UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Statistical Evaluation II:
Dealing with Context Windows

Let's start from the beginning

$$LL = \log L(c_{12}, c_1, p) + \log L(c_2-c_{12}, N-c_1, p)$$
$$- \log L(c_{12}, c_1, p_1) - \log L(c_2-c_{12}, N-c_1, p_2)$$

c_1 = occurrences of word 1 in the text

c_2 = occurrences of word 2 in the text

c_{12} = co-occurrences of word 1 with word 2 in the text

N = number of tokens in the text

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2-c_{12})/(N-c_1)$

Let's start from the beginning

$$\begin{aligned} LL = {} & \log \mathit{L}(c_{12}, c_1, p) + \log \mathit{L}(c_2\text{-}c_{12}, N\text{-}c_1, p) \\ & - \log \mathit{L}(c_{12}, c_1, p_1) - \log \mathit{L}(c_2\text{-}c_{12}, N\text{-}c_1, p_2) \end{aligned}$$

c_1 = occurrences of word 1 in the **text**

c_2 = occurrences of word 2 in the **text**

c_{12} = co-occurrences of word 1 with word 2 in the **text**

N = number of tokens in the **text**

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2\text{-}c_{12})/(N\text{-}c_1)$

Let's start from the beginning

$$\begin{aligned} LL = \; & \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\ & - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2) \end{aligned}$$

c_1 = occurrences of word 1 in the **data**

c_2 = occurrences of word 2 in the **data**

c_{12} = co-occurrences of word 1 with word 2 in the **data**

N = number of tokens in the **data**

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2 - c_{12})/(N - c_1)$

Data, not Text!

- We have abstracted data from the text
- We should no longer refer to the text
- But, instead, to the data
- The DataFrames we have constructed have everything we need

Counts for the target word (word 1)

$$f(t) = \frac{1}{W} \sum_{c} n(c,t)$$

t = the target word (word 1)

c = the co-occurrent (word 2)

W = the size of the window

This equation from Bullinaria and Levy, “Extracting Semantic Representations from Word Co-Occurrence Statistics”, 2007,

What does this mean?

$$f(t) = \frac{1}{W} \sum_{c} n(c,t)$$

$$n(c,t)$$

- Word counts depend on co-occurrence counts!

$$\sum_{c} n(c,t)$$

- Sum all co-occurrence counts for *t*

What does this mean (cont.)?

$$f(t)=\frac{1}{W}\sum_{c} n(c,t)$$

$$\frac{1}{W}\sum_{c} n(c,t)$$

- Finally, divide by the total window size (*L* + *R*)

Counts for the co-occurrent and *N*

$$f(c) = \frac{1}{W} \sum_t n(c,t)$$

- Sum of the co-occurrences of *c* with every *t*

$$N = \frac{1}{W} \sum_t \sum_c n(c,t)$$

- Sum of the counts for every *t*

How to do this in Python

- *df* is a *nxn* DataFrame of co-occurrence counts
- c1 = np.sum(df, axis = 1) / 8 → Series
- c2 = np.sum(df) / 8 → Series
- N = np.sum(df.values) / 8 → float
- *NB*: np.sum(df, axis = 1) == np.sum(df)

And now, proceed as usual

- p = c2 / N
- p1 = c12 / c1
- p2 = (c2-c12) / (N-c1)
- LL = log $L(c_{12}, c_1, p)$ + log $L(c_2\text{-}c_{12}, N\text{-}c_1, p)$

  - log $L(c_{12}, c_1, p_1)$ - log $L(c_2\text{-}c_{12}, N\text{-}c_1, p_2)$

And then interpret the results

- Choose some important/interesting words
- Take look at the top 10 (or more) LL scores
- What do they tell you about that word in that text/corpus?
- What did you see that you expected?
- What did you see that you didn't expect?

An example: θεός in the OT

ἐναντίον	opposite	138.562030837
πατήρ	father	138.871378908
προσκυνέω	to bow down to	142.195328221
σωτήρ	savior	146.211537668
ὅδε	this	153.773856064
φοβέω	to fear	167.118568656
ἄνθρωπος	human being	203.996356483
ἕτερος	other	206.001410962
εὐλογέω	to praise	251.776073598
εὐλογητός	praise	310.444360915
λατρεύω	to serve	328.301017235

Homework I

- Take a text in your native language of at least 50,000 words (i.e., novel length)
- Produce co-occurrence matrices for 2L-2R, 4L-4R, 6L-6R, 8L-8R
- Produce LL matrices for each of these window sizes
- Choose 3 important words in the text
- Examine the top 10 LL scores

Homework II

- Write about a page about these 3 words with:
  - Tables of the top 10 LL scores for each word for each window size (4, 8, 12, 16)
  - Explanation of what these tables show you
  - Your judgment from these tables of which window size is best for this text in this language
    - Make sure to say here what evidence led your decision!

Homework III: What I want

- Your text
- 3 scripts:
  - To produce the 4 co-occurrence tables
  - To produce the 4 LL tables
  - To select the top 10 LL scores for every word in every LL table
- Your page about your results

Homework IV: A piece of advice

- Check your LL answers!
- I will
- And I will do it like this