Information Theory

CS114 Lab 2

Te Rutherford

Information Theory

- Developed by Shannon in the 40's to formalize the fundamental limits on data compression and transmission rate.
- In natural language, we do compress data (e.g. word length) and transmit data (speaking).

Concepts

- Entropy, joint entropy, and conditional entropy
- Mutual Information
- KL Divergence
- Cross entropy
- Perplexity

Entropy

- a measure of the uncertainty associated with a random variable (and its probability distribution function).
- Higher entropy implies …

  = Higher uncertainty (harder to predict)
  = More information content
  = More bits required to encode and communicate

Which one is harder to predict?

X = horse that wins the long distance race

- P(X=1) = 0.25
- P(X=2) = 0.25
- P(X=3) = 0.25
- P(X=4) = 0.25

Y = horse that wins the short distance race

- P(Y=1) = 0.5
- P(Y=2) = 0.3
- P(Y=3) = 0.1
- P(Y=4) = 0.1

Entropy

- Definition

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- What's the entropy of X?
- What's the entropy of Y?

Which one is harder to predict?

- P(X=1) = 0.25 00
- P(X=2) = 0.25 10
- P(X=3) = 0.25 01
- P(X=4) = 0.25 11

H(X) = 2 bits

- harder to predict
- more bits to code (on average)
- more information

- P(Y=1) = 0.5 0
- P(Y=2) = 0.3 10
- P(Y=3) = 0.1 111
- P(Y=4) = 0.1 110

H(Y) = 1.68 bits

- more predictable
- more frequent one gets a shorter code
- we can guess …

Joint Entropy

- The joint entropy of 2 RV X,Y is the number of bits needed on average to code both their values
- Same interpretation as univariate version.

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(X,Y)$$

Conditional Entropy

- The conditional entropy of a RV Y given another X, expresses how many extra bits required on average to communicate Y given that the other party knows X

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(y|x)$$

- What is H(Y|X) if X is a perfect predictor of Y?
- What is H(Y|X) if X is independent of Y?

Mutual Information

- How predictive is X of Y? and vice versa
- I(X,Y) is mutual information between X and Y

$$I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

(there are many other equivalent definitions)

- If X is a perfect predictor of Y, then
  $I(X,Y) = H(Y) - H(Y|X) = H(Y) - 0 = H(Y)$
- If X is independent of Y, then
  $I(X,Y) = H(Y) - H(Y|X) = H(Y) - H(Y) = 0$

Mutual Information in Natural Language

- If we want to predict the author's gender,
  - I(gender, talk about health or not)
  - I(gender, talk about sports or not)
  - I(gender, talk about cars or not)
- then we rank how predictive/relevant each topic to gender.

Pointwise Mutual Information

- In contrast, pointwise mutual information – often called as PMI – is defined for specific values of X and Y

$$PMI(x,y) = \log \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)}$$

- When computed for a pair of words, PMI can measure the semantic relatedness of two words
  e.g.
  PMI (“drink”, “beer”) > PMI (“drink”, “homework”)

Entropy of Natural Language

- How much information is there per word?
- How many bits do we need to communicate in English?

$$\mathrm{H}(\mathrm{X}) = -\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{p}(\mathrm{x})\log_2 \mathrm{p}(\mathrm{x})$$

$$\mathrm{H}(w_{1:n}) = -\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{p}(w_{1:n})\log_2 \mathrm{p}(w_{1:n})$$

- We don't know the true p. what should we do?

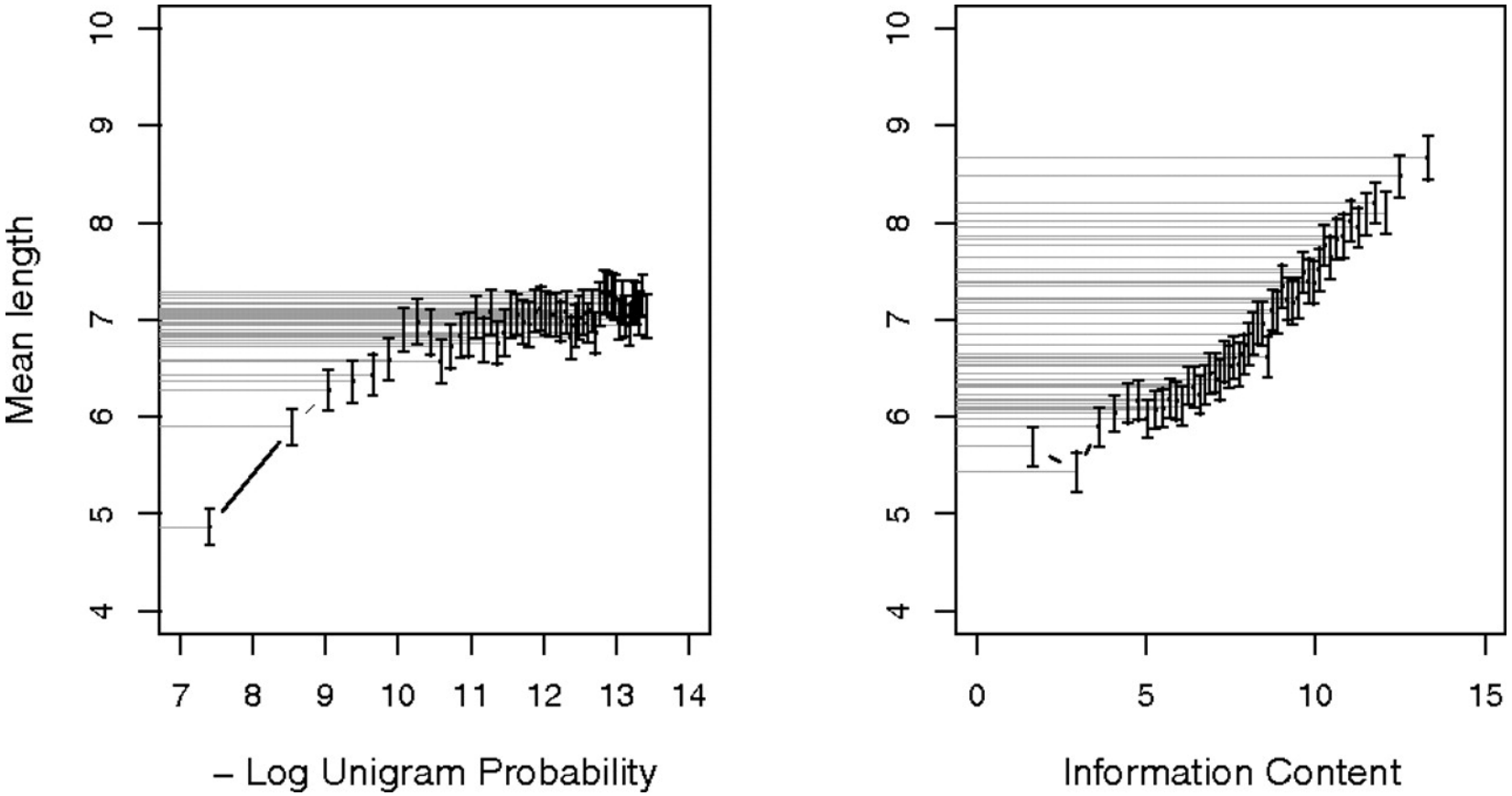
Cross Entropy

- Use the LM that we train instead and then compute the entropy on the test data

$$\mathrm{H}(w_{1:n}) = -\sum_{\mathrm{x}\in\mathrm{X}} \mathrm{p}(w_{1:n})\log_2 \mathrm{p}(w_{1:n})$$

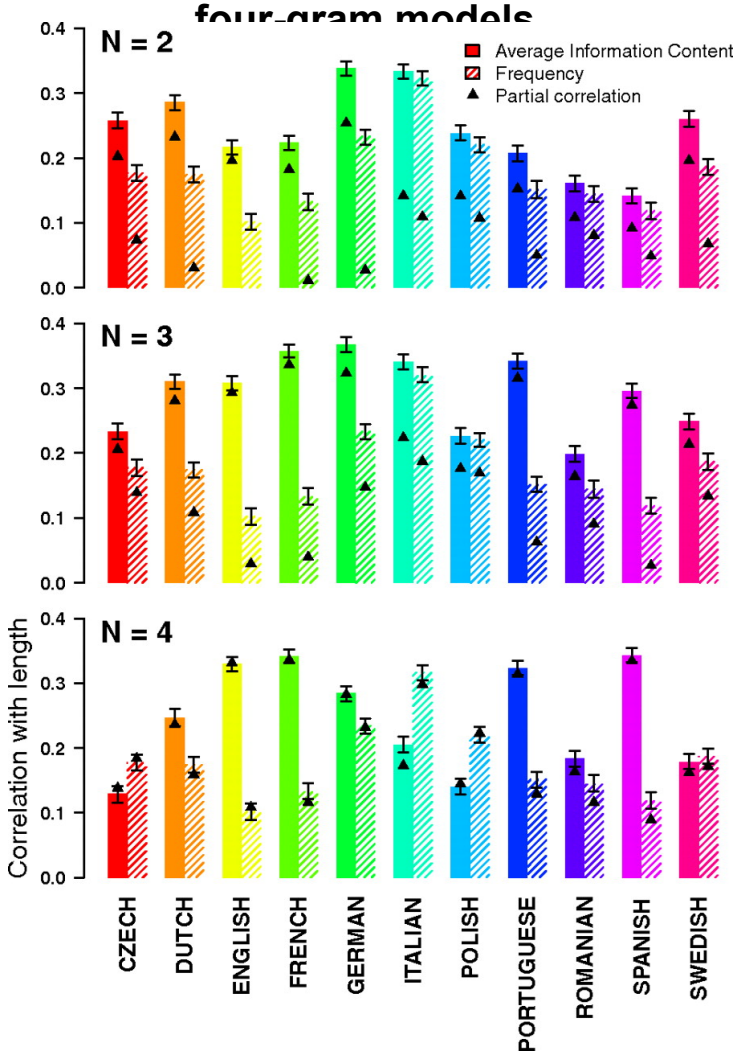
$$\mathrm{H}_{LM}(w_{1:n}) = -\sum_{i=1:n} \frac{1}{n} \log_2 \mathrm{p}_{LM}(w_i \mid w_{i-2}.w_{i-1})$$

Relationship between frequency (negative log unigram probability) and length, and information content and length.



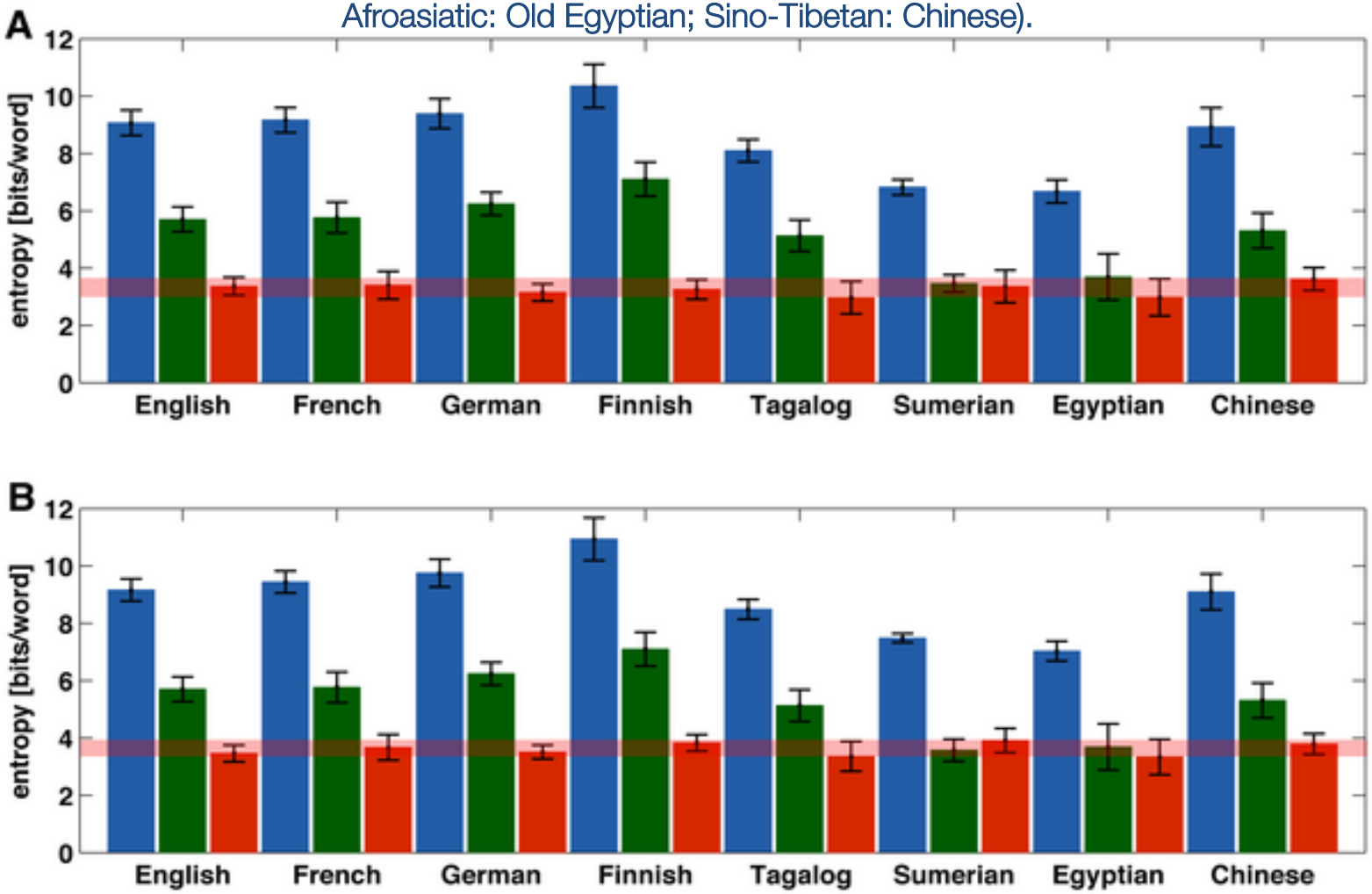
Piantadosi S T et al. PNAS 2011;108:3526-3529

Correlations between information content and word length (solid) and between frequency (negative log unigram probability) and word length (striped) for two-gram, three-gram, and four-gram models



Piantadosi S T et al. PNAS 2011;108:3526-3529

Figure 2. Entropy of eight languages belonging to five linguistic families and a language isolate (Indo-European: English, French, and German; Finno-Ugric: Finnish; Austronesian: Tagalog; Isolate: Sumerian; Afroasiatic: Old Egyptian; Sino-Tibetan: Chinese).



Montemurro MA, Zanette DH (2011) Universal Entropy of Word Ordering Across Linguistic Families. PLoS ONE 6(5): e19875. doi:10.1371/journal.pone.0019875
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0019875

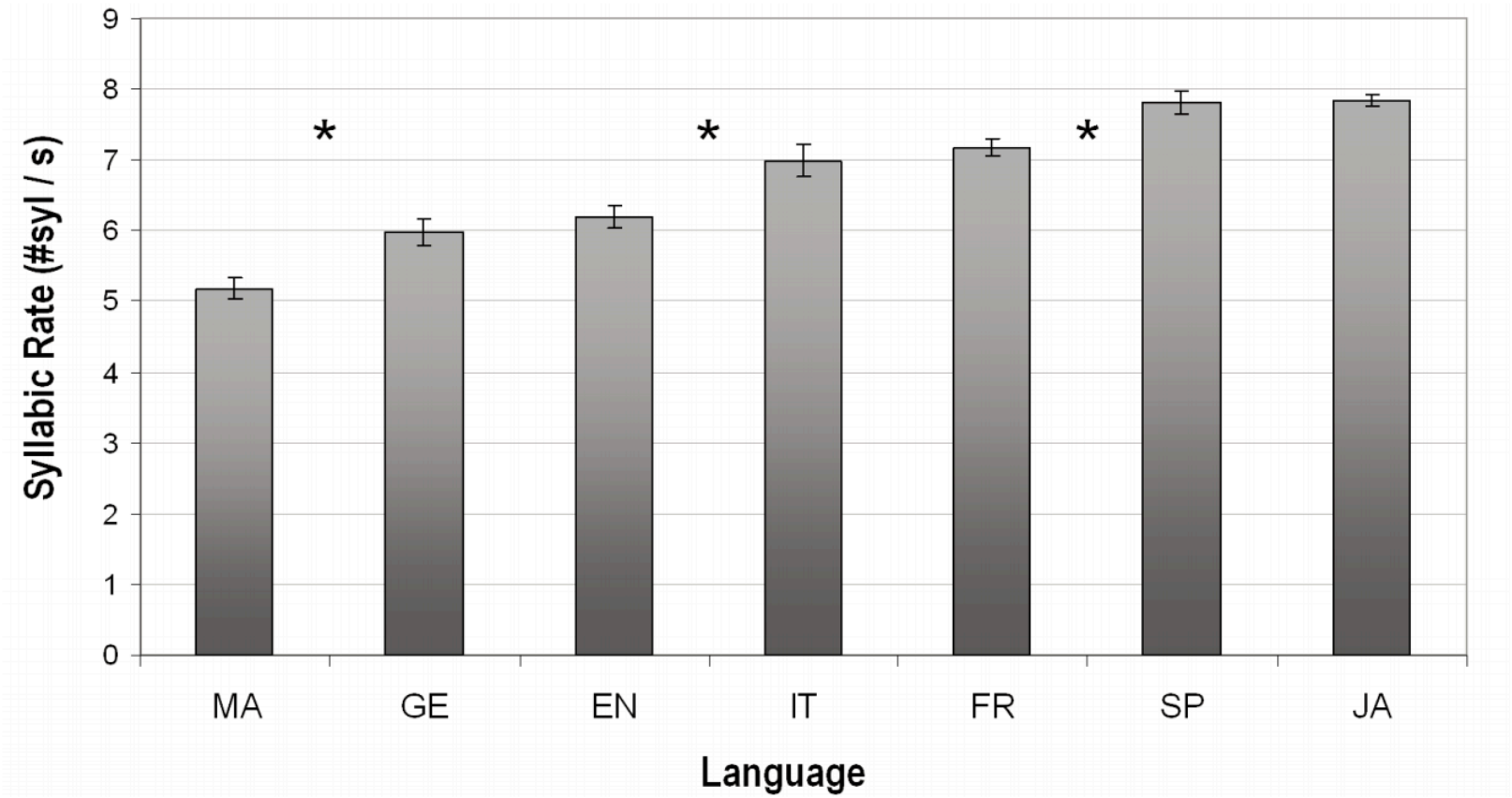


FIGURE 1. Speech rate measured in terms of the number of syllables per second (mean values and 95% confidence intervals). Stars indicate significant differences between the homogeneous subsets revealed by post-hoc analysis.

Pellegrino, 2011

LANGUAGE	INFORMATION DENSITY ID_L	SYLLABIC RATE (#syl/sec)	INFORMATION RATE
English	0.91 (± 0.04)	6.19 (± 0.16)	1.08 (± 0.08)
French	0.74 (± 0.04)	7.18 (± 0.12)	0.99 (± 0.09)
German	0.79 (± 0.03)	5.97 (± 0.19)	0.90 (± 0.07)
Italian	0.72 (± 0.04)	6.99 (± 0.23)	0.96 (± 0.10)
Japanese	0.49 (± 0.02)	7.84 (± 0.09)	0.74 (± 0.06)
Mandarin	0.94 (± 0.04)	5.18 (± 0.15)	0.94 (± 0.08)
Spanish	0.63 (± 0.02)	7.82 (± 0.16)	0.98 (± 0.07)
Vietnamese	1 (reference)	5.22 (± 0.08)	1 (reference)

TABLE 1. Cross-language comparison of information density, syllabic rate, and information rate (mean values and 95% confidence intervals). Vietnamese is used as the external reference.

Pellegrino, 2011

Cross Entropy to Evaluate LM

- Cross entropy measures the difference between the two distributions.
- = H(true LM) + difference(true LM, our LM)
- Cross entropy is the upper bound of the entropy, so it is higher than the true entropy.
  - Bad LM → high cross entropy
  - Good LM → low cross entropy
  - The best possible LM = the true entropy
  - no better than that.

Perplexity

- Perplexity is defined as

$$PP_{LM}(w_{1:n}) = P_{LM}(w_{1:n})^{-\frac{1}{n}} = \sqrt[n]{\frac{1}{P_{LM}(w_{1:n})}}$$

- which is essentially

$$PP_{LM}(w_{1:n}) = 2^{H_{LM}(w_{1:n})}$$

Homework 2

Assignment

1. Show that $PP_M(W) = 2^{H(W)}$ where $PP_M(W)$ is the perplexity of language model M on the sequence of n words W and $H(W)$ is the cross entropy of M on W. (Include the solution in the report)

2. Show that $PP_M(W) = \exp(-\frac{\log P_M(W)}{n})$ where P_M is the language model. (Include the solution in the report)