Natural Language Is Not A Finite-State Process: Evidence from Three Statistical Power Laws

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences
Warsaw

SFB 991 Seminar, Heinrich-Heine-Universität
Düsseldorf, 18th April 2019

My research path

1. MSc in theoretical physics

2. Statistical language modeling (engineering)

3. Power laws of natural language (quantitative linguistics)

4. Information theory (mathematics) $\Longleftrightarrow$ PhD & habilitation

1. Statistical language modeling
2. Is natural language a finite-state process?
3. Power laws of natural language
4. New kinds of stochastic processes
5. Conclusions

1. Statistical language modeling
2. Is natural language a finite-state process?
3. Power laws of natural language
4. New kinds of stochastic processes
5. Conclusions

Statistical language models

- A statistical language model is a probability measure on texts

$$p(x_1, .., x_n) = \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1}),$$

where x_i are consecutive letters or words of the text and the multipliers are conditional probabilities,

$$p(x_i|x_1, ..., x_{i-1}) \geq 0, \quad \sum_{x_i} p(x_i|x_1, ..., x_{i-1}) = 1.$$

- A statistical language model can be used to generate a random text by sampling the subsequent letter or word x_i with probability $p(x_i|x_1, ..., x_{i-1})$, where $x_1, ..., x_{i-1}$ is a part of text generated so far.

Cross entropy and model training

- Usually a statistical language model has a large number of free parameters $\theta_1, ..., \theta_k$,

$$p(x_i|x_1, ..., x_{i-1}) = p(x_i|x_1, ..., x_{i-1}, \theta_1, ..., \theta_k).$$

- Parameters θ_j are estimated by minimizing cross entropy of the model on some training data $y_1, ..., y_n$,

$$\frac{dH(\theta_1, ..., \theta_k)}{d\theta_j} = 0,$$

where cross entropy (= minus log likelihood) is

$$H(\theta_1, ..., \theta_k) := -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i|y_1, ..., y_{i-1}, \theta_1, ..., \theta_k).$$

Some simple classes of language models

- Markov processes or $(k+1)$-gram models:

$$p(x_1, .., x_n) = p(x_1, ..., x_k) \prod_{i=k+1}^{n} p(x_i | x_{i-k}, ..., x_{i-1}).$$

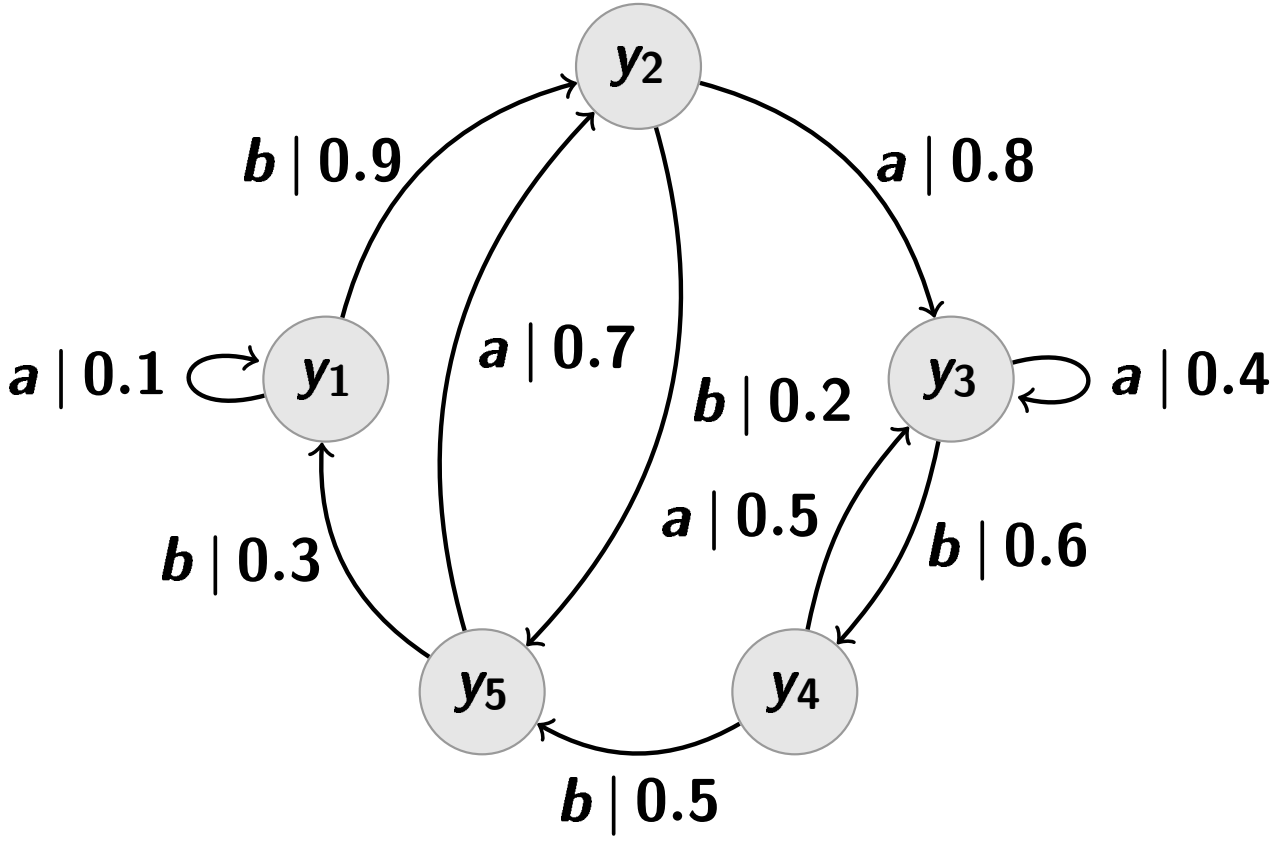
  The order of the Markov process equals k.

- Hidden Markov processes or finite-state models:

$$p(x_1, .., x_n) = \sum_{s_1, ... s_n \in \mathbb{S}} p(s_1) \prod_{i=2}^{n} p(x_i, s_i | s_{i-1}).$$

  Values $x_i \in \mathbb{X}$ — symbols. Values $s_i \in \mathbb{S}$ — states.
  The order of the hidden Markov process equals $|\mathbb{S}|$.

Finite-state models vs. finite-state automata



A hidden Markov process with a finite number of states.
Set of symbols $\mathbb{X} = \{a, b\}$. Set of states $\mathbb{S} = \{y_1, ..., y_5\}$.

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?

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.

We will NOT pursue this idea here.

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:

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

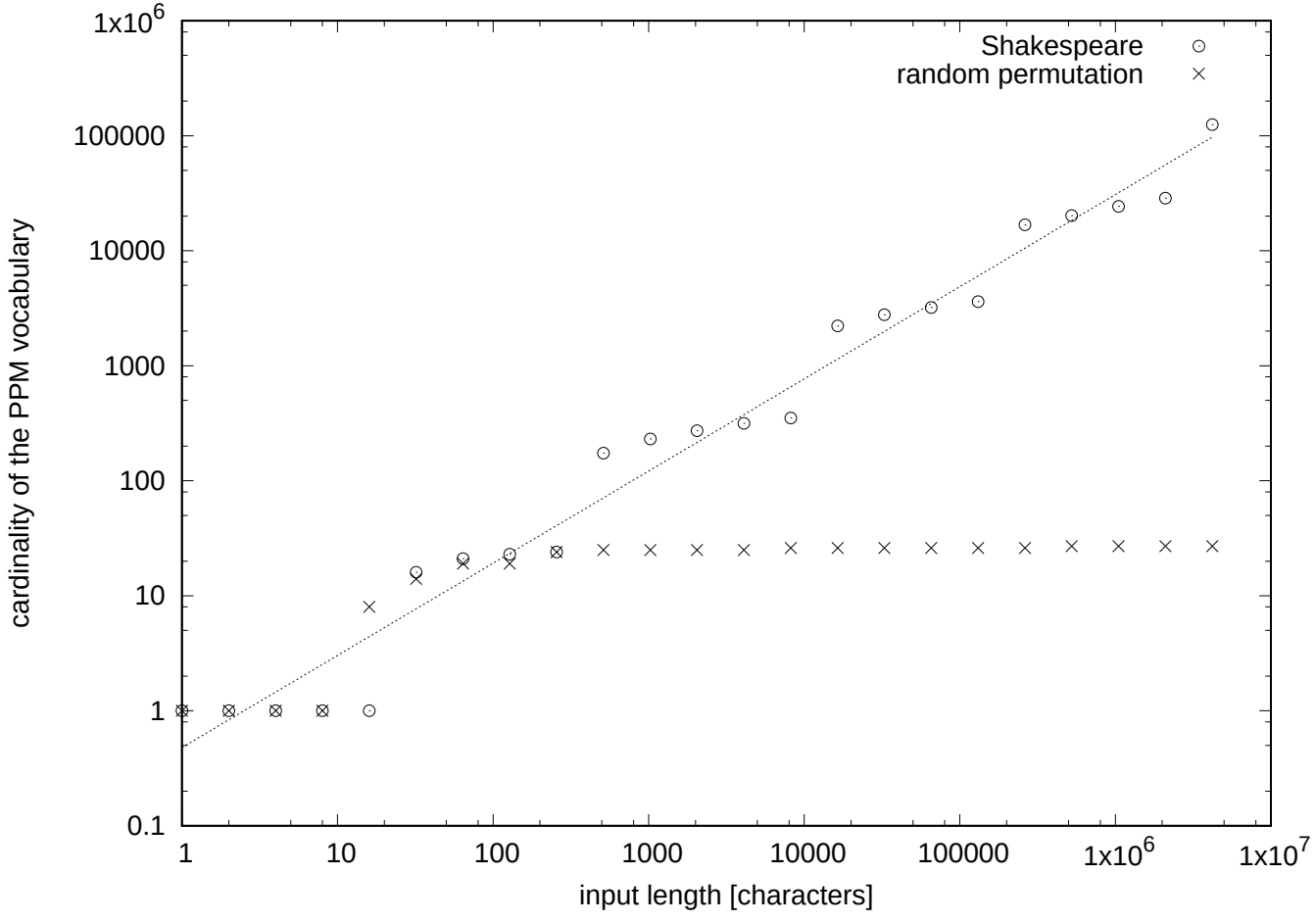
- Hilberg's (1990) hypothesis for natural language:

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

- 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:

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

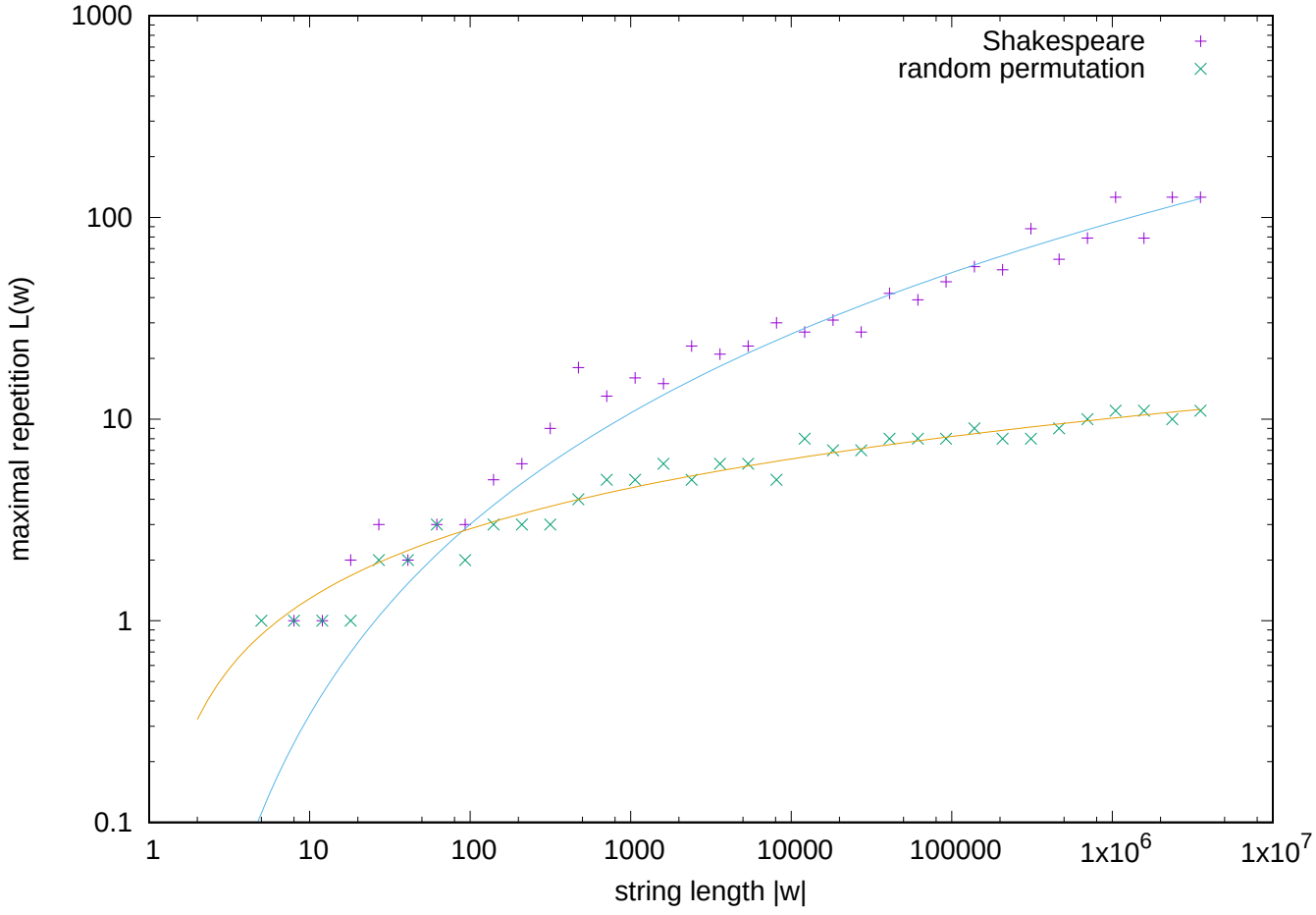
- Bound for typical finite-state processes:

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

- 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:

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

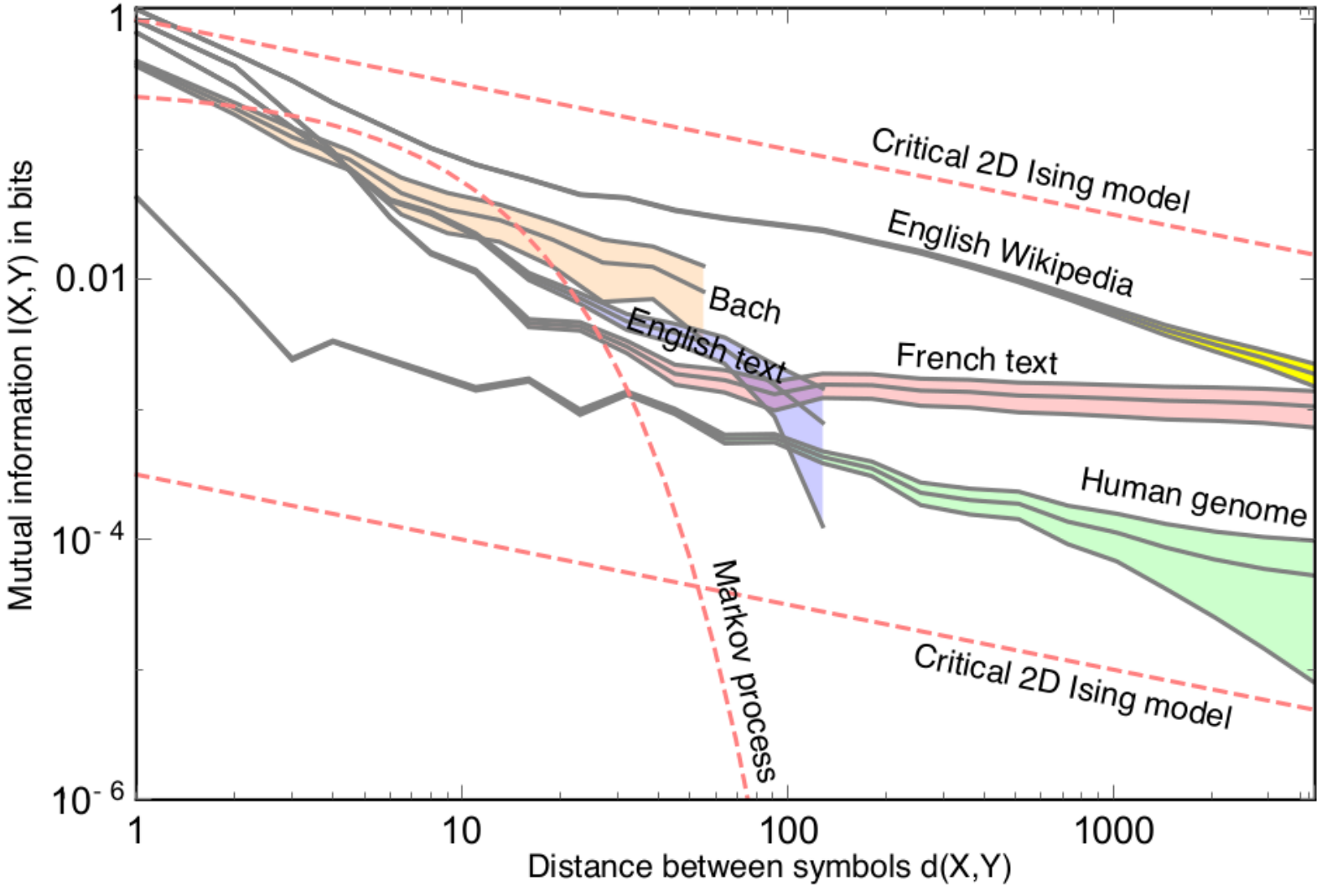
- Bound for typical finite-state processes:

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

- 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



Two new examples of processes

Seeking for mathematical examples of processes
that satisfy the mentioned three power laws of natural language,
I have constructed two classes of processes:

1. Santa Fe processes,
2. Random hierarchical association (RHA) processes.

Santa Fe process I: Motivation

- Each proposition X_i of a text in natural language might be represented as a pair $X_i = (k, z)$ which states that the k-th proposition in some enumeration assumes Boolean value z.
- Moreover, we may suppose that there is:
  — a stochastic process $(K_i)_{i \in \mathbb{Z}}$ (selection process),
  — a stochastic process $(Z_k)_{k \in \mathbb{N}}$ (object described by the text)
  such that if $X_i = (k, z)$ then $K_i = k$ and $Z_k = z$.
- We obtain a power-law growth of block mutual information for an appropriate choice of $(K_i)_{i \in \mathbb{Z}}$ and $(Z_k)_{k \in \mathbb{N}}$.

Santa Fe processes II: Definition

Let us put

$$X_i = (K_i, Z_{K_i}),$$

where

1. processes $(K_i)_{i\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$ are independent,
2. variables K_i are distributed according to a power law

$$P(K_i = k) \propto k^{-1/\beta}, \quad (K_i)_{i\in\mathbb{Z}} \sim \text{IID}, \quad \beta \in (0, 1),$$

3. process $(Z_k)_{k\in\mathbb{N}}$ an IID process with

$$P(Z_k = 0) = P(Z_k = 1) = 1/2.$$

Santa Fe processes III: Mutual information

Theorem

The mutual information for the Santa Fe process obeys

$$\lim_{n\to\infty} \frac{I(X_1^n; X_{n+1}^{2n})}{n^\beta} = \frac{(2-2^\beta)\Gamma(1-\beta)}{[\zeta(\beta^{-1})]^\beta}.$$

RHA processes I: Random selection of blocks

Let integers $(k_n)_{n\in\{0\}\cup\mathbb{N}}$, which we will call perplexities, satisfy

$$0 < k_{n-1} \leq k_n \leq k_{n-1}^2.$$

Next, for each $n \in \mathbb{N}$, let $(L_{nj}, R_{nj})_{j\in\{1,\ldots,k_n\}}$ be an independent random combination of k_n pairs of numbers from the set $\{1, \ldots, k_{n-1}\}$ drawn without repetition. That is,

$$P((L_{n1}, R_{n1}, \ldots, L_{nk_n}, R_{nk_n}) = (l_{n1}, r_{n1}, \ldots, l_{nk_n}, r_{nk_n})) = \binom{k_{n-1}^2}{k_n}^{-1}.$$

Subsequently we define random variables

$$Y_j^0 = j, \qquad j \in \{1, \ldots, k_0\},$$
$$Y_j^n = Y_{L_{nj}}^{n-1} \times Y_{R_{nj}}^{n-1}, \qquad j \in \{1, \ldots, k_n\}, n \in \mathbb{N},$$

where $a \times b$ denotes concatenation.

Hence Y_j^n are k_n distinct random blocks of 2^n numbers.

RHA processes II: Definition

Let $(C_n)_{n\in\{0\}\cup\mathbb{N}}$ be a sequence of independent random variables with uniform distribution

$$P(C_n = j) = 1/k_n, \qquad j \in \{1, \ldots, k_n\}.$$

Definition

The random hierarchical association (RHA) process $\mathcal{X}$ with perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ is defined as

$$\mathcal{X} = Y^0_{C_0} \times Y^1_{C_1} \times Y^2_{C_2} \times \ldots .$$

Sequence $\mathcal{X}$ will be parsed into a sequence of numbers X_j, where

$$\mathcal{X} = X_1 \times X_2 \times X_3 \times \ldots$$

RHA processes III: Entropy and maximal repetition

Theorem

For perplexities $k_n = \lfloor \exp\left(2^{\beta n}\right) \rfloor$, *where* $\beta \in (0,1)$, *the RHA process satisfies the following conditions:*

1. *The block entropy is sandwiched by*

$$C_1 m \left(\frac{1}{\log m}\right)^{1/\beta - 1} \leq H(X_1^m) \leq C_2 m \left(\frac{\log\log m}{\log m}\right)^{1/\beta - 1}$$

2. *The maximal repetition is sandwiched by*

$$C_3 (\log m)^{1/\beta} \leq L(X_1^m) \leq C_4 (\log m)^{1/\beta} \text{ almost surely}$$

Conclusions

- It has been long supposed that natural language cannot be a finite-state process with a small number of hidden states.
- We have exhibited 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:

Ł. Dębowski, Information Theory Meets Power Laws: Stochastic Processes and Language Models, 2020.

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.