Statistical methods of collocation detection

Aleksander Buczyński

2006.04.19

What is a collocation

Collocation detection

Frequency vs dependency

Filtering results

What is a collocation?

- a pair of words often occuring next to each other (*bigram*)?
- a longer string of words (*trigram*, *n-gram*), often occuring next to each other?
- a pair of words occuring close to each other (separated by no more than a few other words)?

What is a collocation? (2)

A pair of words occuring close to each other (separated by zero or more *functional words*) in a given collection of texts:

- **groundhog day**
- **seize** the **day**
- **picture** of the **day**
- **resolve conflicts**, **resolve** the **conflicts**

Types of collocations

1. **unconnected** - random co-occurence of words next to each other
2. **functional** - defining the application domain of words
   - *novel character*, *resolve conflicts*, *miasto stołeczne (capital city)*, *stolica Polski (capital of Poland)*
3. **idiomatic** - carrying meaning that cannot be deduced from the meaning of its components
   - *compact disc*, *couch potato*, *carrot and stick*, *crocodile tears*

Statistical methods of collocation detection

General idea: assign scores to all bigrams in a corpus. The higher the score, the higher the chance that the pair of words form a collocation (or: the stronger the collocation).

Frequency score

Frequency - number of occurences of a pair of words $w_1 w_2$ in a corpus:

$$R_{Freq} = c(w_1 w_2)$$

Most of bigrams with highest frequency are unconnected collocations of frequent words:

- English: *you can*, *if you*, *it is*, *when you*, *to use*, *this is*, *does not*, *that you*, *you have*, *you are*...
- Polish: *się na*, *się do*, *nie ma*, *nie jest*, *jest to*, *nie tylko*, *że nie*, *nie są*...

Conclusion: we should also take into account how often words w_1 and w_2 occur seperately.

Symmetric Conditional Probability

$$R_{SCP} = \frac{c(w_1 w_2)}{c(w_1)} \frac{c(w_1 w_2)}{c(w_2)} = \frac{c(w_1 w_2)^2}{c(w_1)c(w_2)}$$

Where:
$c(w_1 w_2)$ - number of occurences of a pair of words $w_1 w_2$ in a corpus
$c(w)$ - number of occurences of a word w in a corpus
Values between 0 and 1:

- **0** - words never occur next to each other
- **1** - both words occur only next to each other

Frequency vs dependency

Frequency tests:

- Frequency
- Student's t-score
- Log Likelihood Ratio
- Mutual Information

Dependency tests:

- Maximum Mutual Information Ratio
- z-score
- Dice Formula
- Symmetric Conditional Probability

"Broken" test:

- Pointwise Mutual Information

Frequency-like tests

Tabela: Top 20 collocations from Jane Austen books according to different *frequency-like* measures

Freq	Student's	LLR	Lef	MI
to be	to be	to be	to be	to be
it was	it was	had been	had been	had been
she had	she had	have been	have been	have been
she was	had been	it was	it was	it was
had been	it is	could not	could not	could not
it is	she was	she had	she had	she had
to her	could not	it is	it is	it is
could not	have been	my dear	my dear	my dear
have been	he had	did not	did not	did not
he had	do not	do not	am sure	am sure
he was	did not	am sure	do not	do not
do not	she could	they were	they were	they were
she could	he was	she was	sir thomas	sir thomas
did not	would be	he had	she was	she was
that she	they were	sir thomas	he had	he had
would be	must be	would be	would be	would be
that he	my dear	she could	she could	she could
was not	that he	must be	must be	must be
to have	will be	more than	more than	more than
they were	that she	you are	her own	her own

Dependency-like tests

Tabela: Top 20 collocations from Jane Austen books according to different *dependency-like* measures

MxI	Z22	Z-score	Dice	SCP
thornton lacey	thornton lacey	thornton lacey	thornton lacey	thornton lacey
maple grove	maple grove	maple grove	maple grove	maple grove
combe magna	combe magna	combe magna	combe magna	combe magna
de courcy	lovers' vows	lovers' vows	lovers' vows	lovers' vows
brunswick square	de courcy	de courcy	de courcy	de courcy
lovers' vows	brunswick square	brunswick square	frank churchill	brunswick square
de bourgh	captain wentworth	captain wentworth	brunswick square	captain wentworth
sir thomas	frank churchill	frank churchill	captain wentworth	frank churchill
captain wentworth	de bourgh	de bourgh	thousand pounds	de bourgh
colonel brandon	thousand pounds	thousand pounds	de bourgh	thousand pounds
harley street	colonel brandon	colonel brandon	sore throat	colonel brandon
milsom street	sir thomas	sir thomas	colonel brandon	sir thomas
pulteney street	sore throat	sore throat	tete (a) tete	sore throat
charles hayter	tete (a) tete	tete (a) tete	sir thomas	tete (a) tete
abbey mill	dare say	dare say	robert martin	dare say
william larkins	great deal	great deal	am sure	great deal
wimpole street	charles hayter	charles hayter	kellynch hall	charles hayter
lady russell	abbey mill	abbey mill	box hill	abbey mill
berkeley street	captain benwick	captain benwick	great deal	captain benwick
colonel brandon's	am sure	am sure	dare say	am sure

Frequency vs dependency

Tabela: Comparison between frequency and dependency rankings

Freq	LLR	MI	MxI	Z-score	SCP
to be	to be	to be	thornton lacey	thornton lacey	thornton lacey
it was	had been	had been	maple grove	maple grove	maple grove
she had	have been	have been	combe magna	combe magna	combe magna
she was	it was	it was	de courcy	lovers' vows	lovers' vows
had been	could not	could not	brunswick square	de courcy	de courcy
it is	she had	she had	lovers' vows	brunswick square	brunswick square
to her	it is	it is	de bourgh	captain wentworth	captain wentworth
could not	my dear	my dear	sir thomas	frank churchill	frank churchill
have been	did not	did not	captain wentworth	de bourgh	de bourgh
he had	do not	am sure	colonel brandon	thousand pounds	thousand pounds
he was	am sure	do not	harley street	colonel brandon	colonel brandon
do not	they were	they were	milsom street	sir thomas	sir thomas
she could	she was	sir thomas	pulteney street	sore throat	sore throat
did not	he had	she was	charles hayter	tete (a) tete	tete (a) tete
that she	sir thomas	he had	abbey mill	dare say	dare say
would be	would be	would be	william larkins	great deal	great deal
that he	she could	she could	wimpole street	charles hayter	charles hayter
was not	must be	must be	lady russell	abbey mill	abbey mill
to have	more than	more than	berkeley street	captain benwick	captain benwick
they were	you are	her own	colonel brandon's	am sure	am sure

Dependency vs number of occurences

Tabela: Sample SCP ranking for the Polish Language Council website

$w_1 w_2$	$c(w_1)$	$c(w_2)$	$c(w_1 w_2)$	R_{SCP}
stulecie obfitowało	1	1	1	1,000
ekspansywne (a) niepotrzebne	1	1	1	1,000
kodyfikacja zdecentralizowana	1	1	1	1,000
międzywyrazowa fonetyka	1	1	1	1,000
uwięźnie chudzina	1	1	1	1,000
bagienice folwark	1	1	1	1,000
przeanalizowanie stopniowalności	1	1	1	1,000
odzwierciedla apelatywizację	1	1	1	1,000
metajęzykowa (i) performatywna	1	1	1	1,000
trąby mosiężne	1	1	1	1,000
hipotetyczne dwubiegunowe	1	1	1	1,000
inicjalny trzysylabowych	1	1	1	1,000
sobotni zdominowany	1	1	1	1,000
...				
public relations	6	6	6	1,000
...				
punkt widzenia	35	34	26	0,568
...				

R_{SCP} (and other dependency tests) favours rare collocations of rare words.

Typical solution: discarding collocations occuring in a corpus less than n times.

Frequency biased Symmetric Conditional Probability

Alternative approach: instead of binary filter, modify the test. Common sense suggests that among collocations with similar dependency values, those more frequent should be more important.

$$R_{FSCP} = \frac{c(w_1 w_2)^{2+\alpha}}{c(w_1)c(w_2)}$$

$\alpha = 1$ is a bit too strongly biased, $\alpha = 0.5$ seems better.

FSCP - sample results for Polish law

Kodeks cywilny:
chyba że, stosować się, współżycie społeczne, móc żądać, należyta staranność, zakład ubezpieczeń, rażące niedbalstwo, przedsiębiorca składowy, depozyt sądowy, naprawienie szkody, dawać zlecenie, zarobkowo hotel, w razie, druga strona, samorząd terytorialny, skarb państwa, sześć miesięcy, obowiązanym być, jak również...
Kodeks postępowania administracyjnego:
administracja publiczna, organ administracji [publicznej], samorząd terytorialny, wyższy stopień, chyba że, jednostka samorządu, ze względu, interes społeczny, pierwsza instancja, stanowić inaczej, stosować się, siedem dni, od dnia, podstawa prawna, służy zażalenie, przysposobienia opieki, stan prawny, jednostka organizacyjna, z urzędu, niniejszy dział...

Proper names

Proper names are:

- usually strong collocations
- sometimes not the most interesting collocations

Simple filter: a pair of words is considered a proper name if in all its instances in a corpus both words start with upper-case letters.

Proper names filter

Top twenty collocations	Without proper names	Only proper names
piano forte	piano forte	**thornton lacey**
thornton lacey	**de bourgh**	**maple grove**
maple grove	thousand pounds	**combe magna**
combe magna	barouche landau	**lovers' vows**
lovers' vows	sore throat	**vale uske**
vale uske	court plaister	**de courcy**
de courcy	tete (a) tete	**brunswick square**
brunswick square	baked apples	**captain wentworth**
captain wentworth	dare say	**frank churchill**
frank churchill	great deal	**count cassel**
count cassel	am sure	**colonel brandon**
de bourgh	ring (the) bell	**sir thomas**
thousand pounds	my dear	**upper seymour**
colonel brandon	drawing room	**edgar's buildings**
sir thomas	burst forth	**west indies**
upper seymour	young man	**westgate buildings**
barouche landau	ha ha	**blaize castle**
sore throat	depend upon	**charles hayter**
court plaister	**lesley castle**	**abbey mill**
tete (a) tete	had been	**captain benwick**