T-61.5020 Statistical Natural Language Processing

Answers 6 — Collocations
Version 1.0

1. Let's start by calculating the results for pair "valkoinen","talo" manually:

   – Frequency: Bigrams "valkoinen","talo" occurred 710 times.
   – Normalized frequence: Word "valkoinen" occurred 3665 times and "talo" 10 767 times. We get $\frac{710}{3665 \cdot 10767} \approx 1.8 \cdot 10^{-5}$.

   All the results for the frequency method are in Table 1 and for the normalized method in Table 2.

   We see that results are quite good even for as simple method as this.

2. For the collocation "valkoinen", "talo":

$$
\begin{aligned}
Mean(\text{“valkoinen”}, \text{”talo”}) &= \frac{-1 \cdot 710 - 2 \cdot 2 + 1 + 2 \cdot 6}{710 + 2 + 1 + 6} \\
&\approx -0.975
\end{aligned}
$$

$$
\begin{aligned}
&Var(\text{“valkoinen”}, \text{”talo”}) \\
&= \frac{(-1-(-0.975))^2 \cdot 710 + (-2-(-0.975))^2 \cdot 2 + (1-(-0.975))^2 \cdot 1 + (2-(-0.975))^2 \cdot 6}{710 + 2 + 1 + 6} \\
&\approx \quad 0.083
\end{aligned}
$$

   Rest of the results, sorted by the variance, are in Table 3.

   This method has found in practice all the fixed collocations. However, results are not so good with sparse data: "vihainen mielenosoittaja" is definitely not a collocation.

   Size of the window surely affects the results. If it is too large, pairs start to occur together randomly too often, if too small, the collocations with longer effect are not found. If the second word of the collocation can be either before or after the first one, the method will clearly not work at all.

Table 1: *Results for the frequency method*

s_1	s_2	$C(s_1, s_2)$
ja	olla	7329
venäjä	presidentti	717
valkoinen	talo	710
kova	tuuli	279
aste	pakkanen	160
tuntematon	sotilas	154
sekä	myös	138
liukas	keli	106
hakea	työ	31
oppia	lukea	21
ottaa	onki	9
vihainen	mielenosoittaja	7
olla	ula	5
heittää	veivi	3
herne	nenä	3

Table 2: *Results for the normalized frequency method*

s_1	s_2	Normalized frequency$\cdot 10^{-8}$
liukas	keli	1981
aste	pakkanen	386
heittää	veivi	293
herne	nenä	268
valkoinen	talo	180
tuntematon	sotilas	163
vihainen	mielenosoittaja	68
kova	tuuli	35
ottaa	onki	21
venäjä	presidentti	10
oppia	lukea	8
hakea	työ	1
olla	ula	0
sekä	myös	0
ja	olla	0

Table 3: *Results sorted by the smallest variance*

s_1	s_2	Mean	Variance
herne	nenä	-1.000	0.000
vihainen	mielenosoittaja	-1.000	0.000
tuntematon	sotilas	-1.025	0.025
valkoinen	talo	-0.975	0.083
ottaa	onki	-1.250	0.188
venäjä	presidentti	-1.128	0.472
kova	tuuli	-0.880	0.492
liukas	keli	-0.788	0.608
oppia	lukea	-0.606	1.087
heittää	veivi	-0.500	1.250
aste	pakkanen	-0.465	1.347
hakea	työ	-0.433	2.046
olla	ula	-0.250	2.438
sekä	myös	0.252	2.981
ja	olla	-0.083	3.635

3. In statistical test, one should start by forming a null hypothesis. In our case, we assume that the words in a pair are independent: $P(s_1, s_2) = P(s_1)P(s_2)$. The tests will give a level of confidence for the null hypothesis. The significance level, below which the null hypothesis is discarded, is usually at most 0.05.

   In t-test we assume that the probabilities are normally distributed, and check if the expectation value for the observed data differs from the expectation value given by the null hypothesis. The t-values are given by

$$t = \frac{\hat{x} - \mu}{\sqrt{\frac{s^2}{N}}},$$

   where $\hat{x}$ is the sample mean, s^2 is the sample variance, N is the number of samples and μ is the mean of the distribution. In our case

$$\begin{aligned} \mu &= P(s_1)P(s_2) = \frac{C(s_1)}{N}\frac{C(s_2)}{N} \\ \hat{x} &= \frac{C(s_1, s_2)}{N} = \hat{p} \\ s^2 &= p(1-p) = \hat{p}(1-\hat{p}) \approx \hat{p} \end{aligned}$$

   For the pair “valkoinen talo” we get

$$t = \frac{\frac{710}{28181344} - \frac{2665 \cdot 10767}{28181344^2}}{\sqrt{\frac{710}{28181344^2}}} \approx 27.$$

If the t-value is over 6.314, the probability that the sample was from the distribution given by the independence assumption is less than 5%. Consequently, we can mark "valkoinen talo" as a collocation. Table 4 has values for all of the candidates. Note that the last pairs get negative values. This is because they occur together more rarely than the null hypothesis gives.

Table 4: *Results for the t-test*

s_1	s_2	t
valkoinen	talo	27
venäjä	presidentti	26
kova	tuuli	17
aste	pakkanen	13
tuntematon	sotilas	12
liukas	keli	10
oppia	lukea	4
hakea	työ	4
ottaa	onki	3
vihainen	mielenosoittaja	3
heittää	veivi	2
herne	nenä	2
olla	ula	0
sekä	myös	-9
ja	olla	-385

χ^2-test is based on a simple assumption: We look at the separate probabilities and estimate how many times the words should occur together. This is compared to the observed co-occurrence value, and if they differ too much, the pair is likely to be a collocation.

Let's start by collecting the following table (table 5): These values can be used in the two-variable χ^2-test:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

By assigning the numbers:

$$\begin{aligned}\chi^2 &= \frac{28181344(710 \cdot 28167622 - 10057 \cdot 2955)^2}{(710 + 10057)(710 + 2955)(10056 + 28167622)(2955 + 28167622)} \\ &\approx 358771\end{aligned}$$

If the result for χ^2-test is over 3.843, the sample is drawn from an independent distribution with less than 5% probability. "Valkoinen talo" seems to be a collocation.

Table 5: *Quantities needed in the χ^2-test.*

	w_1=valkoinen	$w_1 \neq$ valkoinen
w_2 =talo	710 (valkoinen talo)	$10767 - 710 = 10057$ (punainen talo)
$w_2 \neq$talo	$3665 - 710 = 2955$ (valkoinen mopo)	$28181344 - 710 - 10057 - 2955 = 28167622$ (punainen mopo)

However, if we look at Table 6, we notice that almost every pair would be a collocation according to this. The reason is that the χ^2-test does not test for the pairs to be collocations, but whether they are independent. For example, the pair "ja", "olla" has a high correlation, but actually it is a negative one: They occur together less than they should according to the independence assumption.

4. Mutual information tells how much more information X gives for determining Y. If X and Y are independent, the mutual information is zero. For "valkoinen","talo",

$$
\begin{aligned}
I(x,y) &= \log_2 \frac{P(X,Y)}{P(X)P(Y)} \\
&= \log_2 \frac{\frac{710}{28181344}}{\frac{3665}{28181344}\frac{10767}{28181344}} \\
&\approx 9.0
\end{aligned}
$$

The rest of the results are in Table 7.

The results seem to be good. The course book criticises that this method favours the less frequent words. Reason for this is the way that is used to estimate the probabilities, i.e. maximum likelihood estimation. Better result can be obtained if we instead set a prior for the words to be independent, and let the data modify it.

As a conclusion, we could say the following. The heuristic methods (ex. 1 and 2) are easy to apply and still give fair results. The statistical models applied in exercises 3 and 4 are justifiable as such, but they measure the correlation of the words, not whether they are collocations. However, if this is remembered, the results can be good. The statistical tests (ex. 3) may be harder to piece together, and the assumtions behind the methods must be kept in mind. In methods based directly on probability calculations (ex. 4), these assumptions are usually brought out more explicitly. In all methods based on probability estimates from data, one must choose how to approximate the probabilities. Maximum likelihood estimates can be too susceptible to sparse data. A better choice could be Maximum A Posterior (MAP) estimation with a prior belief that the words are independent.

Table 6: *Results for the χ^2-test*

s_1	s_2	χ^2
liukas	keli	591591
valkoinen	talo	358771
aste	pakkanen	173726
tuntematon	sotilas	70409
ja	olla	29194
kova	tuuli	26644
venäjä	presidentti	18147
heittää	veivi	4120
herne	nenä	2258
vihainen	mielenosoittaja	1321
ottaa	onki	525
oppia	lukea	449
hakea	työ	47
sekä	myös	45
olla	ula	0

Table 7: *Results for the mutual information method*

s_1	s_2	MI
liukas	keli	12.4
aste	pakkanen	10.1
heittää	veivi	9.7
herne	nenä	9.6
valkoinen	talo	9.0
tuntematon	sotilas	8.8
vihainen	mielenosoittaja	7.6
kova	tuuli	6.6
ottaa	onki	5.9
venäjä	presidentti	4.8
oppia	lukea	4.5
hakea	työ	1.7
olla	ula	0.5
sekä	myös	-0.8
ja	olla	-2.5