Distance Measures and Ranking Features

Peter Fankhauser

ESU in Digital Humanities, Leipzig 2015
Workshop: Comparing Corpora

INSTITUT FÜR DEUTSCHE SPRACHE
CLARIN-D
UNIVERSITÄT DES SAARLANDES

Language Variation

- Time, Region
- Cultural/Social Context
- Register: Field, Tenor, Mode
- Genre
- Authorship
- …

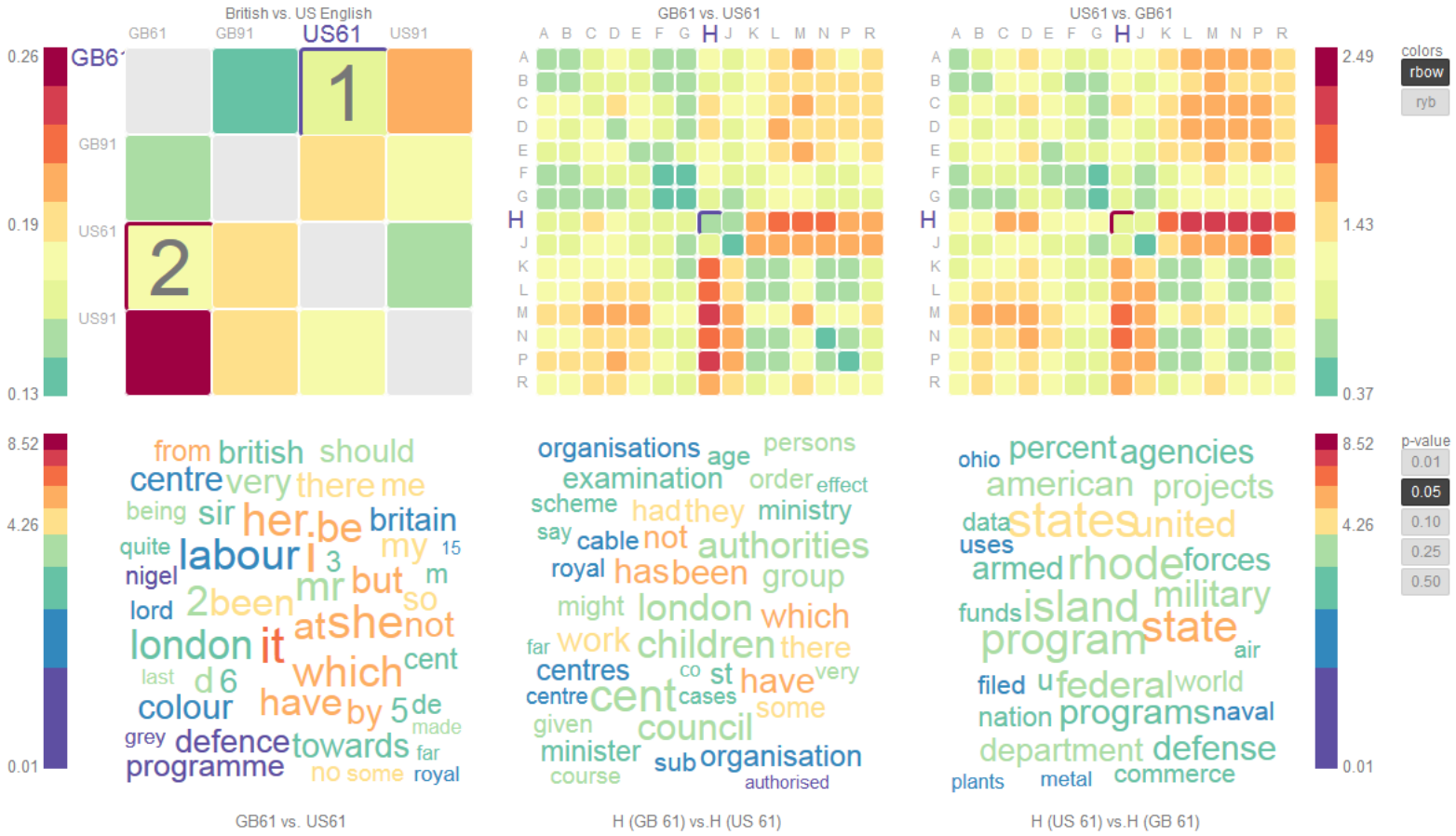
Questions & Desiderata

- Questions
  - Distance between two corpora
  - Most typical/distinguishing feature for a corpus (in comparison)
  - Representativeness of a feature
- Desiderata
  - Distances and feature rankings should be comparable
  - Independent of
    - Size of corpora
    - Inner complexity of corpora

Models & Applications

- Language Models
  - Bag of words: $p(w|Corpus)$;
    of ngrams: $p(w_1 w_2|Corpus)$
  - Topic models: $p(topic|Corpus)$;
    $p(w|topic)p(topic|doc)$
  - Syntactic classes: $p(pos|Corpus)$;
    $p(w|pos)p(pos|pos_{-1}, pos_{-2})$
  - …
- Applications
  - Classification
  - Retrieval
  - *Understanding Variation*

Example
British vs. American English over Time



Chi-square χ^2

- Most basic, widely used “distance measure”
- Contingency Table

	Press Reportage	Press Editorial
„very“	108	60
not „very“	86131	32938

- Observed Frequencies for „very“
- Null Hypothesis H_0: observed frequencies come from the same underlying distribution
- H_1: frequencies come from different distributions

Observed vs. Expected Frequencies

- Observed

	Press Reportage	Press Editorial	Row Totals
„very“	108	60	168
not „very“	86131	32938	119069
Column Totals	86239	32998	119237

- Expected, if H_0 holds

	Press Reportage	Press Editorial	Row Totals
„very“	168*86239/119237	168*32998/119237	168
not „very“	119069*86239/119237	119069*32998/119237	119069
Column Totals	86239	32998	119237

- Expected = RowTotals*ColumnTotals/Total

$$e_{ij} = (o_{i1} + o_{i2})(o_{1j} + o_{2j})/(o_{11} + o_{12} + o_{21} + o_{22})$$

Observed vs. Expected Frequencies

- Observed

	Press Reportage	Press Editorial	Row Totals
„very“	108	60	168
not „very“	86131	32938	119069
Column Totals	86239	32998	119237

- Expected, if H_0 holds

	Press Reportage	Press Editorial	Row Totals
„very“	121.5	46.5	168
not „very“	86117.5	32951.5	119069
Column Totals	86239	32998	119237

- Expected = RowTotals*ColumnTotals/Total

$$e_{ij} = (o_{i1} + o_{i2})(o_{1j} + o_{2j})/(o_{11} + o_{12} + o_{21} + o_{22})$$

Chi-square Value and Error Probability

- $\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

	Press Reportage	Press Editorial
„very“	$\frac{(108 - 121.5)^2}{121.5}$	$\frac{(60 - 46.5)^2}{46.5}$
not „very“	$\frac{(86131 - 86117.5)^2}{86117.5}$	$\frac{(32938 - 32951.5)^2}{32951.5}$

Chi-square Value and Error Probability

- $\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 5.433$

	Press Reportage	Press Editorial
„very“	1.502	3.924
not „very“	0.002	0.006

- 1 Degree of Freedom:
  Error probability: $pvalue = 0.02$
- We can reject H_0 with 98% confidence
  - The frequency of “very” differs significantly
  - “very” is relatively more frequent in Press Editorial

Influence of sample size on Chi-square

- Compare

	Reportage	Editorial
„from“	455	150
not „from“	85784	32848

$\chi_1^2 = 2.521$
$p_1 = 0.112$
not significant

	Reportage	Editorial
„from“	910	300
not „from“	171568	65696

$\chi_2^2 = 2\chi_1^2 = 5.043$
$p_2 = 0.025$
significant

- (Almost) everything gets kind of significant given large enough corpora
- Chi-square values not comparable for corpora of different size
- Pearson‘s $\phi^2 = \frac{1}{n}\chi^2$

Other issues with Chi-square

- Bias of expected values by larger corpus
- Bad approximation of Chi-distribution for small samples
- Bad approximation of Chi-distribution for large differences ($o_{ij} > 2e_{ij}$)
- David MacKay
  „Message to teachers: more Bayes‘ theorem, less chi-squared.“

Language Models

- Unigram Language Models: Multinomial Distributions

$$\hat{p}(w|C) = \frac{f(w, C)}{\sum_i f(w_i, C)}$$

  The observed probability $\hat{p}(w|C)$ of word w in corpus C equals its frequency $f(w, C)$ divided by the total number of words in C.

- Smoothing: avoid zeroes

$$p(w|C) = \lambda\hat{p}(w|C) + (1 - \lambda)p(w)$$

  Jelinek-Mercer: Estimate the probability $p(w|c)$ by a mixture of its observed probability $\hat{p}(w|C)$ and its overall probability $p(w)$ in some background corpus.

- Other Smoothing Methods: Laplace, Dirichlet, etc.

Information Theory: Basics (1)

- Length of *optimal* encoding for a word (in bits)

$$L(w|C) = -log_2 p(w|C)$$

- Frequent words, few bits

$$p(the) = 0.08, L(the) = 3.64\ bits$$

- Rare words, many bits

$$p(several) = 0.00062, L(several) = 10.66\ bits$$

- Number of bits to encode the entire corpus C

$$L(C) = -\sum_i f(w, C) log_2 p(w|C)$$

Information Theory: Basics (2)

- Entropy: average number of bits per word

$$H(C) = -\sum_i p(w_i|C) log_2 p(w_i|C)$$

- Cross-Entropy: encoding corpus C_1 with an optimal encoding based on corpus C_2

$$H(C_1; C_2) = -\sum_i p(w_i|C_1) log_2 p(w_i|C_2)$$

Information Theory: Basics (3)

- Relative Entropy (Kullback-Leibler Divergence)
- *Additional* bits needed when encoding C_1 with the optimal encoding for C_2

$$D_{KL}(C_1||C_2) = H(C_1; C_2) - H(C_1)$$
$$= \sum_i p(w_i|C_i) log_2 \frac{p(w_i|C_1)}{p(w_i|C_2)}$$

- Minimum: $D_{KL}(C_1||C_2) = 0$ for $C_1 = C_2$
- Asymmetry: $D_{KL}(C_1||C_2) \neq D_{KL}(C_2||C_1)$

Example

- Frequencies

	Press Reportage	Press Editorial
„very“	108	60
„from“	455	150
Corpus Size	86239	32998

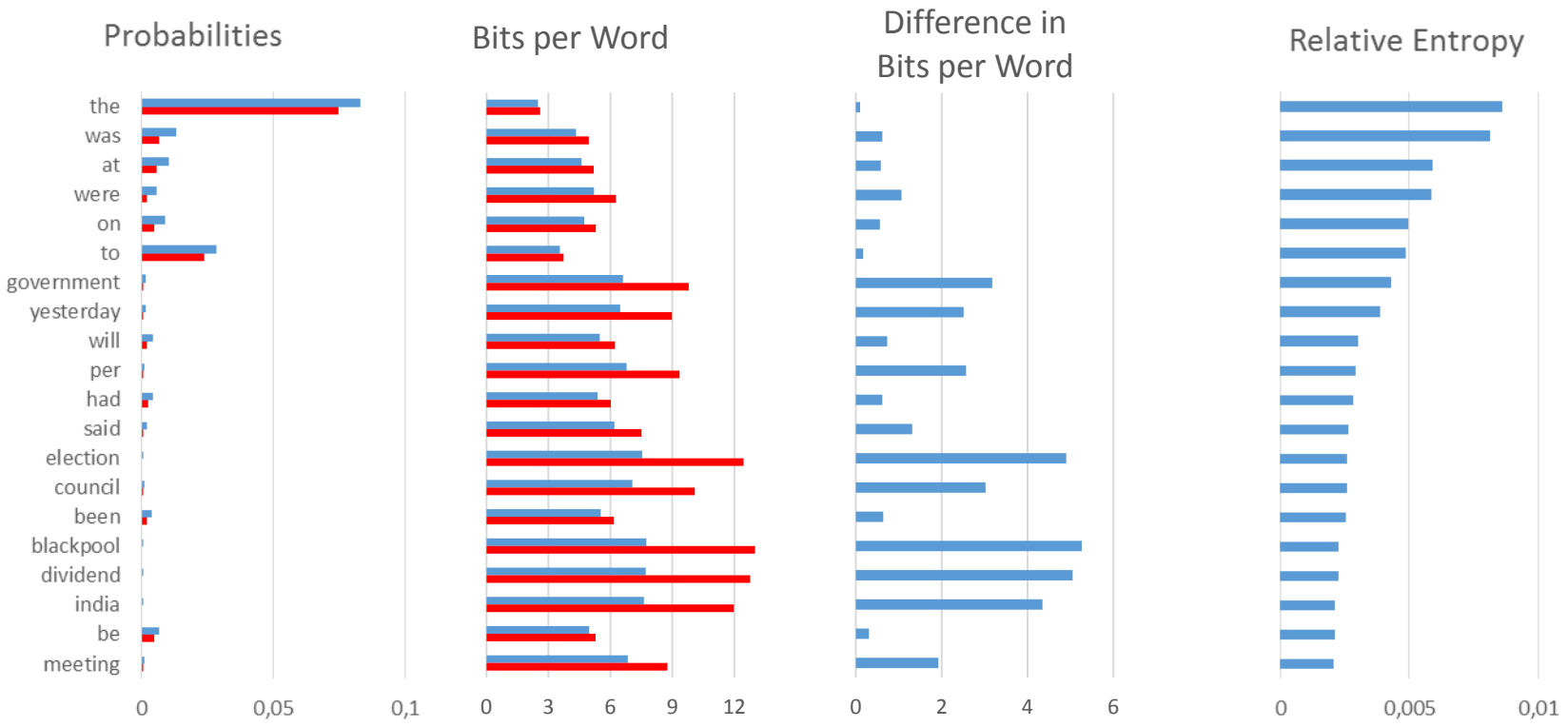
- (Unsmoothed) Probabilities (%)

w	$p(w\|Reportage)$	$p(w\|Editorial)$
„very“	0.13	0.18
„from“	0.53	0.45

- Bits

w	$-log\ p(w\|R)$	$-log\ p(w\|E)$	$log\frac{p(w\|R)}{p(w\|E)}$	$log\frac{p(w\|E)}{p(w\|R)}$	$p(w\|R)\ log\frac{p(w\|R)}{p(w\|E)}$	$p(w\|E) log\frac{p(w\|E)}{p(w\|R)}$
„very“	9.64	9.10	-0.54	0.54	-0.00067	0.00098
„from“	7.57	7.78	0.21	-0.21	0.00113	-0.00098

Top words for Reportage vs. Editorial



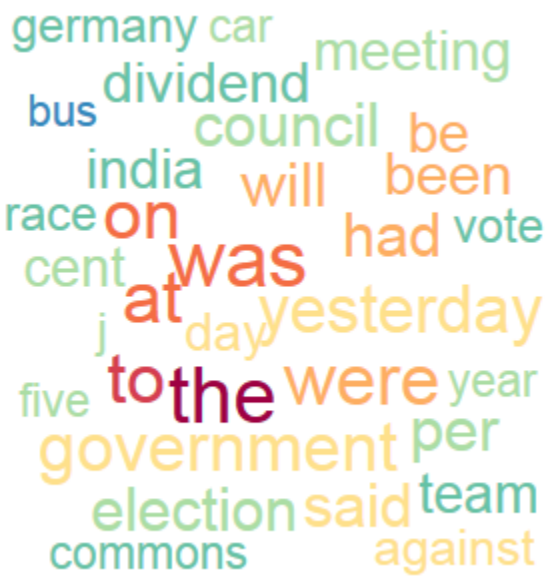
$p(w|R)$
$p(w|E)$

$-log_2 p(w|R)$
$-log_2 p(w|E)$

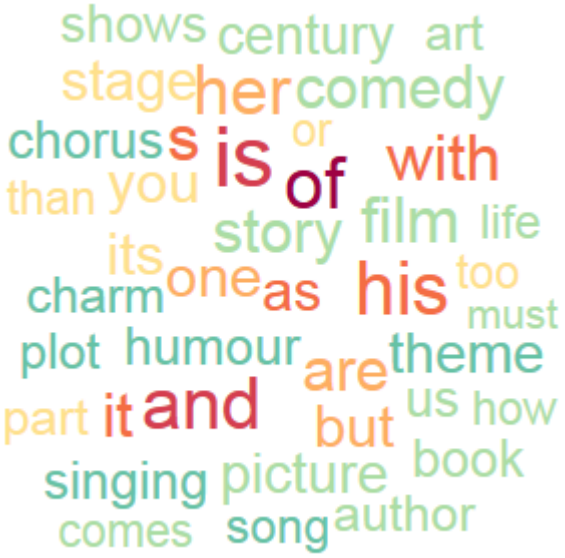
$log_2 \frac{p(w|R)}{p(w|E)}$

$p(w|R) log_2 \frac{p(w|R)}{p(w|E)}$

Both directions



Reportage vs. Editorial



Editorial vs. Reportage

What about Symmetry?

- Mean probability

$$p(w|M) = (p(w|C_1) + p(w|C_2))/2$$

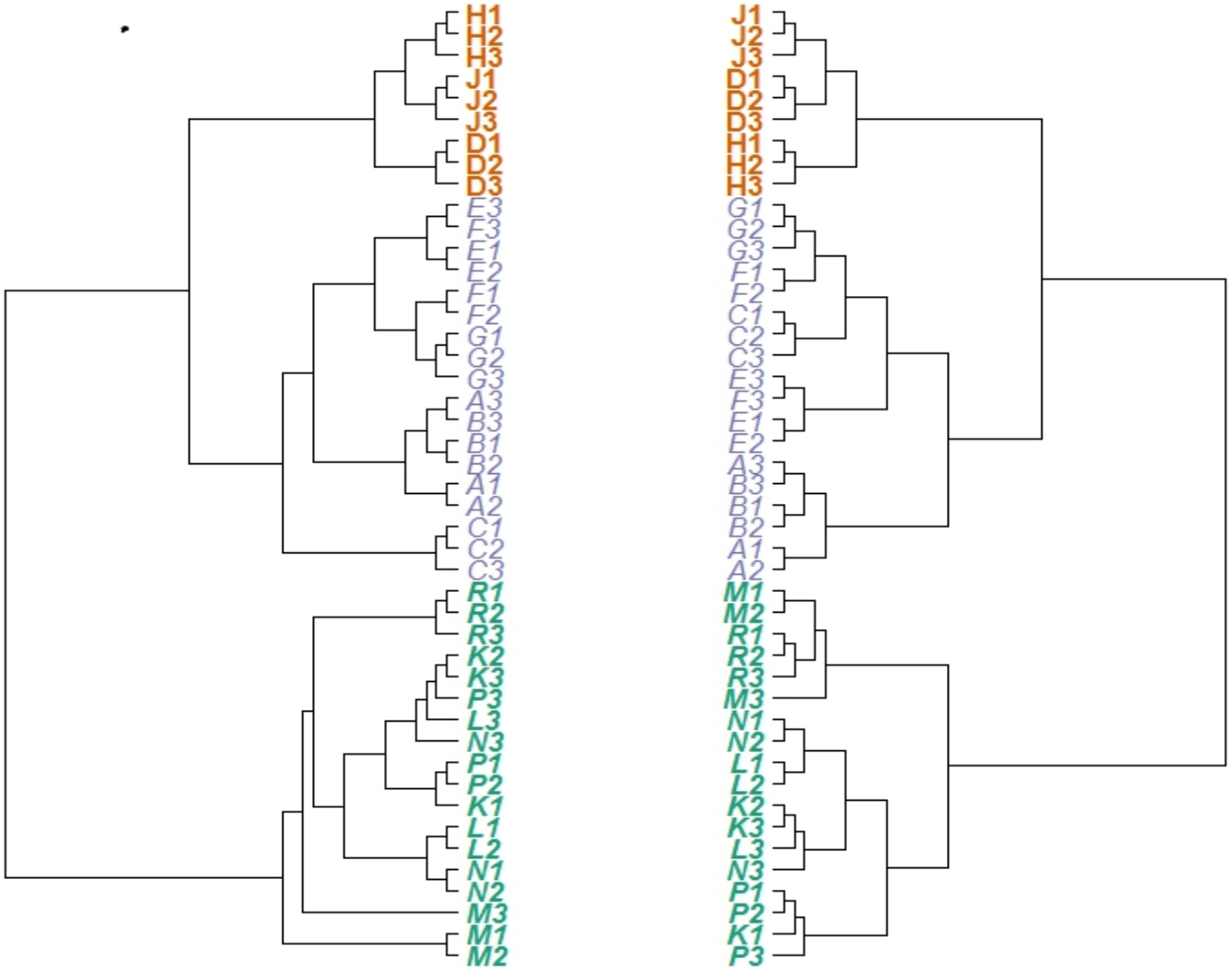
- Jensen-Shannon Divergence

$$D_{JS}(C_1||C_2) = (D_{KL}(C_1||M) + D_{KL}(C_2||M))/2$$

- Properties
  - Reflexive: $D_{JS}(C||C) = 0$
  - Symmetric: $D_{JS}(C_1||C_2) = D_{JS}(C_2||C_1)$
  - (Triangle Inequality)

$$\sqrt{D_{JS}(C_1||C_2)} \leq \sqrt{D_{JS}(C_1||C_3)} + \sqrt{D_{JS}(C_3||C_2)}$$

Cluster by D_{JS}vs. Pearson‘s ϕ^2



Comparison of Measures (1)

	Normalized	Unnormalized
Jensen-Shannon	$\frac{1}{2}\sum_{k=1,2}\sum_{i} p(w_i\|C_k)\ln\frac{p(w_i\|C_k)}{p(w_i\|M)}$	$\frac{1}{2}\sum_{k=1,2}\sum_{i} f(w_i,C_k)\ln\frac{p(w_i\|C_k)}{p(w_i\|W)}$
Chi-Square	$\sum_{k=1,2}\sum_{i}\frac{(p(w_i\|C_k)-p(w_i\|M))^2}{p(w_i\|M)}$	$\sum_{j=1,2}\sum_{i}\frac{(f(w_i,C_k)-e(w_i))^2}{e(w_i)}$

- Do C_1 and C_2come from the same distribution?
- $\chi^2(C_1,C_2) \approx 4\, D_{JS}(C_1||C_2)$
- $p(w|W) = (f(w,C_1)+f(w,C_2)/(\sum_{k=1,2}\sum_i f(w_i,C_k))$
  the mean probability *weighted* by size
- Natural logarithm *ln* instead of log_2
- Jensen-Shannon (unnormalized) a.k.a G-Test, Dunning's LogLikelihood Ratio

Comparison of Measures (2)

	Normalized	Unnormalized
Kullback-Leibler	$\sum_i p(w_i\|C_1)\ln\frac{p(w_i\|C_1)}{p(w_i\|C_2)}$	$\sum_i f(w_i,C_1)\ln\frac{p(w_i\|C_1)}{p(w_i\|C_2)}$
One-Way Chi-Square	$\sum_i \frac{(p(w_i\|C_1)-p(w_i\|C_2))^2}{p(w_i\|C_2)}$	$\sum_i \frac{(f(w_i,C_1)-e(w_i\|C_2))^2}{e(w_i\|C_2)}$

- Does C_1 come from the distribution of C_2?
- How surprising is C_1 when using a lexicon/vocabulary optimized for C_2?
- $\chi^2(C_1,C_2) \approx 2\ D_{KL}(C_1||C_2)$
- Expected frequency

$$e(w|C_2) = p(w|C_2) * \sum_i f(w_i,C_1)$$

Cheat Sheet

- Log Likelihood

$$2\sum_i o_i ln\frac{o_i}{e_i}$$

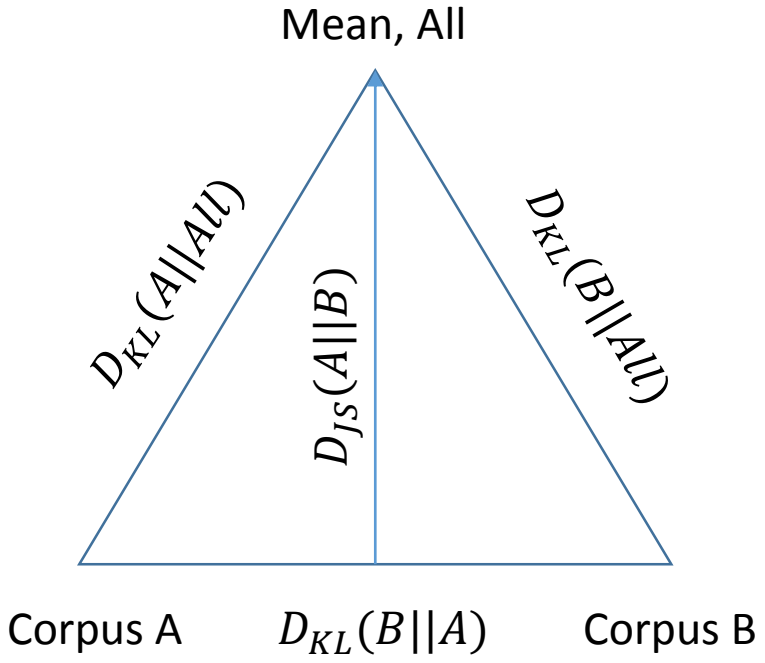
- Chi-Square

$$\sum_i \frac{(o_i - e_i)^2}{e_i}$$

- Definitions of observed o_i and expected e_i depend on normalization and direction

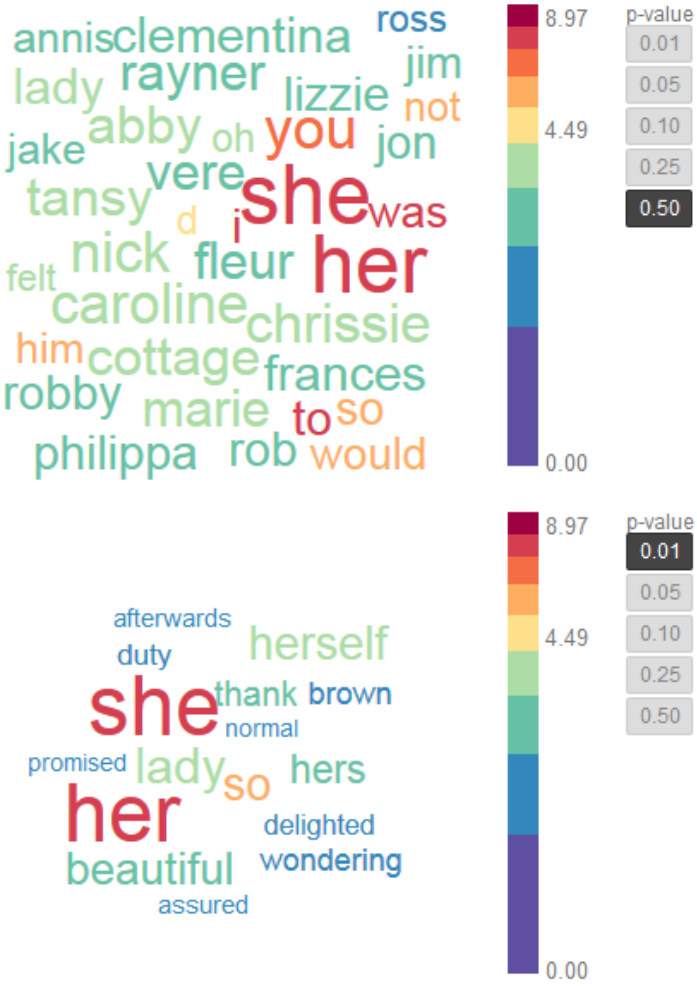
The big picture

- Relative Entropy D_{KL} for
  - Feature Ranking
  - Comparing subcorpus with corpus
  - Comparing document with corpus
  - Diachronic Change
- Jensen-Shannon D_{JS} for
  - Testing Null-Hypothesis (with caution!)
  - Comparing corpora/documents with each other

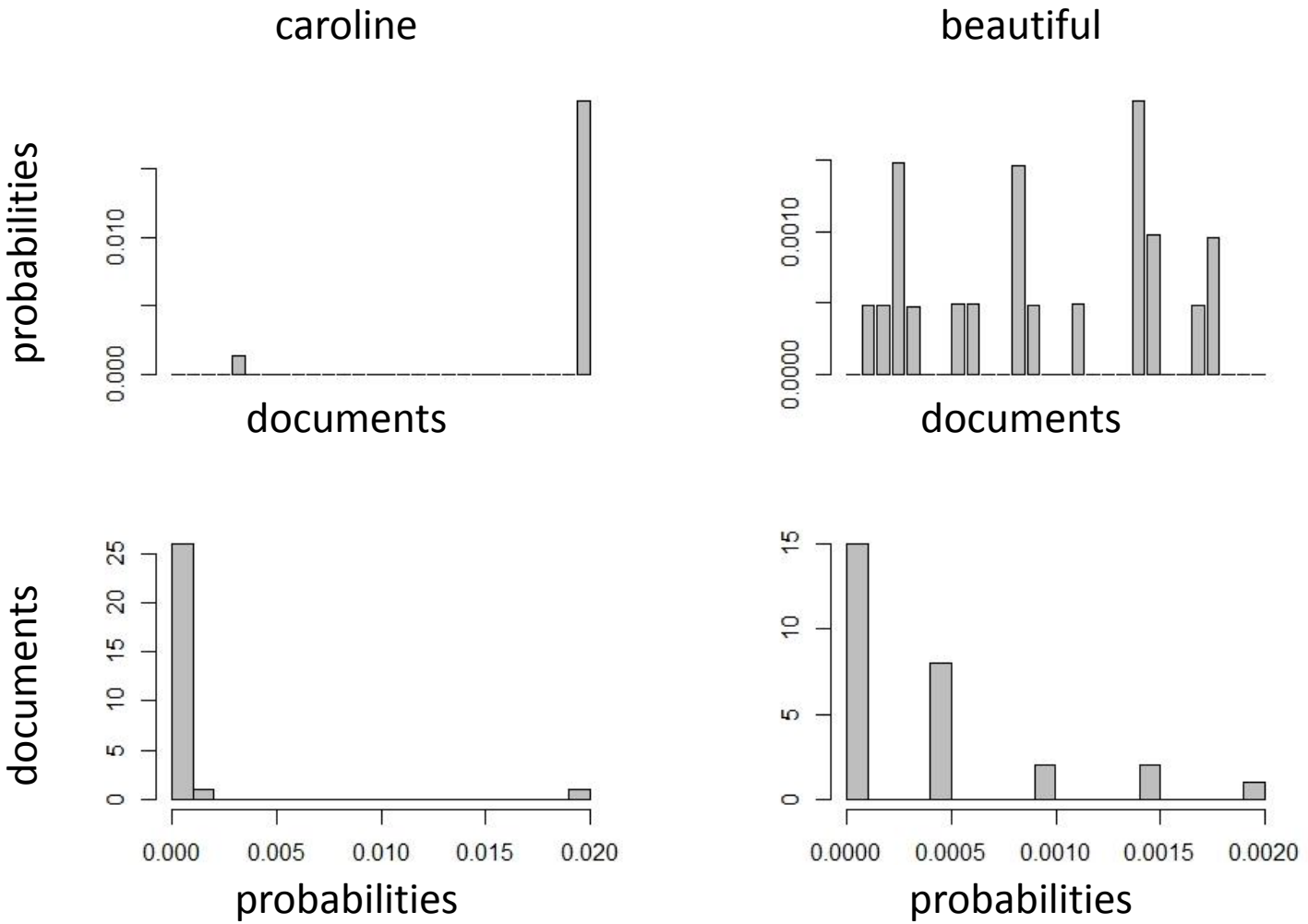


Significance/Representativeness

- Example
  Romance (P91) vs. Adventure (N91)
- 28 documents each
- Frequencies
  f(caroline,P91)= 45
  f(beautiful,P91)=29
  f(caroline,N91)=f(beautiful,N91)=0
- Contribution to $D_{KL}(P91||N91)$
  caroline: 0.00078
  beautiful: 0.00040



Distributions in Romance 91



(Welch) T-Test

- Mean of *individual* probabilities of a word in each *document* (Micro Average)
- Basic idea: Difference between *means* taking into account their *variance*
- Requires: normal distributions (not the case here)

data: carolineP91 and carolineN91
t = 1.0712, df = 27, p-value = 0.2936

not significant

data: beautifulP91 and beautifulN91
t = 3.7838, df = 27, p-value = 0.0007818

significant

- Assumption of normal distribution overestimates influence of variance -> overestimates error probability

(Mann-Whitney) U-Test

- Non-parametric test:
  no assumption about the distributions
- Basic idea
  - Sort probabilities from *P91* and *R91* in descending order
  - Compare sum of ranks for *P91* vs. *R91*
  - Take care of ties

data: carolineP91 and carolineN91
W = 420, p-value = 0.1611

Still not significant
Smaller error probability

data: beautifulP91 and beautifulN91
W = 574, p-value = 5.818e-05

Even more significant

Wrap up

- Distance Measures
  - One way: asymmetric
  - Two way: symmetric
  - Role of (length)*normalization*: comparability
  - Intuitive information theoretic basis
  - Many names for the same thing:
    LLR, G-Test, KL-Divergence/JS-Divergence
- Ranking Features by Indicativeness/Typicality
  - Contribution to Distance by Feature
  - Representativeness: Beware of Dispersion/Variance

One more Thing

- Multinomial Naive Bayes Classifier

$$p(C|d) \propto p(C) \prod_i p(w_i|C)^{f(w_i,d)}$$

- Classify document d into class C_k with $\arg\max_k p(C_k|d)$
- Logarithm

$$\log p(C|d) \propto \log p(C) + \sum_i f(w_i, d) \log p(w_i|C)$$

- Bonus Questions
  - How does this relate to cross entropy?
  - Why don‘t we need to subtract the entropy of d (for classification)?
  - Why is $\log p(w_i|C)$ not really a good indicator for the importance of a word?