UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Log-Likelihood Ratio and Co-occurrence Analysis

Why measure significance?

- What is the most frequent word in your native language?
  - Probably an article (the, der, die, das) or conjunction (and, und)
- This is probably also the most frequent co-occurrent with almost every word
- It doesn't make much sense that this would be the most important word semantically for every other word
- So we seek to measure significant co-occurrence

Significant Co-occurrence

- “Significant collocation is regular collocation between two items, such that they co-occur more often than their respective frequencies.” (León, 14)
- “log-likelihood measures the strength of association between words by comparing the occurrences of words respectively and their occurrences together.”
- “It is…a number that tells us how much more likely one hypothesis is than another” (Manning and Schütze, 172)

Why Log Likelihood?

- It deals better with sparsity than many other significance measures (e.g., chi-squared)
- Gives more easily interpretable results (you don't need a chi-squared table)
- It is computationally complex, but doesn't require as much in the way of resources as some other measures

UNIVERSITÄT LEIPZIG

Alexander von Humboldt Stiftung/Foundation

The Log-Likelihood Formula

$$\text{Log } \lambda = \log \frac{L(H_1)\ \#\text{independence}}{L(H_2)\ \#\text{dependence}}$$

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

Where $L(k,n,x) = x^k(1\text{-}x)^{n\text{-}k}$

So $\log L(c_{12}, c_1, p) = \log(p^{c_{12}}(1\text{-}p)^{c_1\text{-}c_{12}})$

c_1 = occurrences of word 1
c_2 = occurrences of word 2
c_{12} = co-occurrences of word 1 with word 2
N = number of tokens in the text

$p = c_2/N$
$p_1 = c_{12}/c_1$
$p_2 = (c_2\text{-}c_{12})/(N\text{-}c_1)$

Analyzing the LL Formula I

- $p = c_2/N$
  - probability of any word in the text being word 1
  - higher value suggests independence
- $p_1 = c_{12}/c_1$
  - probability that word 2 will appear if word 1 appears
  - higher value suggests dependence
- $p_2 = (c_2-c_{12})/(N-c_1)$
  - probability that word 2 appears apart from word 1
  - higher value suggests independence

Analyzing the LL Formula II

- How does LL deal with sparse data?
- As an example, $\log(p^{c12}(1-p)^{c1-c12})$
  - Notice what happens here as p gets larger
  - p^{c12} will get larger while $(1-p)^{c1-c12}$ will get smaller
  - So the maximum for this whole member of the equation is ~ p = 0.5

What does the LL formula produce?

- It produces a ratio comparing independence to dependence
- Since it is a ratio of natural logs, it can be positive or negative
- If it is positive, the words are more independent
- If negative, more dependent
- But we usually have one more step

The last step in the process

- Normally, the last thing done in the LL process is to multiply the answer by -2
- This gives us a number that we can compare to a chi-squared table to find a *p* value
- Once we have done this, dependence will be represented by a positive value, independence negative

The Chi-squared Table

- http://passel.unl.edu/Image/Namuth-CovertDeana956176274/chi-sqaure%20distribution%20table.PNG
- For LL tests, we always use 1 degree of freedom
- But this is not so important for us
- All that we want is the ratio of independence to dependence
- We are not hypothesis testing but hypothesis weighting
- This makes LL perfect for us

Conclusion

- We need to statistically weight our co-occurrence counts
- The log-likelihood ratio is good for this for several reasons
  - Deals well with sparsity
  - Gives more easily interpretable results
  - Requires only moderate computational resources, if done correctly
- And so, let's do it!

Literature

Manning, Chris and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: 1999.

Léon, Jacqueline. “Meaning by Collocation: the Firthian Filiation of Corpus Linguistics.” Online. http://htl.linguist.univ-paris-diderot.fr/leon/firth2007pdf.pdf