Statistics of collocations

Serge Sharoff

Centre for Translation Studies
University of Leeds

s.sharoff@leeds.ac.uk

Outline

Collocations

- 'Collocations of a given word are statements of the habitual or customary places of that word' (Firth, 1957)
  'you shall know a word by the company it keeps'
- collocations – constructions; collocates – words in collocations,
- Non-compositionality: *strong* vs. *powerful*: *tea* or *car*
  released from prison vs *discharged from hospital*
  full knee replacement vs *total knee replacement*

Examples of collocations

- Terms: *stiff breeze, weapons of mass destruction*
- Phrasal verbs: *get off, tell off, look up, make up*
- Support verb constuctions: *take a shower, make sense,*
- Stock phrases: *the rich and powerful, by and large*

Methods for counting

- N-grams: sequences of N words (bi-, trigram)
- *to be or not to be*
  unigrams → *to*, *be* (2), *or*, *not* (1)
  bigrams → *to be* (2); *be or* (1); *or not* (1); *not to* (1)
  trigrams → *to be or* (1); *be or not* (1); *or not to* (1)
- Skip-grams: pairs in a window of N words:
  W2: *to be* (2); *to or*; *be or*; *be not*

Counting bigrams

Bigrams		Trigrams	
of the	7211.67	i do not	522.24
in the	5167.19	there be a	401.55
it be	4050.64	it be a	372.39
to the	2617.17	one of the	356.03
be a	2366.99	it be not	348.88
do not	2230.41	there be no	292.65
on the	2181.97	be able to	241.46
have be	2151.05	do not know	232.90
there be	2017.23	the end of	213.57

last year	107.22
prime minister	97.18
last night	84.95
first time	83.27
other hand	56.12
last week	51.27
other people	42.01
next year	40.35
soviet union	38.95
young man	38.29

Outline

Statistics of surprise

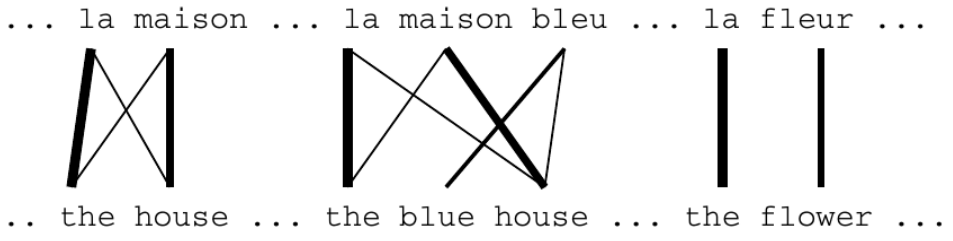
- Null hypothesis: words are distributed at random
- F_i — number of occurrences of word$_i$
- F_{ij} — number of joint occurrences of the two words (i and j)
- N — corpus size
- O_{ij} — observed probability, E_{ij} — expected probability,
- $O_{ij} = \frac{F_{ij}}{N}$ — observed probability,
- $E_{ij} = \frac{F_i}{N} \times \frac{F_j}{N}$ — expected probability,

	1	2	3	4	5	6
1	.	.	.	**x**	.	.
2	.	.	.	.	.	.
3	.	.	.	.	.	.
4	.	.	.	.	.	.
5	.	.	.	.	.	.
6	.	.	.	.	.	.

Notation for the probabilities

- The space of events (S)
  What is our event space?
- $p(x|\text{partial knowledge})$
- Conditional independence:
  Knowing about X doesn't tell me about Y
  $p(Y|X) = p(Y)$
  $p(X|Y) = p(X)$
- Conditional probability

$$p(X|Y) = \frac{p(X\&Y)}{p(Y)}$$



- $p(house|maison) = 0.476$
  $p(home|maison) = 0.104$
  $p(parent|maison) = 0.077$
  $p(flower|fleur) = 0.039$

Measures of collocations

- $O_{ij} = \frac{F_{ij}}{N}$ — observed probability,
- $E_{ij} = \frac{F_i}{N} \times \frac{F_j}{N}$ — expected probability,
- $MI_{ij} = \log\left(\frac{O_{ij}}{E_{ij}}\right)$ — Mutual Information score,
- $Dice_{ij} = 2 \times \frac{O_{ij}}{E_i + E_j}$ — Dice score,
- $T_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{O_{ij}}}$ — T-score
- Log-likelihood (LL) score from contingency table

	word2	¬ word2
word1	F_{ij}	$F_i - F_{ij}$
¬ word1	$F_j - F_{ij}$	$N - F_{ij}$

$loglike = 2(a \ln(\frac{F_i}{E1}) + b \ln(\frac{F_j}{E2}))$; $E1 = c\frac{a+b}{c+d}$; $E2 = d\frac{a+b}{c+d}$

Examples of predictions

- *new company*,
  $F_{ij} = 358, F_i = 105,645, F_j = 57,118, N = 100,000,000$
- *private company*,
  $F_{ij} = 423, F_i = 16,357, F_j = 57,118, N = 100,000,000$
- *post office*,
  $F_{ij} = 1,425, F_i = 10,871, F_j = 29,132, N = 100,000,000$

	MI score	Dice	T-score	LL-score
new company	6.19	2.82	15.97	761.32
private company	5.74	7.61	20.18	2,548.55
post office	8.59	9.44	25.11	6,354.51

Outline

Corpus analysis

- Multiword terminology
  Multiterm Extract
- One sense per collocation hypothesis
  take kindly
- Queries for collocations:
  strong N.*

  Right window of 3: *to offer* N.*
- Collocations for other languages
  den Vorteil *eines persönlichen Kontaktes über die Stimme* ***bietet****.*
  offer the advantage

Automation through word sketches

- Word sketches in `https://app.sketchengine.eu/`
- Fixed set of queries:
  Modifiers: ADV .. V.*
  Objects: V.* .. N.* or N.* *to be* VVN
- Sketches for other languages
  bieten

Basic points

- Collocations and collocates
- Statistics for measuring surprise
- Human judgment vs. computer model

For the seminar

Study collocation properties for words in your projects
Use their immediate left/right contexts and spans;
Try filtering collocates by their POS tags
Use word sketches

For further classes

Please either install Jupyter Lab with Python on your own laptop:
`https://jupyter.org/install`
OR ensure you have access to Google Drive and Google Colab:
`https://colab.research.google.com/`