. Statistical approach works fairly well to find collocations.

Compositionality

- A phrase is compositional if the meaning can predicted from the meaning of the parts.
- Collocations are not fully compositional in that there is usually an element of meaning added to the combination. Eg. *strong tea*.
- Idioms are the most extreme examples of non-compositionality. Eg. *to kick the bucket*

Non-Substitutability

- We cannot substitute near-synonyms for the components of a collocation. For example, we can't say *yellow wine* instead of *white wine* even though *yellow* is as good a description of the color of white wine as *white* is
- Many collocations cannot be freely modified with additional lexical material or through grammatical transformations (**Non-modifiability**).

Collocational Window

- Many collocations occur at variable distances. A collocational window needs to be defined to locate these. Freq based approach can't be used.
  - she knocked on his door
  - they knocked at the door
  - 100 women knocked on Donaldson's big door
  - a man knocked on the metal front door

Principal Approaches to Finding Collocations

- Dictionary
- Selection of collocations by **frequency**
- Selection based on **mean and variance** of the distance between focal word and collocating word
- **Mutual information**

Collocations in Dictionaries: *strength vs power*

strength	power
to build up ~	to assume ~
to find ~	emergency ~
to save ~	discretionary ~
to sap somebody's ~	~ over [several provinces]
brute ~	supernatural ~
tensile ~	to turn off the ~
the ~ to [do X]	the ~ to [do X]
[our staff was] at full ~	the balance of ~
on the ~ of [your recommendation]	fire ~

Frequency

- Finding collocations by counting the number of occurrences.
- Usually results in a lot of function word pairs that need to be filtered out.
- Pass the candidate phrases through a part of-speech filter which only lets through those patterns that are likely to be “phrases”.

Frequency

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

Most frequent bigrams in an Example Corpus

Except for *New York*, all the bigrams are pairs of function words.

Frequency

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Part of speech tag patterns for collocation filtering.

Frequency: filtering with POS patterns

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

The most highly ranked phrases after applying the filter on the same corpus as before.

strong *challenge* vs powerful *computer*

w	$C(strong, w)$	w	$C(powerful, w)$
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	9	magnet	4

CLUL

Mean and Variance

- The mean μ is the average offset between two words in the corpus.
- The variance σ^2

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

where n is the number of times the two words co-occur, d_i is the offset for co-occurrence i, and μ is the mean.

Offset = nb words between two words + 1

Mean and Variance: Interpretation

- The mean and variance characterize the distribution of distances between two words in a corpus.
- We can use this information to discover collocations by looking for pairs with low variance.
- A low variance means that the two words usually occur at about the same distance.
- A negative mean indicates that they permute

Mean and variance

σ	μ	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Mean and Variance: An Example

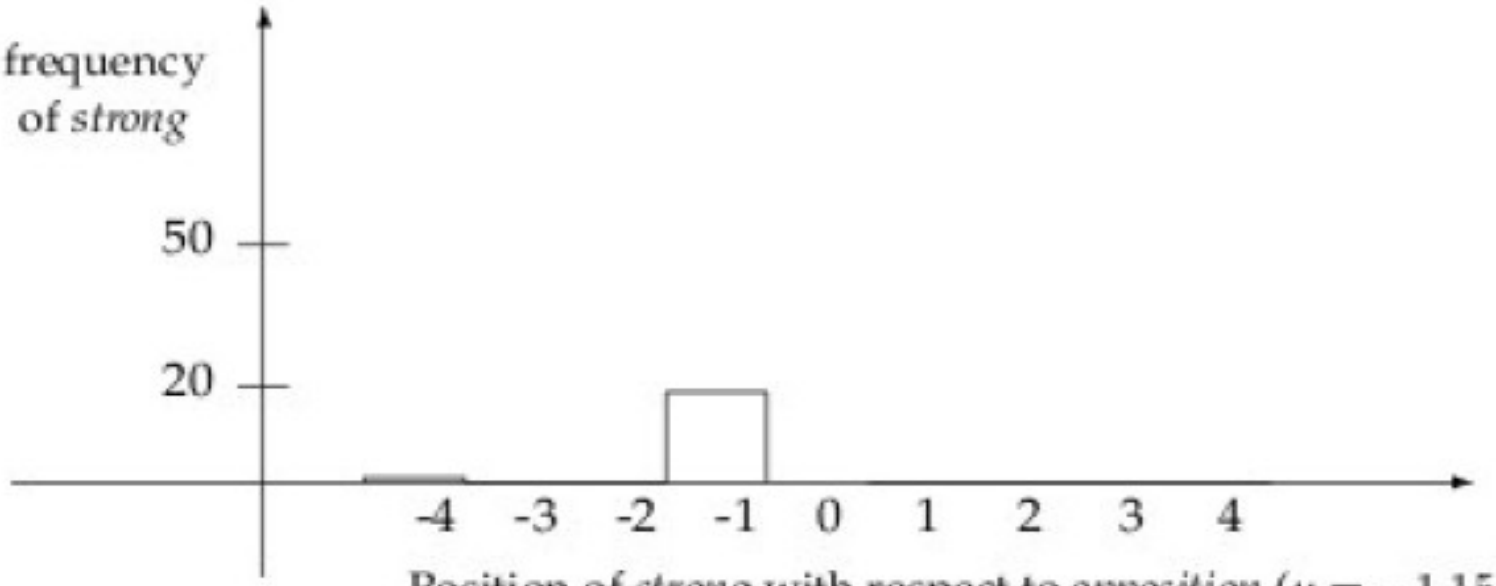
- For the *knock, door* example sentences the mean is:

$$\frac{1}{4}(3+3+5+5) = 4.0$$

- And the variance:

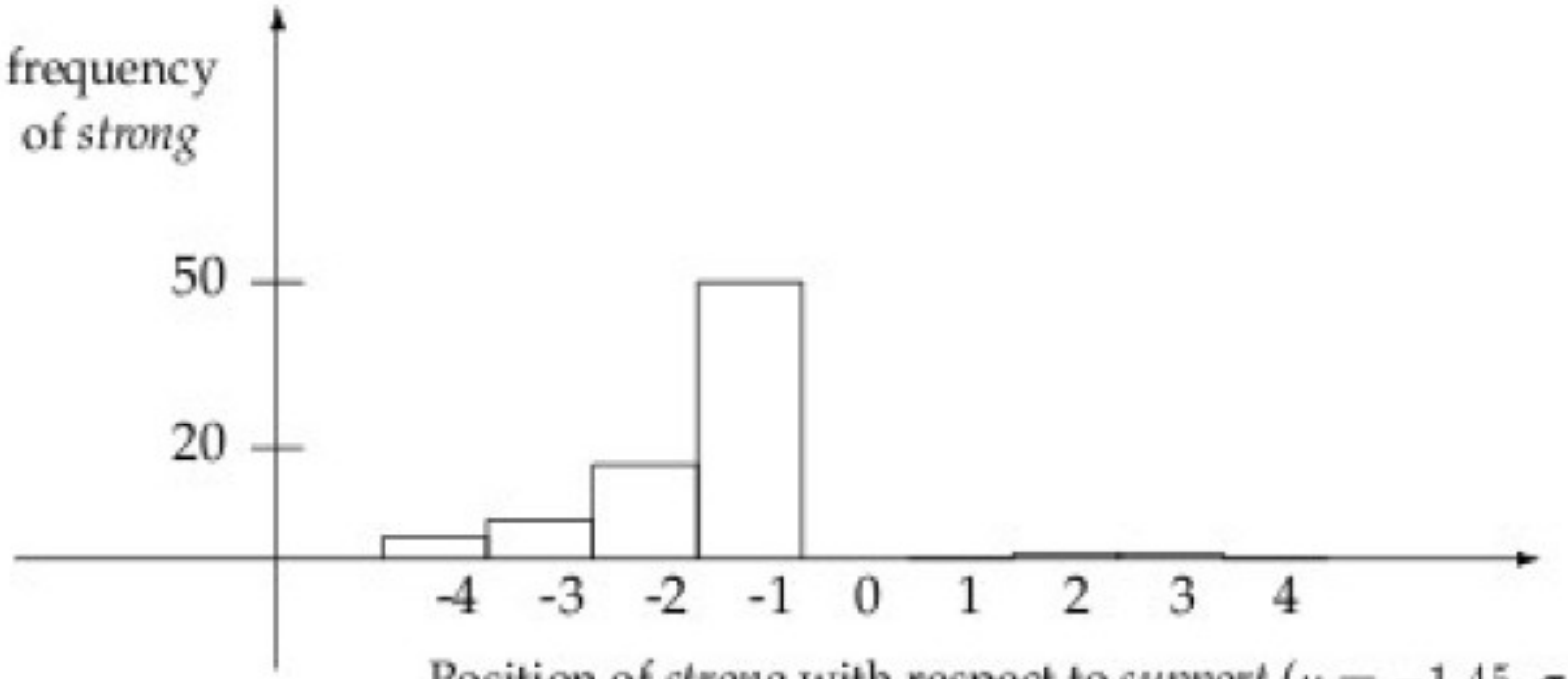
$$\sigma = \sqrt{\frac{1}{3}((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)} \approx 1.15$$

strong...opposition



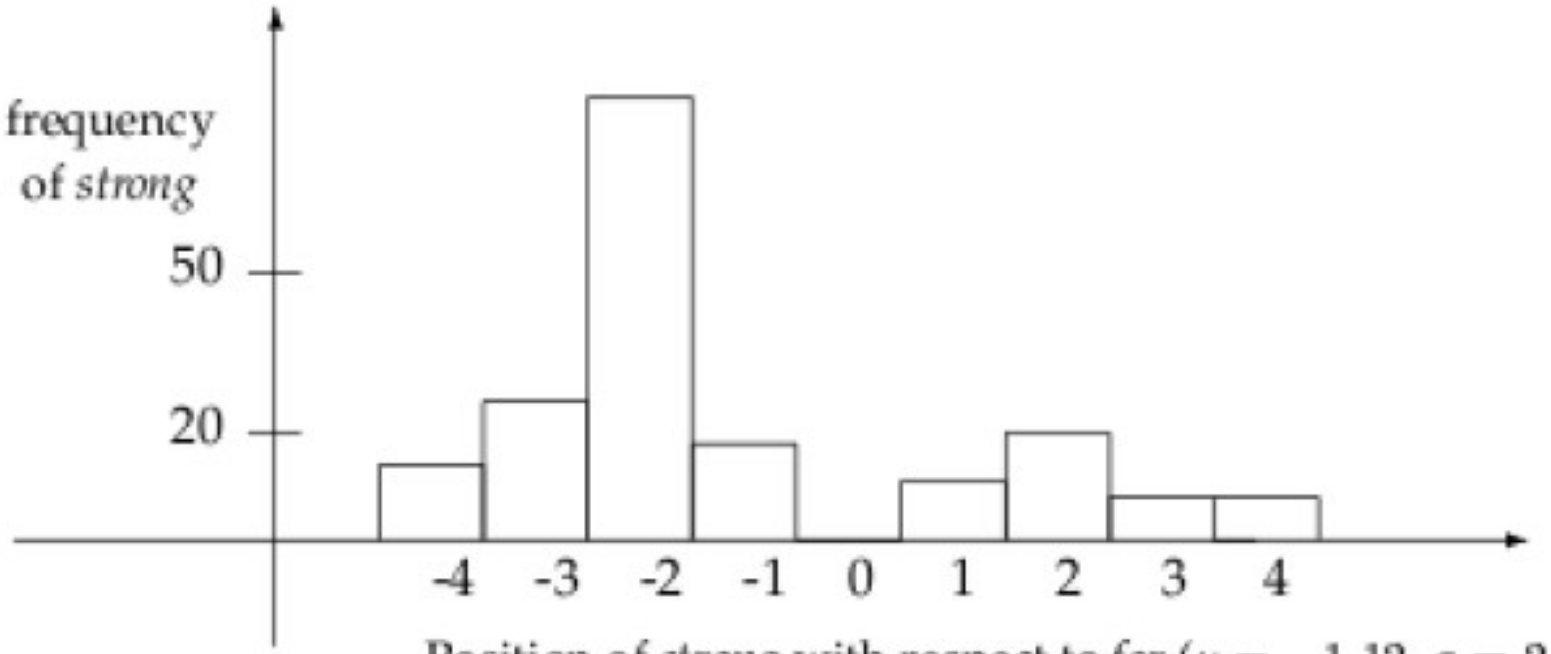
Position of *strong* with respect to *opposition* ($\mu = -1.15, \sigma = 0.67$).

strong...support



Position of *strong* with respect to *support* ($\mu = -1.45, \sigma = 1.07$).

strong and for



Position of *strong* with respect to *for* ($\mu = -1.12, \sigma = 2.15$).

Pointwise Mutual Information

- An information-theoretically motivated measure for discovering interesting collocations is *pointwise mutual information* (Church et al. 1989, 1991; Hindle 1990).
- It is roughly a measure of how much one word tells us about the other.
- It takes into account the fact that words can also co-occur by chance.
- Positive indicates co-occurrence
- Negative indicates that they tend not to co-occur

Pointwise Mutual Information (Cont.)

- Pointwise mutual information between particular events x' and y', in our case the occurrence of particular words, is defined as follows:

$$
\begin{aligned}
I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\
&= \log_2 \frac{P(x'|y')}{P(x')} \\
&= \log_2 \frac{P(y'|x')}{P(y')}
\end{aligned}
$$

PMI

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

$$I(Ayatollah, Ruhollah) = \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \approx 18.38$$

Corpus size = 14307668 words

CLUL

PMI

I_{1000}	w^1	w^2	w^1w^2	bigram	I_{23000}	w^1	w^2	w^1w^2	bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Problems with MI (see previous slide)

Some collocations move up as we have more data

- *marijuana growing*

Some non-collocations move down as we have more data

- *Reds survived*

But still many bigrams with an inflated MI

- *fewest visits*
- *Schwartz eschews*

(Experiment with CQPweb)

Salience

Comparing frequencies: term salience
The *Log-odds ratio*

- The log odds ratio measure compares the frequency of occurrence of each n-gram in a given specialized corpus with its frequency of occurrence in a reference corpus, where a is the frequency of a word in the specialized corpus, b is the size of the specialized corpus minus a, c is the frequency of the word in the general corpus and d is the size of the general corpus minus c. High positive log odds scores indicate strong salience, while high negative log odds scores indicate word irrelevant for the class.

$$r = ln(ad/cb) = ln(a) + ln(d) - ln(c) - ln(b)$$

Discriminative power

Tf-idf weighting

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

$$\text{idf}_t = \log \frac{N}{\text{df}_t}.$$

tf-idf assigns to term t a weight in document d in a corpus of N documents that is:

1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);

2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);

3. lowest when the term occurs in virtually all documents.

Reference

- Foundations of statistical natural language processing, chapter 5, *Collocations*, C. D. Manning and H. Schutze

Practical session

Download assignment_ut-hood.tar.gz:

$ wget http://alfclul.clul.ul.pt/crpc/curso/CL2012/files/collocations.tar.gz

Unzip it:

$ tar xvzf assignment_ut-hood.tar.gz

You will now have:

Five directories:

- /RC-en: a reference corpus in English
- /RC-pt: a reference corpus in Portuguese
- /Politics: a specialized corpus in English
- /Misterioso: a specialized corpus in Portuguese
- /lib: library of Perl programs

Three files: compute_salience.sh, idf.pl and mi.pl

A script and two programs

- compute_salience.sh: a script to compute the salience of expressions in a corpus

  $ ./compute_salience.sh -h

- idf.pl: a PERL program to compute the Inverse document frequency of terms

  $ ./idf.pl -h

- mi.pl: a PERL program to compute the Mutual Information value of collocations

  $ ./mi.pl -h

Salience on the English corpus

Compute the salience of n-grams from the specialized Politics corpus

$ ./compute_salience.sh en n Politics *will create*

en_n_Politics.model

- *the salience values (r in the logg-odds ratio formula)*

en_reference_tok_fqs_n_Politics

- n-gram token frequencies (c *in the logg-odds ratio formula*)

en_specialize_tok_fqs_n_Politics

- n-gram token frequencies (a *in the logg-odds ratio formula*)

en_reference_doc_fqs_n_Politics

- n-gram document frequencies (to compute IDF)

en_specialize_doc_fqs_n_Politics

- n-gram document frequencies (to compute IDF)

Loking at the most/least salient terms

The most salient n-gram:

$ head en_n_Politics.model

The least salient n-gram:

$ tail en_n_Politics.model

The most/least frequent terms

The most frequent n-gram:

```
$sort -k2gr en_specialized_tok_fqs_n_Politics | head
```

The least frequent n-gram:

```
$sort -k2gr en_specialized_tok_fqs_n_Politics | tail
```

Terms appearing in most/least documents

The most frequent n-gram

$sort -k2gr en_specialize_doc_fqs_n_Politics | head

The least frequent n-gram

$sort -k2gr en_specialize_doc_fqs_n_Politics | tail

Inverse Document Frequency

The most discriminating n-grams:

```
$ ./idf.pl en_specialize_doc_fqs_n_Politics | sort -k2gr | head
```

The least discriminating n-grams:

```
$ ./idf.pl en_specialize_doc_fqs_n_Politics | sort -k2gr | tail
```

Finding interesting collocations (2 tokens only)

The most interesting collocations for the specialized corpus:

```
$ ./mi.pl en_specialize_tok_fqs_1_Politics
  en_specialize_tok_fqs_2_Politics | sort -k2nr | head
```

The most interesting collocations for the reference corpus:

```
$ ./mi.pl en_reference_tok_fqs_1_Politics
  en_reference_tok_fqs_2_Politics | sort -k2nr | head
```