Aalto University
School of Science

Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping

Jefrey Lijffijt, Panagiotis Papapetrou,
Kai Puolamäki, and Heikki Mannila
Department of Information and Computer Science
Aalto University, Finland

Application

- Given two corpora (=collections of texts), *S* and *T*
- Find all words that are *significantly* more frequent in *S* than in *T*

Word	Freq in *S*	Freq in *T*
sergeant	57	32
Total	400.000	410.000

- Is this significant?

Motivation

- Find differences between groups
  - Age groups
    - S = 20-30, T = 40-50
  - Text types
    - S = newspaper, T = magazines
  - Author gender
    - S = male, T = female
  - News items
    - S = today, T = past year

Data

- Text = sequence of words
- Male vs. female authors
  - British National Corpus (BNC XML edition 2007)
  - 100 million words
  - 4000 texts
- Dates of important events
  - San Francisco Call Newspaper Corpus (Lappas et al. 2009)
  - 63 million words (after stop words removed)
  - 380.000 articles

Problem setting

- Input:
  - Two corpora: *S* and *T*
  - A significance threshold: *α (0 < α < 1)*

- Word *q* is dominant in *S* at level *α* if and only if

$$p = \Pr\left(\frac{freq(q,S)}{size(S)} \leq \frac{freq(q,T)}{size(T)} \right) \leq \alpha$$

Binomial test (bag-of-words model)

- **Assume all words are independent**
- Significance test using 2x2 table

Word	Freq in *S*	Freq in *T*
sergeant	57	32
Total	400.000	410.000

- $$p = \sum_{k=freq(q,S)}^{size(S)} \underbrace{\binom{size(S)}{k} p_{q,T}{}^{k} (1 - p_{q,T})^{size(S)-k}}_{\text{Probability of exactly } k \text{ occurrences}} \approx 2.2 \cdot 10^{-5}$$

Binomial test (bag-of-words model)

- **Assume all words are independent**
- Significance test using 2x2 table

Word	Freq in *S (male)*	Freq in *T (female)*
sergeant	57	32
Total	400.000	410.000

- $$p = \sum_{k=freq(q,S)}^{size(S)} \underbrace{\binom{size(S)}{k} p_{q,T}{}^{k} (1 - p_{q,T})^{size(S)-k}}_{\text{Probability of exactly } k \text{ occurrences}} \approx 2.2 \cdot 10^{-5}$$

Binomial test (bag-of-words model)

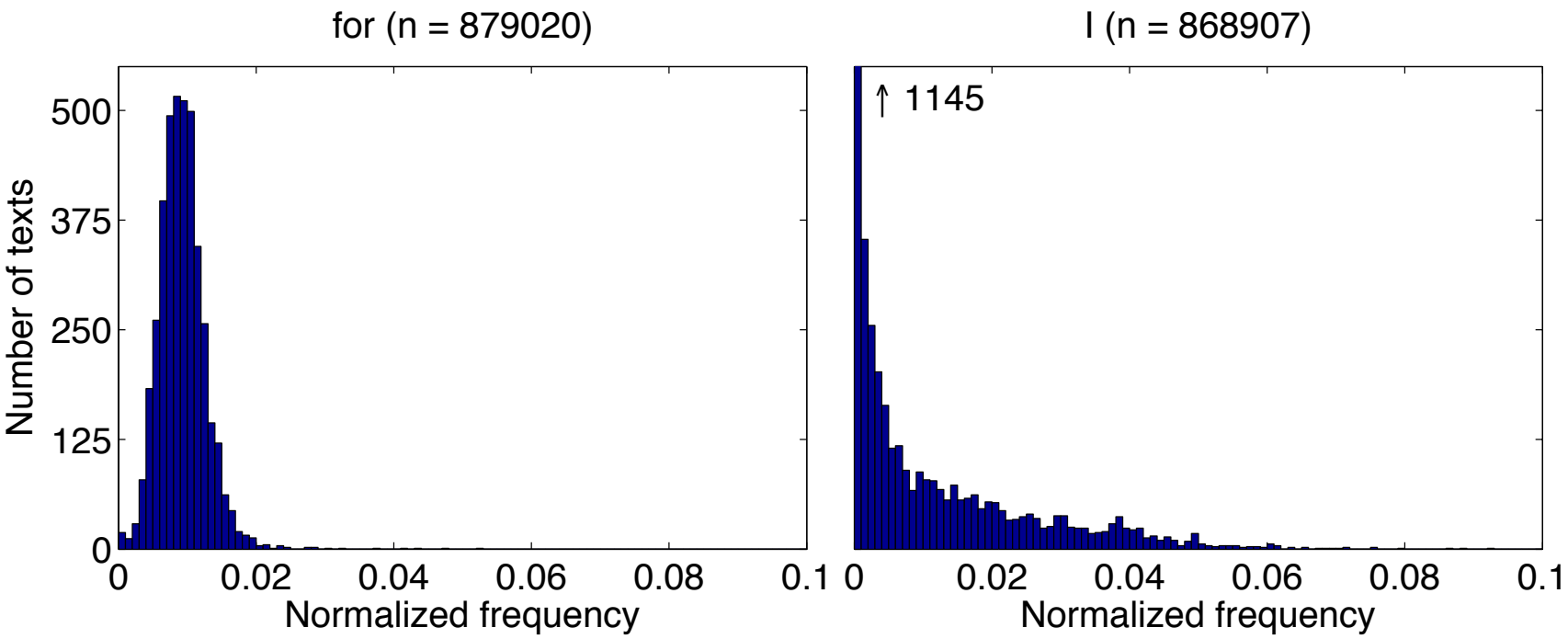
- **Assume all words are independent**
- *However:* texts have structure!

- Why the bag-of-words model then?
  - Mathematically simple
  - Computationally efficient

- Core questions:
  - Can we provide more realistic models?
  - Does it matter?

Many words are bursty

Data: British National Corpus, 4049 texts



- Frequency distribution differs per word
  - Depends on frequency and word ‘type’

Proposed method 1: Inter-arrival times

- Count space between consecutive occurrences of **and**

Finnair believes that it will be able to resume its scheduled service to **and** from New York on Monday, after two days of cancellations caused by hurricane Irene. All three airports serving New York City have been closed because of the hurricane **and** Finnair was forced to cancel flights on Saturday **and** Sunday. The airline is not certain when its scheduled service can be resumed, but the assumption is that Monday afternoon's flight from Helsinki will depart. Some Finnair passengers whose final destination is not New York have been rerouted **and** some have delayed travel plans. The company has also offered ticket holders a refund. *YLE*

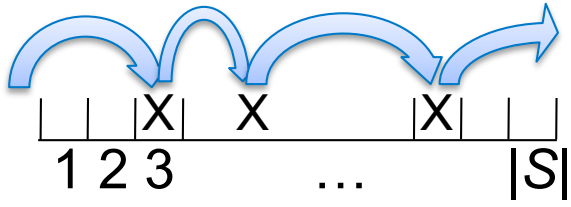
- $IA_{and} = \{29, 9, 39, 29\}$

- Hypothesis: this captures the behavior pattern of words

Proposed method 1: Inter-arrival times

- Count space between consecutive occurrences
  → Inter-arrival time distribution

- Resampling based on observed distribution in *T*
  0. *Optional*: fit Weibull distribution
  1. Pick random first index
  2. Sample random inter-arrival time
  3. Repeat 2. until size of corpus *S* exceeded



- Produce *N* random corpora: $R_1, \ldots, R_N$

- $$\hat{p} = \frac{1 + \sum_{i=1}^{N} I\left(freq(q,S) \leq freq(q,R_i)\right)}{1 + N}$$

Proposed method 2: Bootstrapping

- Based directly on word frequency distribution
- Resampling based on observed distribution in *T*
  - Number of samples equal to number of samples in *S*
- Repeat resampling to obtain *N* random corpora
- $$\hat{p} = \frac{1 + \sum_{i=1}^{N} I\left(\frac{freq(q,S)}{size(S)} \leq \frac{freq(q,R_i)}{size(R_i)} \right)}{1 + N}$$

Comparison for *sergeant*

Word	Freq in *S (male)*	Freq in *T (female)*
sergeant	57	32
Total	400.000	410.000

- $p_{binomial}$ = 0.000022
- $p_{IA\text{-}Weibull}$ = 0.08
- $p_{bootstrap}$ = 0.11

- Maybe the difference is not so significant!

Example: frequency thresholds

- $\alpha \leq 0.01$ in a text of 2000 words

Word	Freq in BNC ($x10^6$)	Weibull β	Binomial	Weibull Inter-arrival	Bootstrap
a	2.2	1.01	61	61	72
for	0.9	0.93	29	30	37
I	0.9	0.57	29	48	110

- β is the shape parameter of the Weibull fit
- Smaller β gives larger differences

Finding significant news events

- San Francisco Call Newspaper Corpus (Lappas et al. 2009)
- Term query: *Jacksonville*
- Event: great fire at Jacksonville, May 3rd, 1901

1901	0427	0428	0429	0430	0501	0502	0503	0504	0505	0506	0507	0508	0509	0510	0511	0512	0513	0514	0515	0516	0517	0518	0519	0520
Boot	.	.	.	.	.	.	.	.	x	x	.	x	.	.	.	.	.	x	.	.	x	.	.	.
IA_W	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
IA_E	.	.	.	.	.	.	.	.	x	x	.	x	.	.	.	.	.	x	.	.	.	.	.	.
Bin	.	.	.	.	.	.	.	.	x	x	x	x	.	.	.	.	x	x	.	.	x	.	.	.
Lappas	x	.	x	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	.	.	x	x	x	x

- Result: bootstrap/IA ≤ bag-of-words ≤ thresholding

Conclusion

- Bag-of-words model poorly represents frequency distributions (of *bursty* words)
- New models: inter-arrival times and bootstrap method
  - More conservative p-values
  - Weibull β predicts difference between models
- Future work:
  - Use inter-arrivals in other settings, e.g., information retrieval
  - Other statistics than word frequencies
- http://users.ics.tkk.fi/lijffijt/

Comparing p-values of all words