Word occurrence distributions, and their implications for statistical analysis and comparison of corpora

Jefrey Lijffijt

Data Science Lab @ Ghent University – iMinds
Outline

- Word frequencies
- Comparing corpora
- How about collocates
In a nutshell

The accuracy of any inference depends on model structure

Have to choose the right error term (distribution)

What ‘randomness’ do we observe in text corpora?
Examples of token counts
Example: British National Corpus

'\textit{the}' (BNC, $\sigma = 6043900$)

Number of texts

\begin{itemize}
\item Observed
\item Binomial
\end{itemize}
Example: British National Corpus

'\textit{the}' (BNC fiction prose, $\sigma = 820573$)
Example: British National Corpus

'for' (BNC, \( \sigma = 879020 \))
Example: British National Corpus

'for' (BNC fiction prose, $\sigma = 108950$)
Example: British National Corpus

'I' (BNC fiction prose, $\sigma = 279056$)
How about word choice?
Example: British National Corpus (Conversations)

'someone' vs. 'somebody' (σ = 2654)
Example: British National Corpus
(Fiction Prose)

'someone' vs. 'somebody' ($\sigma = 7875$)
Problem revisited

This is *not* binomial [Katz 1996, Church & Gale 1995, Church2000]

- Still typical assumption in Natural Language Processing, Machine Learning, Corpus Linguistics

What is appropriate then?

- General solution not known (afaik)
Solution for limited settings

Spatial distributions can be modelled fairly well

- Weibull distribution [Altmann et al. 2009]

Several attempts to model burstiness in topic models

- (Exponential) Dirichlet Compound Multinomial [Madsen et al. 2005, Elkan 2006]

Deep learning

- Long Short Term Memory [Hochreiter & Schmidhuber 1997]
Inter-arrival times

An inter-arrival time is the distance between two events

Here: two occurrences of the same word

Modelled ‘well’ by a stretched exponential (=Weibull) distribution [Altmann et al. 2009]
Inter-arrival time distributions: ‘for’
Inter-arrival time distributions: ‘I’

- Conversation
- Fiction
- Newspaper
- Academic

![Graphs showing inter-arrival times for different categories: Conversation, Fiction, Newspaper, Academic.](image-url)
Weibull dispersion parameter

Frequency

Weibull $\beta$ exponential distribution

0
0.5
1
1.5
2
2.5
Comparing corpora
What if we compare corpora

For example, we want to know whether one construct is more frequent in one setting than another.

How:

- Construct two sets of samples (texts)
- Run a statistical test
What not to do

Make a 2 x 2 table

\[
\begin{array}{ccc}
\text{Corpus} & \text{Form A} & \text{Form B} \\
X & a & b \\
Y & c & d \\
\end{array}
\]

Run $\chi^2$ test

- Test whether A vs. B is independent of X vs. Y

What goes wrong here?
Most texts in BNC would have significant over or underuse of ‘I’, when compared to the whole corpus.

I (σ = 868907)

- Green bars: Observed
- Purple line: Bag-of-words prediction

Number of texts

0 0.01 0.02 0.03 >=0.04

0 250 500 750 1000 1250
Quantification of the problem

Suppose we randomly split a corpus into two parts

Since the split is at random, the differences are pure noise

By definition of the p-value, an unbiased statistical test should then give uniform p-values
What should we do?

Two preferred solutions: [Lijffijt et al. online ahead of print]

- Permutation test or t-test
- In common: count frequencies per text, do not pool them

The distributions are not Normal, but t-test does not require that;

- Mean of a sum of bounded variables is approximately Normal
- Prefer permutation test if almost all counts are zero
Can be used to study language change

In Lijffijt, Säily & Nevalainen (2011), we study lexical change around the English Civil War (1642-1651).

Clearly, war is a more frequent topic after the war; increase in e.g., armies, colonel, militia, officers, ordnance, regiment, restoration and war.

Likelihood-ratio test gives about 20% false positives.

- Non-parametric variants are similar; non-parametric ≠ without assumptions [Lijffijt et al. online ahead of print]
Collocates
Degenerate example (BNC Fiction Prose)

Is ‘Peter Yeo’ a collocate?

Given a rate of 111 in 15M (Yeo);
the binomial tail probability = $1.2212 \cdot 10^{-15}$

However, this combination occurs only in 1 text.

What can we do here?
Take home message
The accuracy of any inference depends on model structure.

Have to choose the right error term (distribution).

Actual language modelling problem not solved.