Contrastive analysis

Computational Methods for Text Analysis

Pestova Alena Sergeevna

НИУ ВШЭ Санкт-Петербург

25.09.2021 / 03

Corpus-based contrastive analysis

The task is to extract vocabulary specific to a given corpus

- **Reference corpus** represents word usage in a language in general or in some subject area
- Build the frequency lists for the corpus of interest and the reference corpus
- Sort words by the difference in frequency of the studied corpus with the reference corpus
- Keywords of the studied corpus are at the top of the list - the words that are more specific to this particular corpus

Keywords of the corpus

Simple maths (by Adam Kilgarriff)

«this word is twice as common in this corpus as in that corpus»

- The simplest way
  - Normalize the frequencies
    - metric Instances per million (IPM)
  - Calculate the ratio of the normalized frequencies
  - Sort the lists by the calculated reatio

For example:

- Two corpora, the size of each corpus is one million tokens
- We do not need to normalize frequencies

Fc focus corpus — the studies corpus

Rc reference corpus

Problem 1: we cannot divide by 0

word	fc	rc	ratio
rarity	10	0	?
stir	100	0	?
yummy	1000	0	?

Standard solution - add 1:

word	fc	rc	ratio
rarity	11	1	11
stir	101	1	101
yummy	1001	1	1001

Problem 2: there are two many big ratios because of rare words

The frequency is also important. Solution: add n.

- $n = 1$

word	fc	rc	fc+n	rc+n	ratio	rank
rare	10	0	11	1	11,00	1
sometimes	200	100	201	101	1,99	2
frequent	12000	10000	12001	10001	1,20	3

- $n = 100$

word	fc	rc	fc+n	rc+n	ratio	rank
rare	200	300	300	400	0.75	3
sometimes	10	0	110	100	1,10	2
frequent	12000	10000	12100	10100	1,20	1

Normality and words distribution

In the paradigm of the standard statistical tests, there were problems with comparing frequiencies

- The normality assumption is unlikely in the case of words frequency distribution
- There are too many rare events in the language (remember Zipf's Law)
- Inapplicability of tests based on the assumption of normality (e.g. chi-square), at least to rare events (frequency < 5)

Dunning log-likelihood: motivation

Log likelihood ratio

A way to incorporate word frequencies into the statistical test paradigm:
Ted Dunning "Precise Surprise and Coincidence Statistics Methods (1994)

- Dunning log-likelihood is less dependent on an assumption of the distribution normality
- Therefore, it does not overestimate the detection of rare events so much and can be used for evaluation of not only the most frequent words

Dunning log-likelihood: formulas

	Corpus 1	Corpus 2	Total
Word Frequency	a	b	a+b
Frequency of other words	c	d	c+d
Total	a+c	b+d	a+b+c+d

Expected frequencies:

- Eij $\frac{R_i C_j}{N}$
- E1 $\frac{(a+b)(a+c)}{(a+b+c+d)}$
- E2 $\frac{(a+b)(b+d)}{(a+b+c+d)}$
- E3 $\frac{(c+d)(a+c)}{(a+b+c+d)}$
- E3 $\frac{(c+d)(b+d)}{(a+b+c+d)}$

Dunning log-likelihood: formulas

old:

$LL = G^2 = 2(a \log(a/E1) + b \log(b/E2)$

new (almost equal to the previous one):

$LL = G^2 =$

$2(a \log(a/E1) + b \log(b/E2) + c \log(c/E3) + d \log(d/E4))$

If we calculate the log-likelihood ratio test for two words in two corpora, then

$$G^2 \approx X^2(1)$$

We can calculate statistical significance of the difference (on the level of significance 0.05):

$$\text{p-value} : P(X^2 >= 3.84)$$

$$\text{CDF}(3.84) = 0.95 \text{for} X^2$$

Log-Likelihood ratio

- The practical effect of this improvement is that statistical textual analysis can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions
- It allows comparisons to be made between the significance of the occurrences of both rare and common phenomenon.
- More sensitive to frequent events (words) than to less frequent one [underestimates the degree of difference for less frequent words]

Log-odds

$$\mathrm{LR} = \log \frac{(a/sum(a))}{(b/sum(b))}$$

- G^2 for stat. significance and log-odds for effect size, sorting by log-odds
- sorting by log-odds and/or G^2, finding a threshold for cutting lists

Collocations

Collocations is a term that denotes pairs of words that not only stand togther frequently, but their appearence next to each other is above the mere chance. That means, there is statistically significant association between these words.

Conditional Probability

$$P(Event|Condition)$$

Conditional Probability

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \tag{1}$$

$$P(\text{intelligence}|\text{artificial}) = \frac{P(\text{artificial intelligence})}{P(\text{artificial})} =$$

$$= \frac{\frac{4}{46804371}}{\frac{121}{46804371}} = \frac{4}{121} = 0.033$$

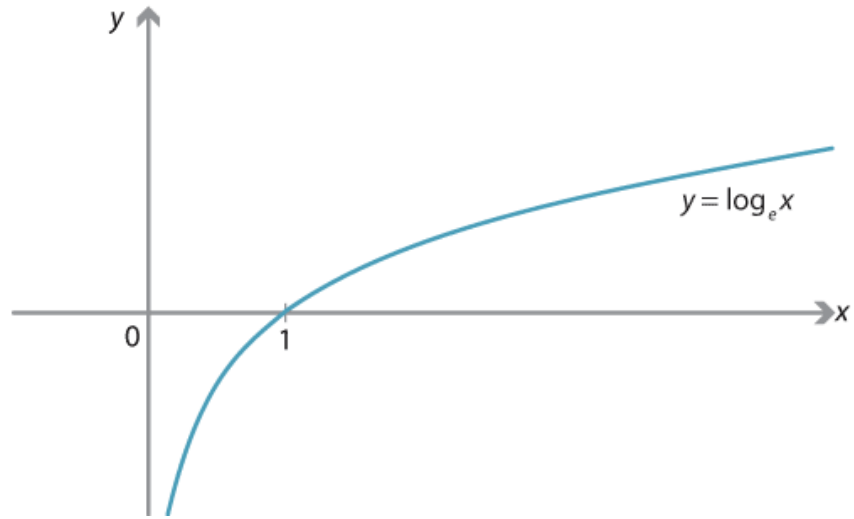
Pointwise mutual information

PMI

$$pmi(x;y) = log \frac{p(x,y)}{p(x)p(y)} =$$

$$= log \frac{p(x|y)}{p(x)} =$$

$$= log \frac{p(y|x)}{p(y)}$$



Positive PMI (negative values do not make much sense for word association measurements)

$$ppmi(x;y) = max(pmi(x;y), 0)$$

The problem of PMI

- Very sensitive to rare events that are highly informative relative to each other (words always occur together) [increase the degree of difference for rare words]

PMI can be used:

- For extracting vocabulary specific to a given corpus (as well as G2, log-odds, frequency ratios)
- For finding collocations in the text (as well as G2 and log-odds)

Contrastive analysis

We have seen how to calculate metrics for contrastive analysis with g2 and log-odds. Let's use PMI for the same thing.

word	fc	rc
sometimes	200	100
other words	10000	20000

- x - meeting the word "sometimes"
- y - corpus 1 (size of the corpus)
- $p(x|y) = 200/10200 = 0.0196$ - probablity of meeting word "sometimes"in corpus 1
- $p(x) = 300/(300 + 10000 + 20000) = 0.0099$ - probablity of meeting word "sometimes"in total (in corpus 1 and corpus 2 in our case)

$$pmi(x; y) = log\frac{p(x|y)}{p(x)} = log\frac{0.0196}{0.0099} = 0.9854$$

Collocations

bigram	word1 freq	word2 freq	bigram freq
united states	200	100	50
all other bigrams	10000	10000	...

- x - meeting the word "united"in the corpus
- y - meeting the word "states"in the corpus
- $p(x) = 200/10000 = 0.02$ - probability of meeting the word "united"
- $p(x|y) = 50/1000 = 0.05$ - probability of meeting the word "united"together with the word "states"

$$pmi(x; y) = log \frac{p(x|y)}{p(x)}$$

Dunning log-likelihood: formulas

	Word 2	not Word 2	Total
Word 1	a	b	a+b
not Word 1	c	d	c+d
Total	a+c	b+d	a+b+c+d

Expected frequencies:

Eij $\frac{R_i C_j}{N}$

E1 $\frac{(a+b)(a+c)}{(a+b+c+d)}$

E2 $\frac{(a+b)(b+d)}{(a+b+c+d)}$

E3 $\frac{(c+d)(a+c)}{(a+b+c+d)}$

E3 $\frac{(c+d)(b+d)}{(a+b+c+d)}$

$LL = G^2 =$

$2(a\log(a/E1) + b\log(b/E2) + c\log(c/E3) + d\log(d/E4))$

Collocations

Now we can not only extract frequent bigrams, but also we can find words that significantly associated with each other.
We may be interested in collocations with some specific word.

tidylo by Julia Silge: weighted log odds

1. Log odds ratio:

$$O_1 = \frac{f_{(w,c1)}}{N_{c1} - f_{(w,c1)}}$$

$$O_2 = \frac{f_{(w,c2)}}{N_{c2} - f_{(w,c2)}}$$

$$LO = log \frac{O_1}{O_2}$$

2. Weighted by uninformative Dirichlet prior:

$$\delta = \frac{\frac{f_{(w,c1)}+\alpha_{(w,c1)}}{N_{c1}+\alpha_{c1}-f_{(w,c1)}-\alpha_{(w,c1)}}}{\frac{f_{(w,c2)}+\alpha_{(w,c2)}}{N_{c2}+\alpha_{c2}-f_{(w,c2)}-\alpha(w,c2)}}$$

package tidylo in R