Edge Hill University

</TECHUP_WOMEN>

Natural Language Processing

Part 2

Collocation Extraction

Durham University

Edge Hill University

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

University of York

institute of CODING

Office for Students

Overview

- Completed introduction to NLP
- Completed Activity 1

- Collocation Extraction
  - ❑ Collocations
    - Definition
    - Characteristics
    - Applications
  - ❑ Automatically detecting collocations
    - Frequency
    - Pointwise Mutual Information (PMI)

What is a collocation?

Manning and Schütze (1999):
"A collocation is a phrase consisting of two or more words that corresponds to some conventional way of saying things"

New York

Strong tea

make up

Kick the bucket

new company

Not a fixed phrase

Choueka (1988):
"A collocation is a phrase that:

i. *has characteristics of syntactic and semantic unit*
ii. *whose meaning cannot be derived from the meaning of its components"*

Collocations: Limited Compositionality

The meaning of a collocation is **not** derived by the meaning of its constituent words

kick the bucket : *pass away*

spill the beans : *reveal a secret*

Collocations: Limited Substitutability

The constituent words of collocations cannot be replaced by other synonymous words

strong tea	**Powerful** tea
Kick the **bucket**	Kick the **can**
real estate	**actual** estate
weapons of **mass** destruction	weapons of **bulk** destruction

Collocations: Limited Modifiability

Collocations cannot be augmented with additional words

kick the bucket

kick the ***wooden*** bucket

Spill the beans

spill the ***red*** bins

Properties of collocations

Limited compositionality

The meaning of a collocation is **not** derived by the meaning of Its constituent words

- kick the bucket : *pass away*
- spill the beans : *reveal a secret*

Limited substitutability

The constituent words of collocations cannot be replaced by other synonymous words

- Kick the bucket != punch the bucket
- spill the beans != spill the lentils

Limited modifiability

cannot be augmented with additional words

- Kick the *wooden* bucket
- spill the red beans

What are collocations used for?

- **Automatic Summarisation:** automatically produced summary sounds natural
- **Machine Translation:** Correct erroneous translations (e.g. strong tea vs powerful tea)
- **Computational Lexicography:** Identify new collocations to be included in a dictionary
- **Internet Search:** Index documents using important collocations

NLP methods for collocation extraction

- Collocations exhibit some statistical properties

- For extracting collocations we will use the following statistical methods:
  1. **Co-occurrence frequency**
  2. **Mutual Information**

Co-occurrence Frequency

"Words that occur frequently together should have a special function"

Frequency of word 1 (unigram)	Frequency of words 2 (unigram)	**Co-occurrence frequency of word 1 and word 2 (bigram)**
Frequency (New) = 20,000	Frequency (York) = 15,000	Frequency (**New York**) = 11,428
Frequency (Old) = 22,000	Frequency (York) = 15,000	Frequency (Old York) = 141

Co-occurrence frequency of 'New' and 'York' is very high, so 'New York' should be a collocation

Co-occurrence frequency of 'Old' and 'York' is very low, so 'Old York' is NOT a collocation

Co-occurrence Frequency Results on a large corpus

of	the	80,871
in	the	58,841
to	the	26,430
on	the	21,842
for	the	21,839
and	the	18,568
that	the	16,121
at	the	15,630
to	be	15,494
in	a	13,899

Results are not interesting

of	a	13,689
by	the	13,361
with	the	13,183
from	the	12,622
New	York	11,428
he	said	10,007
as	a	9,775
is	a	9,231
has	been	8,753
for	a	8,573

20 most frequent bigrams in an example newswire corpus and their frequencies

Improving Co-occurrence Frequency Results

Part-of-speech filtering

- Keep only those unigrams that follow a pre-determined POS pattern
  - "first_word:Noun, second_word:Noun"
  - "first_word:Adjective, second_word:Noun"

~~of /Prep~~	~~the /Det~~	80,871
~~in /Prep~~	~~the /Det~~	58,841
~~to /To~~	~~the /Det~~	26,430
~~on /Prep~~	~~the /Det~~	21,842
~~for /Prep~~	~~the /Det~~	21,839
~~and /Conjuc~~	~~the /Det~~	18,568
~~that /Prep~~	~~the /Det~~	16,121
~~at /Prep~~	~~the /Det~~	15,630
~~to /TO~~	~~be /VB~~	15,494
~~in /Prep~~	~~a /Det~~	13,899

~~of /Prep~~	~~a /Det~~	13,689
~~by /Prep~~	~~the /Det~~	13,361
~~with /Prep~~	~~the /Det~~	13,183
~~from /Prep~~	~~the /Det~~	12,622
New /Adj	York /Noun	11,428
~~he /Pronoun~~	~~said /VB~~	10,007
~~as /Prep~~	~~a /Det~~	9,775
~~is /VB~~	~~a /Det~~	9,231
~~has /VB~~	~~been /VB~~	8,753
~~for /Prep~~	~~a /Det~~	8,573

Co-occurrence Frequency Using POS filter

New	York	11,428
United	States	7,261
last	year	3,301
Saudi	Arabia	3,191
vice	president	2,514
Persian	Gulf	2,387
San	Francisco	2,161
Middle	East	2,001
President	Bush	1,942
Soviet	Union	1.867

Results are significantly better using the POS filter

oil	prices	1,328
next	year	1,210
chief	executive	1,074
from	the	12,622
New	York	11.428
he	said	10,007
as	a	9,775
is	a	9,231
has	been	8,753
for	a	8,573

20 most frequent bigrams after applying a POS filter

Limitation of Co-occurrence frequency

…	…	…
Agatha	Christie	20
Bette	Midler	20
videocassette	recorder	20
unsalted	butter	20
…	…	…

Even if we apply filtering Collocations might be ranked lower in the list

Pointwise Mutual Information (PMI)?

- Statistical metric motivated by information theory
- Given x,y events:

$$pmi(x;y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

Joined probability: Probability that x, y occur together

Probability of x

Probability of y

P(x)P(y): probability that x and y occur independently

- Intuition: *How likely is that two events will appear together and not separately*
  - Collocations: *How likely is that two words will appear together and not separately*

How do we calculate Mutual Information?

$$pmi(x;y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

- P(x) = P(word_1) = $\frac{\text{frequency of word 1}}{\text{Number of words in the document}}$
- P(y) = P(word_2) = $\frac{\text{frequency of word 2}}{\text{Number of words in the document}}$
- P(x,y) = P(word_1, word_2) = $\frac{\text{frequency of word 1 co−occuring with word 2}}{\text{Number of words in the document}}$

How do we calculate Mutual Information?

Calculate the PMI of *computer science*

Words : w_1= computer, w_2= science

Frequencies: c_1= 42, $c_2 = 20$, c_{12}=20

Corpus size (# of uni-grams): N = 14307668

Answer

$$\text{pmi(computer;science)} = log_2 \frac{P(computer,science)}{P(computer)P(science)} =$$

$$= log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668}\frac{20}{14307668}} \approx 18.38$$

Mutual Information Activity

- Calculate the PMI of the words “New York” given in the following sentence

The City of New York, often called New York City or simply New York, is the most populous city in the United States

- $P(New) = \frac{\text{frequency of word "New"}}{\text{Number of words in the document}} = \_\_\_\_\_$
- $P(York) = \frac{\text{frequency of word "York"}}{\text{Number of words in the document}} = \_\_\_\_\_$
- $P(New, York) = \frac{\text{frequency of "New" co–occuring with "York"}}{\text{Number of words in the document}} = \_\_\_\_\_$
- $PMI(New,York) = \frac{P(New,York)}{P(New)P(York)} = \_\_\_\_\_$

Mutual Information Solution

The City of New York, often called New York City or simply New York, is the most populous city in the United States

- P(New) = $\frac{\text{frequency of word "New"}}{\text{Number of words in the document}} = \frac{3}{23}$
- P(York) = $\frac{\text{frequency of word "York"}}{\text{Number of words in the document}} = \frac{3}{23}$
- P(New, York) = $\frac{\text{frequency of "New" co−occuring with "York"}}{\text{Number of words in the document}} = \frac{3}{23}$
- PMI(New,York) = $\frac{P(New,York)}{P(New)P(York)} = \frac{\frac{3}{23}}{\frac{3}{23}*\frac{3}{23}} = \frac{23}{3}$

Mutual Information

w1	w2	c1	c2	c12	pmi(w1, w2)
Ayatollah	Ruhollah	42	20	20	18.38
Agatha	Christie	30	117	20	16.31
videocassette	recorder	77	59	20	15.94
Unsalted	butter	24	320	20	15.19

PMI can identify low frequency collocations

Collocation Extraction - Summary

- Methods for detecting collocations

	Advantages	Disadvantages
Co-occurrence Frequency	• Easy to implement • Performs well using POS filter	• High co-occurrence frequency does not always determine collocations (e.g., new company) • Cannot identify collocations that occur with a low frequency
PMI	• Takes into consideration degree of correlation between two words • Can detect collocations that present low co-occurrence frequency	• It overestimates rare phrases.

Activity

- Completed topic 5.3, and introduction to NLP
- Activities to complete to aid understanding
  - Solutions provided