Large Scale Entropy Rate Estimation

A New Law that Governs the Complexity of Language

Łukasz Dębowski
ldebowsk@ipipan.waw.pl
(joint work with Ryosuke Takahira and Kumiko Tanaka-Ishii)

Instytut Podstaw Informatyki PAN
Warszawa

Cognitive Systems Modeling
5th Peripatetic Conference
Zakopane, 6–8 October 2016

What is the (Shannon) entropy rate?

An informal operational definition

Entropy rate h is the average number of yes/no questions that a person who knows the language needs to guess a letter of a text while knowing all the preceeding letters.

We have a stationary stochastic process $X_1, X_2, X_3, \ldots$.
For blocks $X_1^n = (X_1, \ldots, X_n)$ we define block entropy

$$H(X_1^n) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n)$$

and $h = \lim_{n\to\infty} H(X_1^n)/n = \lim_{n\to\infty} \left[H(X_1^n) - H(X_1^{n-1}) \right]$.

Entropy rate h measures the ultimate amount of
information=unpredictability=randomness
in text per unit symbol.

How to estimate the entropy rate?

A psycholinguistic approach

Take some human subjects and let them actually guess the consecutive letters of some carefully chosen text.

This above approach is quite costly and, because of large variation of estimates, it does not yield a very precise number.

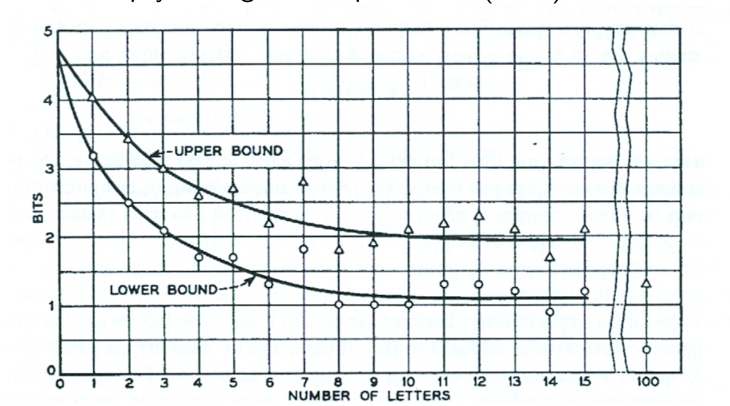
A computational approach

Write a universal computer program that tries to predict the the consecutive symbols of any stationary sequence of symbols and apply it to large corpora of texts (such as a few GB).

The above approach yields a very definite estimate of the entropy rate but some systematic error is buried in the assumption that a computer program can predict text as good as human subjects.

Some famous plot

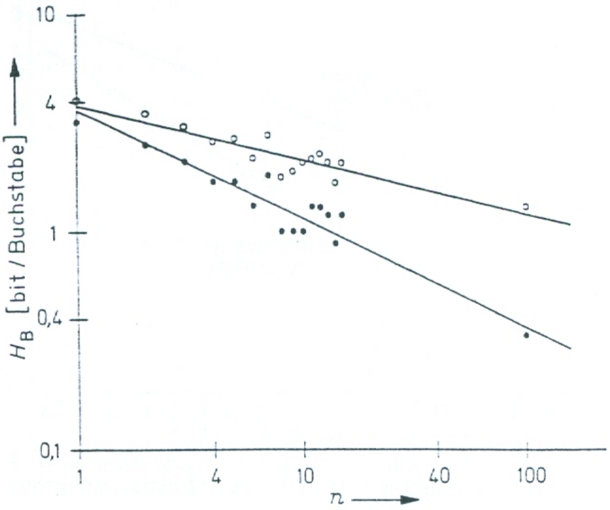
Shannon's psycholinguistic experiments (1951):



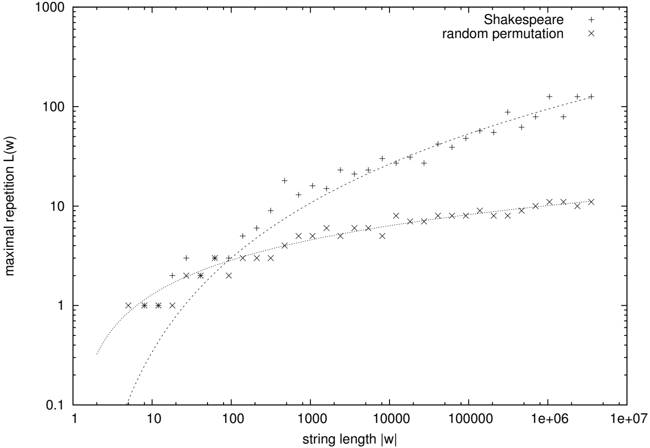
Many later studies claimed $h \approx 1$ bpc (bit per letter).

Wait, but isn't the entropy rate zero?

Hilberg (1990) replotted Shannon's figure in a log-log scale:



Some other evidence—Maximal repetition



This plot proves that conditional Renyi entropy rate is zero for natural language. Is the Shannon entropy rate also zero?

1. Entropy rate
2. Our experiment

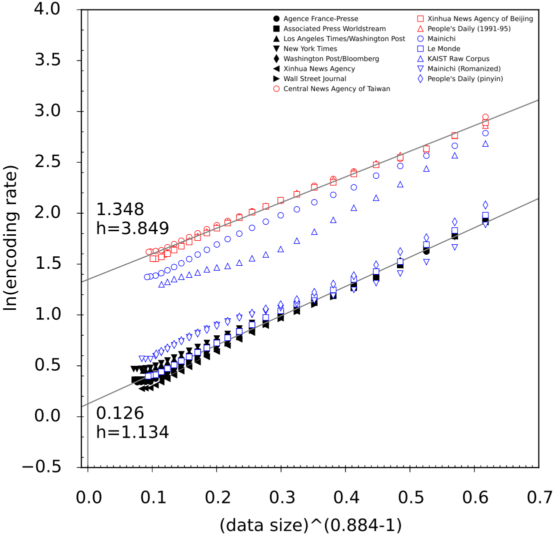
The experimental setup

- We want to clearly show for natural language that, while the conditional Renyi entropy rate is zero, the Shannon entropy rate is positive.
- For this, we need to estimate the Shannon entropy rate as precisely as possible.
- We chose to run a standard universal prediction procedure called PPM (prediction by partial match) on 20 corpora of up to 7.8 gigabytes across six languages (English, French, Russian, Korean, Chinese, and Japanese).
- We had to address the problem of slow convergence of the entropy rate estimates. We considered a careful extrapolation given by a carefully chosen ansatz function.

Ansatz functions

- Many works report only a single value of the encoding rate $\boldsymbol{h(n)}$ for the maximal size of the available data $\boldsymbol{n}$.
- Whereas any computation can handle only a finite amount of data, the entropy rate is the limit $\boldsymbol{h = \lim_{n\to\infty} h(n)}$.
- Since the probabilistic model of natural language is unknown, we have to extrapolate $\boldsymbol{h(n)}$ using an ansatz such as:
  - Hilberg (1990), Crutchfield and Feldman (2003) (power law):
  $$h_1(n) = An^{\beta-1} + h.$$
  - Ebeling and Nicolis (1991), Grassberger (1989):
  $$h_2(n) = An^{\beta-1} \ln n + h.$$
  - Our new proposal (stretched exponential):
  $$h_3(n) = \exp(An^{\beta-1} + h').$$

Our experimental data



Conclusions (1)

Baffled whether the entropy rate $\boldsymbol{h}$ of natural language is zero, we estimated $\boldsymbol{h}$ using the PPM method.

Compared to previous works:

- We calculated the encoding rates $\boldsymbol{h(n)}$ for six different languages by using much larger corpora (up to 7.8 gigabytes).
- We extrapolated the encoding rates $\boldsymbol{h(n)}$ to estimates of $\boldsymbol{h}$ using a novel ansatz (stretched exponential).

We obtain estimates of $\boldsymbol{h}$ which are 20% smaller than reported previously but positive, which falsifies a hypothesis by Hilberg.

Conclusions (2)

But there remains something true in Hilberg's ideas.
He seemed to suppose that all languages are equally hard to learn.

We can see that!

- Exponent β controls the speed of convergence of the encoding rate $\boldsymbol{h(n)}$ to the entropy rate $\boldsymbol{h}$.
- While entropy rate $\boldsymbol{h}$ measures how hard it is to predict texts, exponent β measures how hard it is to learn to predict texts.
- Whereas the entropy rate $\boldsymbol{h}$ strongly depends on the kind of the script, the exponent β turned out to be approximately constant, $\beta \approx \mathbf{0.884}$, across six languages.

`www.ipipan.waw.pl/~ldebowsk`