Statistical Structure in Natural Language

Tom Jackson

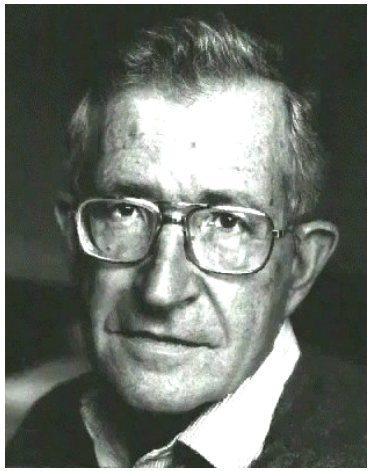
April 4th PACM seminar

Advisor: Bill Bialek

Some Linguistic Questions

- Why does language work so well?
  - Unlimited capacity and flexibility
  - Requires little conscious computational effort
- Why does language work at all?
  - Wittgenstein's dilemma
  - Rather accurate transmission of meaning from one mind to another
- How do we acquire language so easily?
  - Poverty of the stimulus
  - Everyone learns 'the same' language

Some Linguistic Answers



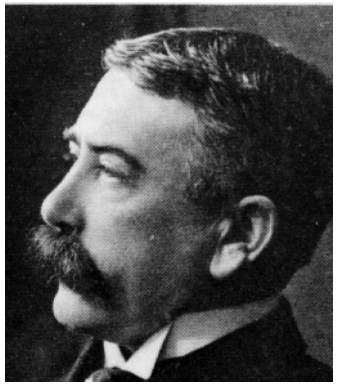
Noam Chomsky

- Language as 'words and rules'
  - Semantic and syntactic components of language
- Hierarchial structure
  - Morphology (intra-word)
  - Syntax (intra-sentence)
  - Compositional rules and paragraph structure
- Chomsky: Grammar must be 'universal'
- Nowak: evolution of syntactic communication
  - More efficient to express an increasing number of concepts using combinatorial rules.

Computational Linguistics

- Language performance vs. language competence
  - The 'Chomskyan Revolution'
  - Linguistics as psychology
  - Corpus-based methods outside the mainstream
- Inherent drawbacks
  - No access to contextual meaning
  - Word meaning only defined through relationships to other words
  - Vive la différance!

Ferdinand de Saussure



- Relevant for applications
- What are appropriate metrics to use?

Information Theory

Claude Shannon

- Entropy: information capacity

$$H(X) = -\sum_{\{x\}} p_x \log p_x$$

- Mutual information: Information X provides about Y and vice versa.

$$I(X;Y) = H(X) + H(Y) - H(X \cdot Y)$$

$$= \sum_{\{x,y\}} p_{xy} \log \frac{p_{xy}}{p_x p_y}$$

Naïve Entropy Estimation

$$\hat{H}(\{n_i\}) = -\sum_{i=1}^{K} \frac{n_i}{N} \log \frac{n_i}{N}$$

- Pro: Entirely straightforward
- Con: Useless for interesting case K<N
  - In the undersampled regime, just because we don't see an event doesn't mean it has zero probability.
  - Analogous to overfitting a curve to data
  - More appropriate to consider a "smooth" class of distributions
- Problem: How to apply Occam's principle in an unbiased way?

Bayesian Inference

- Bayes' Theorem: $\Pr(\text{model}\,|\,\text{data}) = \dfrac{\Pr(\text{data}\,|\,\text{model})\Pr(\text{model})}{\Pr(\text{data})}$

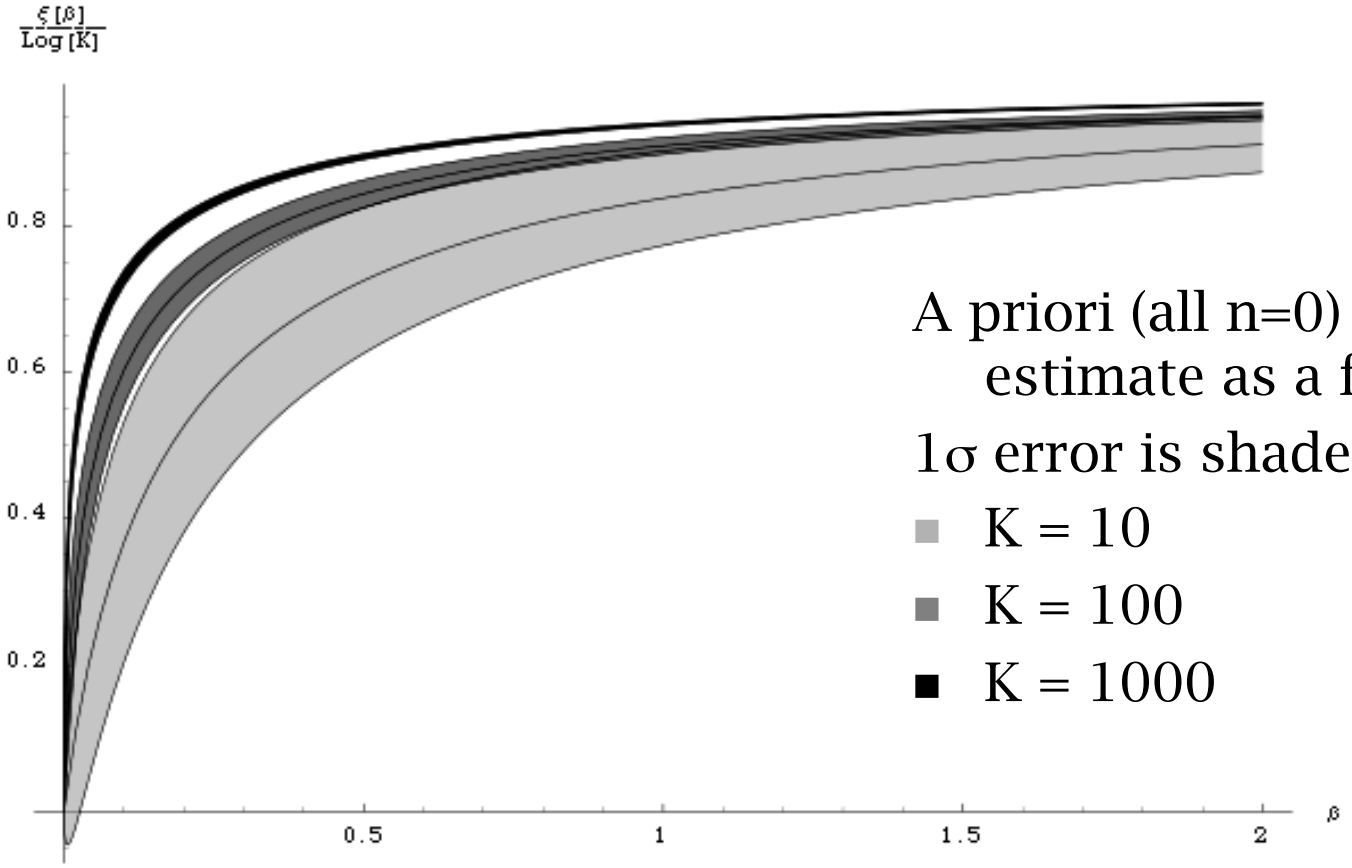
- Here our prior is on the space of distributions…

$$\frac{1}{Z}\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1}$$

- Generalize uniform prior to the Dirichlet prior parameterized by β.

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + K\beta}$$

Bias in the Dirichelet Prior



A priori (all n=0) entropy estimate as a function of β

1σ error is shaded

- K = 10
- K = 100
- K = 1000

- Prior highly biased for fixed values of β
- Obtain an unbiased prior by averaging over β

The estimator

$$\hat{S}^m = \frac{1}{Z(\{n_i\})} \int d\beta \frac{d\xi}{d\beta} \rho(\{n_i\},\beta) \left\langle S^m \right\rangle_\beta$$

$$\rho(\{n_i,\beta\}) = \frac{\Gamma(K\beta)}{\Gamma(N+K\beta)} \prod_{i=1}^{K} \frac{\Gamma(n_i+\beta)}{\Gamma(\beta)}$$

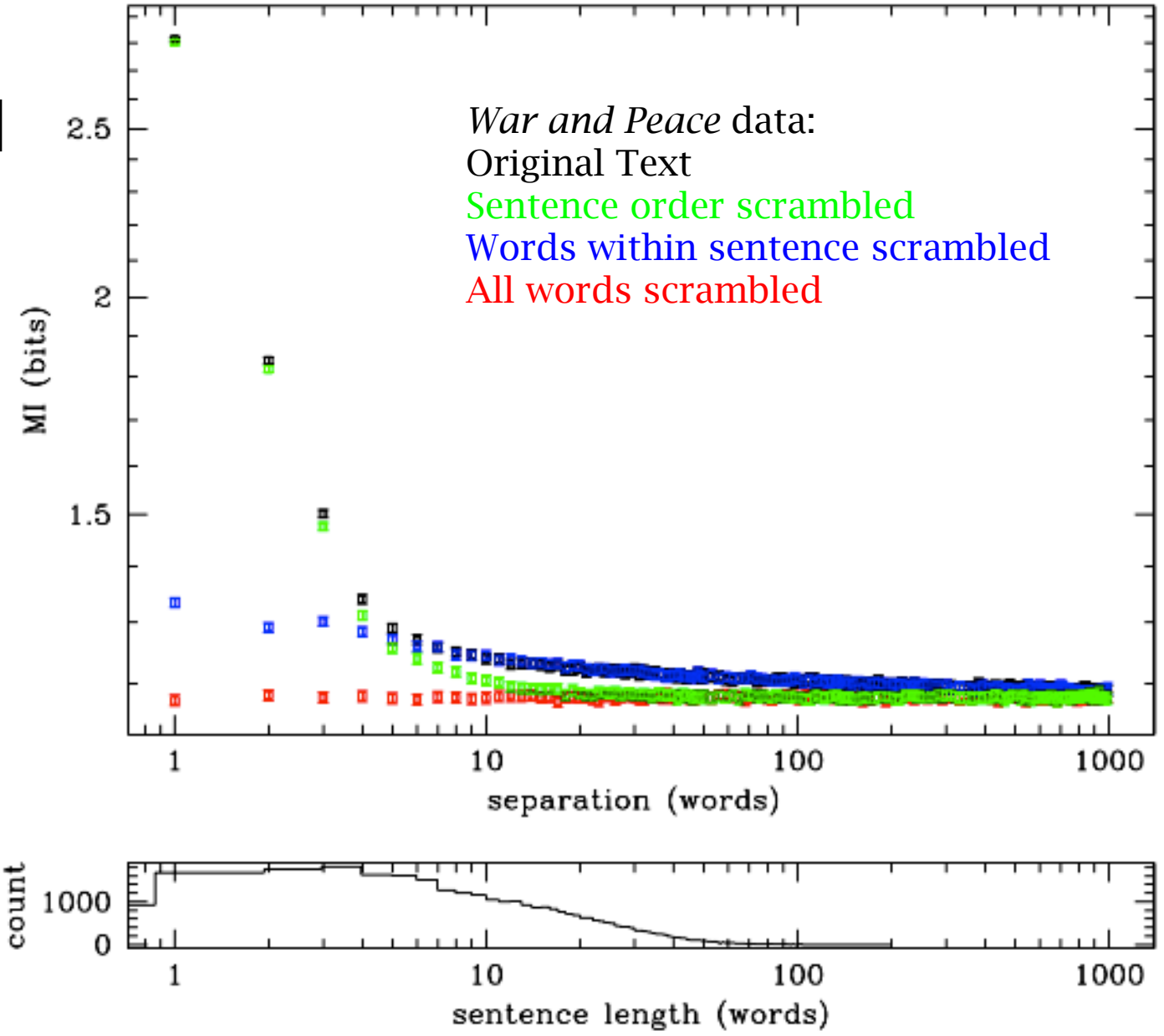
$$\left\langle S \right\rangle_\beta = -\sum_{i=1}^{K} \frac{n_i+\beta}{N+K\beta} \left(\Psi(n_i+\beta+1) - \Psi(N+K\beta+1) \right)$$

- Computational complexity still linear in N

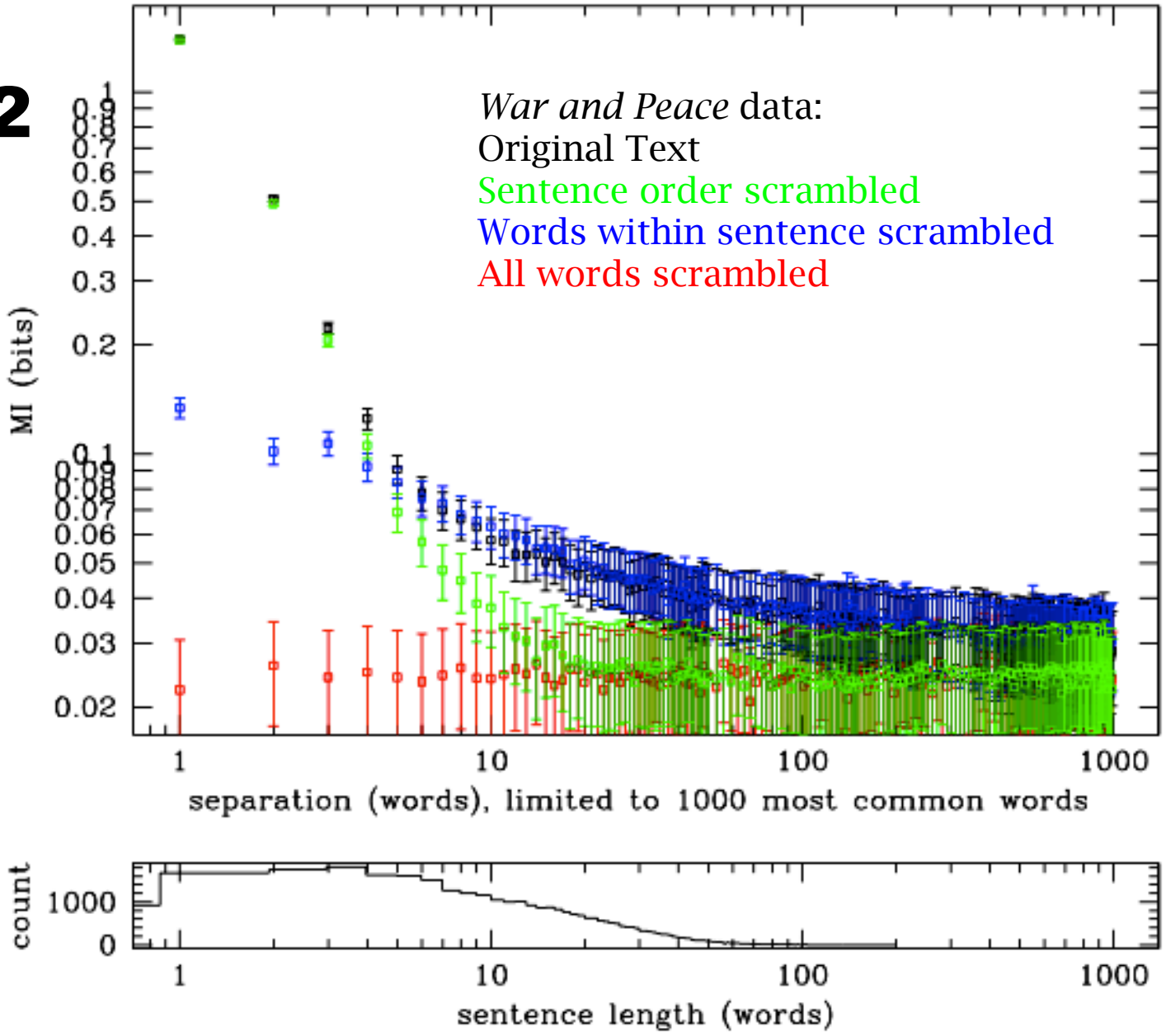
Getting to Work

- Summer '03 research project
- MI estimation algorithm implemented in ~5K lines of C code
- Texts used:
  - Project Gutenberg's *Moby Dick*
  - Project Gutenberg's *War and Peace*
    - K=18004, N = 567657
  - Reuters news articles corpus
    - K=40696, N = 2478116

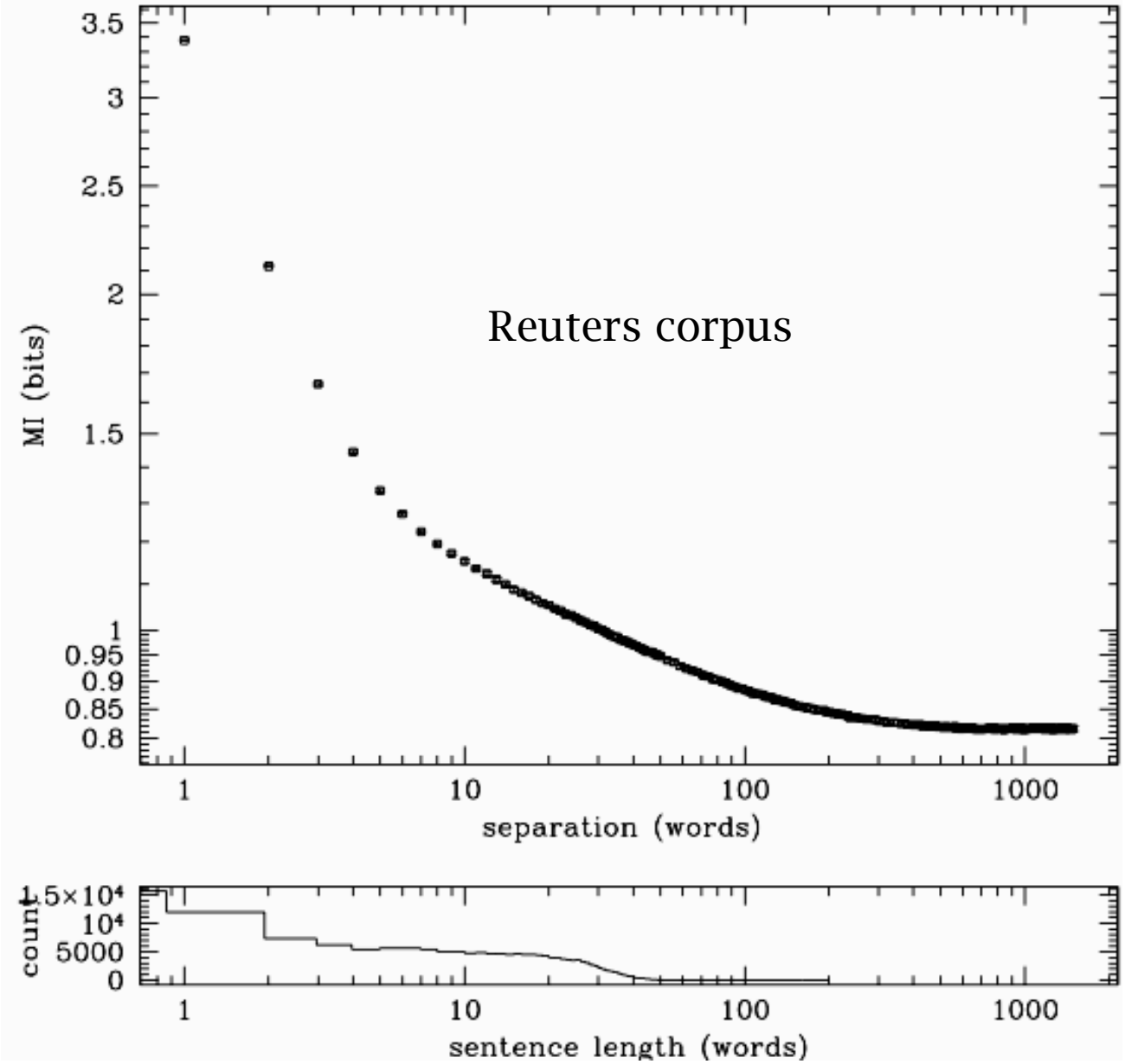
Results 1


War and Peace data:
Original Text
Sentence order scrambled
Words within sentence scrambled
All words scrambled
MI (bits)
2.5
2
1.5
1
10
100
1000
separation (words)
count
1000
0
1
10
100
1000
sentence length (words)

Results 2


War and Peace data:
Original Text
Sentence order scrambled
Words within sentence scrambled
All words scrambled
MI (bits)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
1
10
100
1000
separation (words), limited to 1000 most common words
count
1000
0
1
10
100
1000
sentence length (words)

Results 3


Reuters corpus
MI (bits)
3.5
3
2.5
2
1.5
1
0.95
0.9
0.85
0.8
1
10
100
1000
separation (words)
count
1.5×10⁴
10⁴
5000
0
1
10
100
1000
sentence length (words)

Information Bottleneck

- Are we really observing a crossover between syntactic and semantic constraints?
  - What "carries" the information at a given separation?
- Information Bottleneck method: compress X into X' while maximizing I(X' ;Y).
  - Two algorithms to determine p(x' |x):
    - Simulated Annealing (top down clustering)
    - Agglomerative (bottom up clustering)
- Difficult to implement in this case
  - Start with K~10^5; expect relevant clusters at K' ~10
  - Considerable degeneracy from *hapax legomena*

Conclusions

- Robust estimation of probability distribution functionals is possible even with undersampling
- Pairwise mutual information displays power law behavior, with breaks at semantically significant length scales

Possible Applications

- Semantic extraction
  - Automatic document summarizing
  - Search engines
- Data compression
  - Shannon compression theorem
  - Existing algorithms work at finite range only

For further reading

- Language
  - *Words and Rules: the ingredients of language*, by Steven Pinker
  - Nature 417:611
- Entropy estimation
  - Physical review E 52:6841
  - xxx.lanl.gov/abs/physics/0108025

End