Three Power Laws Which Show That Language Is Not A Finite-State Process

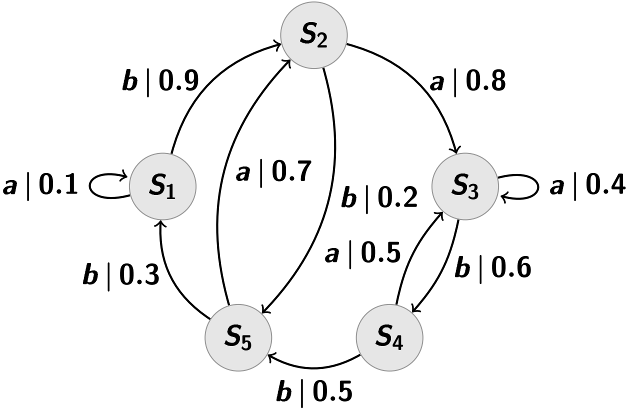
Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences
Warsaw

Cognitive Systems Modeling
Around information, information around
7th Peripatetic Conference
Małe Ciche, 18–21 October 2018

1. Finite-state processes
2. Three power laws
3. Conclusions

Finite-state automata and processes



A hidden Markov process with a finite number of hidden states S_i.

The HM order of a process is the min-number of hidden states.
Finite-state processes are processes of a finite HM order.

Is natural language a finite-state process?

- **Yes:**
  B. F. Skinner. *Verbal Behavior*. Prentice Hall, 1957.
  Skinner-like argument: Human brain consists of a billion of neurons (a finite number). Assuming that each neuron can be in two states, we obtain that the verbal behavior can be modeled by a finite-state automaton with $\mathbf{2^{10^9}}$ states.
- **No:**
  N. Chomsky. A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58, 1959.
  Chomsky-like argument: There appear nested utterances of structure $\boldsymbol{a^n b^n}$ in human language with $\boldsymbol{n}$ arbitrarily large. Hence the natural language cannot be modeled by a finite-state automaton and should be modeled at least by a context-free grammar (push-down automaton).

More convincing empirical evidence?

- Chomsky-like argument is based on our rational understanding of how natural langauge works.
- Observing structures $a^n b^n$ with n large enough is difficult.
- Is there another computational method of showing that natural language is not a finite-state process?
  - Can a mathematical theory (information theory and statistics) provide a method of showing that a given stream of data cannot be generated by a finite-state process?
  - Can we estimate the HM order of a process?
  - Can we apply these methods to human language corpus data?

Three power laws

We will demonstrate three information-theoretic power laws, probably satisfied by language, which disprove that language is a finite-state process with a small number of hidden states.

It is still possible that natural language is a finite-state process with a very large number of hidden states ($\mathbf{2^{10^9}}$ or more if we take into account interaction with the environment and other individuals).

Law 1: Block-wise mutual information

- Entropy: $H(X) = -\sum_x P(X = x) \log P(X = x)$
- Strings of letters: $X_j^k = (X_j, X_{j+1}, ..., X_k)$
- Mutual information:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \geq 0$$

- Bound for finite-state processes:

$$I(X_{-n}^0; X_1^n) \leq I(S_0; S_1) \leq H(S_1) \leq \log(\#\text{ of hidden states})$$

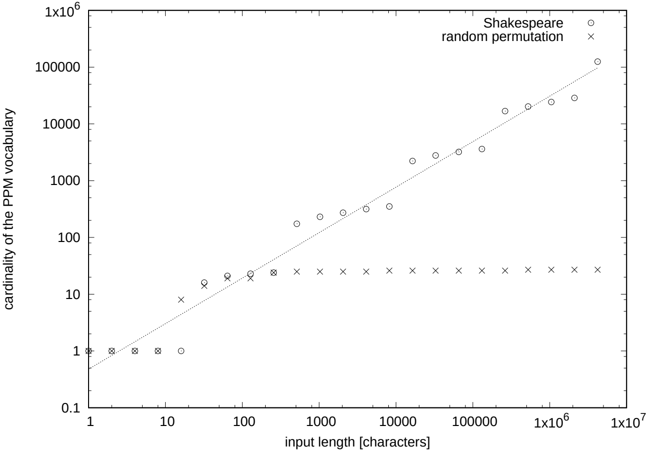
- Hilberg's (1990) hypothesis for natural language:

$$I(X_{-n}^0; X_1^n) \propto n^\beta, \quad \beta \approx 1/2$$

- Mutual information $I(X_{-n}^0; X_1^n)$ is hard to lower-bound.

The growth of PPM vocabulary

— an upper bound for mutual information $I(X_{-n}^{0}; X_{1}^{n})$



$$\ln \operatorname{card} V(x_1^n) \approx -0.737 + 0.801 \ln n$$ for Shakespeare

Law 2: Maximal repetition

- Strings of letters: $x_j^k = (x_j, x_{j+1}, ..., x_k)$
- Maximal repetition:

$$L(x_1^n) = \max \left\{ k : x_{i+1}^{i+k} = x_{j+1}^{j+k} \text{ for some } 0 \leq i < j \leq n - k \right\}$$

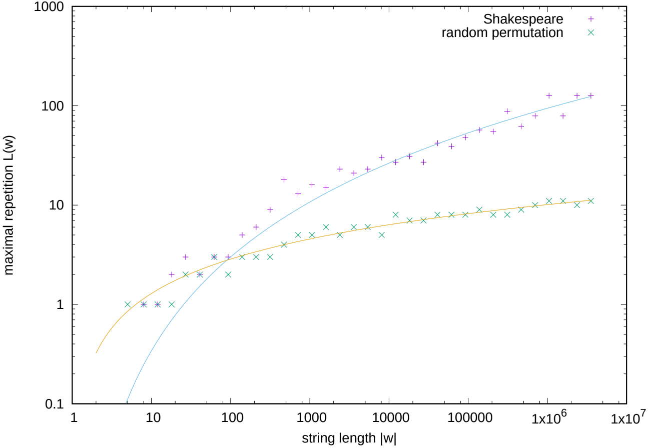
- Bound for typical finite-state processes:

$$L(X_1^n) \leq A \log n \text{ almost surely}$$

- Dębowski's (2012, 2015) observation for natural language:

$$L(x_1^n) \propto (\log n)^\alpha, \quad \alpha \approx 3$$

The growth of maximal repetition



- $L(x_1^n) \approx 0.02498 \, (\log n)^{3.136}$ for Shakespeare
- $L(x_1^n) \approx 0.4936 \, (\log n)^{1.150}$ for random permutation of chars

Law 3: Symbol-wise mutual information

- Entropy: $H(X) = -\sum_x P(X = x) \log P(X = x)$
- Mutual information:

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \geq 0$$

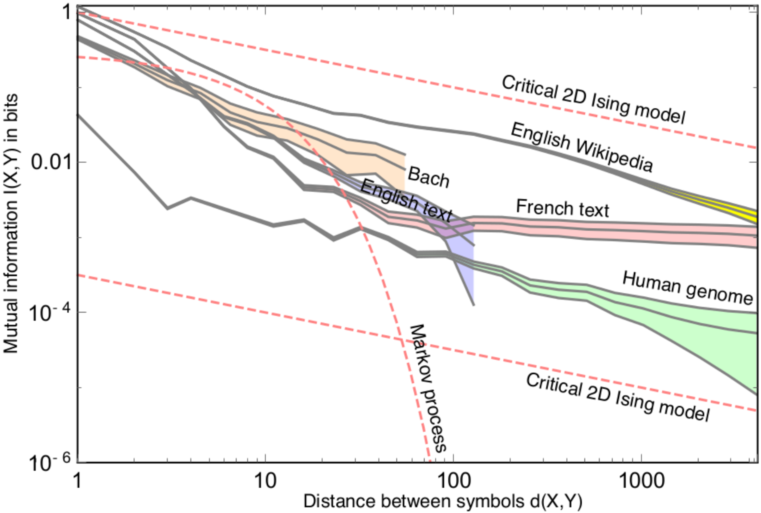
- Bound for typical finite-state processes:

$$I(X_1; X_n) \leq A\lambda^n, \quad \lambda < 1$$

- Lin and Tegmark's (2017) observation for natural language:

$$I(X_1; X_n) \propto n^{-\gamma}, \quad \gamma \approx 1/2$$

Lin and Tegmark's (2017) plot



Estimating the HM order of a process directly

We can also consistently estimate the number of hidden states of a given process (Lehéricy 2017), but the algorithm is quite complicated and the speed of its convergence is unknown.

The estimated number of hidden states can be
a quickly growing function of the text sample size.
It may be interesting to investigate this functional dependence.

Conclusions

- It has been long supposed that natural language cannot be a finite-state process with a small number of hidden states.
- We have shown three information-theoretic power laws, probably satisfied by natural language, which also disprove this hypothesis.
- Contrary to Chomskyan thought, rejecting finite-state processes does not mean eradicating any probability models from linguistic considerations.
- The world of stochastic processes is much richer than just finite-state processes — see my book in progress, entitled:

Ł. Dębowski, Information Theory Meets Power Laws: Stochastic Processes and Language Models, 2018±1.

References

N. Chomsky. A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58, 1959.

Dębowski, Ł. (2015). Maximal repetitions in written texts: Finite energy hypothesis vs. strong Hilberg conjecture. *Entropy*, 17:5903–5919.

Dębowski, Ł. (2018). Maximal repetition and zero entropy rate. *IEEE Trans. Inform. Theory*, 64(4):2212–2219.

Ebeling, W. and Nicolis, G. (1991). Entropy of symbolic sequences: the role of correlations. *Europhys. Lett.*, 14:191–196.

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248.

L. Lehéricy. Consistent order estimation for nonparametric Hidden Markov Models. `https://arxiv.org/abs/1606.00622v5`, 2017.

H. W. Lin and M. Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19:299, 2017.

B. F. Skinner. *Verbal Behavior*. Englewood Cliffs: Prentice Hall, 1957.