Probability for Language Modeling

Part III: Learning

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Quantitative Cognitive Linguistics Network
5th September 2024

A (not so) rare specimen of a stochastic parrot

The rise of large language models
with their strengths
(fluent relevant grammatical replies)
and weaknesses
(factual hallucinations)
reopens the old question whether
statistical language modeling
makes sense for language science.

Motivating question

Does statistical language modeling make sense for linguistics?

I think one should replace "Does" with "How can".

There is a related question of a theoretical importance:

How much randomness is there in language and speech ...
... and, precisely, how does it interfere with structure?

This question cannot be answered without a certain understanding of mathematical models of randomness.

These slides provide an intro.

Neural scaling law and Hilberg's law

- Several recent large-scale computational experiments in statistical language modeling reported power-law tails of learning curves [Takahira et al., 2016, Hestness et al., 2017, Kaplan et al., 2020, Henighan et al., 2020, Hernandez et al., 2021, Tanaka-Ishii, 2021].
- This observation can be implied by Hilberg's law, a power-law growth of mutual information between increasing blocks of text [Hilberg, 1990, Crutchfield and Feldman, 2003].
- This power-law growth occurs for languages as diverse as English, French, Russian, Chinese, Korean, and Japanese.
- We observe a language-independent value of the power-law exponent: the mutual information between two blocks of length n is proportional to $n^{0.8}$ [Takahira et al., 2016, Tanaka-Ishii, 2021].

Theorem about facts and words

- We advertise a mathematical theory of Hilberg's law that we have been developing for several years. Most of our results were resumed in works [Dębowski, 2011, 2021a,b].
- The focal point is the theorem about facts and words:

The number of independent facts described in a finite text is roughly less than the number of distinct words used in this text.

- This theorem pertains to a general stationary process and it links ergodic decomposition with semantics and statistics.
- This result seems paradoxical since we might think that combining words we could express more independent facts.
- However, this theorem can be proved easily, by adopting quite natural definitions of facts and words.

Entropy rate and excess entropy

We write blocks of random variables: $X_j^k := (X_j, X_{j+1}, ..., X_k)$.

Let a finite alphabet $\mathbb{X} = \{1, 2, ..., D\}$.

Consider a stationary process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\mathbb{X}$.

We denote its entropy rate

$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{k\to\infty} H(X_i|X_{i-k}^{i-1}),$$

where:

- $H(X) := \mathbb{E}\left[-\log P(X)\right]$ is the entropy of X,
- $H(X|Y) := \mathbb{E}\left[-\log P(X|Y)\right]$ is the entropy of X given Y.

We will bound the sublinear excess entropy $H(X_1^n) - hn$.

The Santa Fe process

- A Santa Fe process is a stochastic process $(X_i)_{i\in\mathbb{Z}}$ where individual variables can be decomposed as pairs

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

  with two processes $(K_i)_{i\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$.
- The narration $(K_i)_{i\in\mathbb{Z}}$ consists of topics $K_i : \Omega \to \mathbb{N}$.
- The knowledge $(Z_k)_{k\in\mathbb{N}}$ consists of facts $Z_k : \Omega \to \{0, 1\}$.
- Process $(X_i)_{i\in\mathbb{Z}}$ is a simple model of a non-contradictory text: Whenever a certain topic is discussed again ($K_i = K_j$), the same fact is reported ($Z_{K_i} = Z_{K_j}$).
- That said, we may assume narration $(K_i)_{i\in\mathbb{Z}}$ and knowledge $(Z_k)_{k\in\mathbb{N}}$ to be pretty arbitrary processes and investigate consequences of our particular choices.

The number of described facts

- We say that a finite text x_1^n describes m initial facts by means of a function g if

$$m = U_g(x_1^n) := \min \{k \in \mathbb{N} : g(k, x_1^n) \neq Z_k\} - 1.$$

- Let knowledge $(Z_k)_{k\in\mathbb{N}}$ be a Bernoulli$(\frac{1}{2})$ process (fair coin).
- Let narration $(K_i)_{i\in\mathbb{Z}}$ be an IID process in natural numbers with Zipf's distribution $P(K_i = k) \sim k^{-\alpha}$, where $\alpha > 1$.
- Then for the Santa Fe process, putting $g(k, x_1^n) := z$ if $(k, z) \in x_1^n$ and $(k, 1 - z) \notin x_1^n$, whereas $g(k, x_1^n) := 2$ for other (k, x_1^n), we obtain a power law

$$\mathbb{E}\, U_g(X_1^n) \sim n^{1/\alpha}.$$

The number of described facts in general

- A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called strongly non-ergodic if the invariant σ-field $\mathcal{I}$ is non-atomic.
- Let $(Z_k)_{k\in\mathbb{N}}$ be an $\mathcal{I}$-measurable Bernoulli($\frac{1}{2}$) process.
- Variables Z_k are called facts since they don't depend on time.
- We say that a finite text x_1^n describes m initial facts by means of a function g if

$$m = U_g(x_1^n) := \min \{k \in \mathbb{N} : g(k, x_1^n) \neq Z_k\} - 1.$$

The number of described facts and excess entropy

- We denote $U_n := U_g(X_1^n)$. We observe

$$H(Z_1^{U_n}|U_n) = H(Z_1^{U_n}) - H(U_n),$$

where

$$H(U_n) \le 2\log(\mathbb{E}\, U_n + 2), \qquad \mathbb{E}\, U_n \le H(Z_1^{U_n}) \le H(X_1^n).$$

- Hence by the data-processing inequality,

$$I(X_1^n; Z_1^\infty) \ge I(X_1^n; Z_1^{U_n}|U_n) - H(U_n) = H(Z_1^{U_n}|U_n) - H(U_n).$$

- We have also an upper bound by the excess entropy

$$\begin{aligned} I(X_1^n; Z_1^\infty) &\le I(X_1^n; X_{n+1}^\infty) \\ &= H(X_1^n) - H(X_1^n|X_{n+1}^\infty) = H(X_1^n) - hn. \end{aligned}$$

- Thus the number of described facts bounds excess entropy

$$\mathbb{E}\, U_g(X_1^n) - 4\log(H(X_1^n) + 2) \le H(X_1^n) - hn.$$

Source coding

Let P be the probability measure of a stationary process $(X_i)_{i\in\mathbb{Z}}$.

Let Q be an incomplete measure: $\sum_{u\in\mathbb{X}^*} Q(u) \le 1$.

By Barron's inequality and the Shannon-McMillan-Breiman theorem, we obtain the lower bound

$$\liminf_{n\to\infty} \frac{[-\log Q(X_1^n)]}{n} \ge \lim_{n\to\infty} \frac{[-\log P(X_1^n)]}{n} = h \text{ a.s.}$$

if process $(X_i)_{i\in\mathbb{Z}}$ is ergodic. The analogous source coding inequality lower bounds the expectation

$$\liminf_{n\to\infty} \frac{\mathbb{E}\,[-\log Q(X_1^n)]}{n} \ge \lim_{n\to\infty} \frac{\mathbb{E}\,[-\log P(X_1^n)]}{n} = h$$

without the requirement of ergodicity.

Universal distributions

An incomplete measure Q is called universal if for any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over alphabet $\mathbb{X}$, we have

$$\lim_{n\to\infty} \frac{[-\log Q(X_1^n)]}{n} = h \text{ a.s.,}$$

$$\lim_{n\to\infty} \frac{\mathbb{E}\,[-\log Q(X_1^n)]}{n} = h.$$

Theorem (conditional universality criterion)

An incomplete measure Q is universal if for any $k \ge 1$, any conditional distribution $\tau : \mathbb{X} \times \mathbb{X}^k \to [0,1]$, and any $x_1^n \in \mathbb{X}^*$,

$$-\log Q(x_1^n) \le C(k,n) - \log \prod_{i=k+1}^{n} \tau(x_i|x_{i-k}^{i-1}),$$

where $\lim_{k\to\infty} \limsup_{n\to\infty} C(k,n)/n = 0$.

Maximum likelihood (ML)

We define the maximum likelihood (ML) in the class of k-th order Markov processes over alphabet $\mathbb{X} = \{1, 2, ..., D\}$ as

$$\hat{Q}(k|x_1^n) := \begin{cases} 1, & k \geq n, \\ \max_\tau \prod_{i=k+1}^n \tau(x_i|x_{i-k}^{i-1}), & k < n, \end{cases}$$

where the maximum is taken across all k-th order transition matrices $\tau : \mathbb{X} \times \mathbb{X}^k \to [0, 1]$.

The maximizing τ is called the maximum likelihood distribution for string x_1^n and denoted $\hat{\tau}(\cdot|x_1^n)$.

Empirical entropy

Let us write the frequency of string a_1^k in string x_1^n as

$$N(a_1^k|x_1^n) := \sum_{i=1}^{n-k+1} 1\left\{x_i^{i+k-1} = a_1^k\right\}.$$

Subsequently, let us denote the k-th order empirical entropy

$$\mathcal{H}(k|x_1^n) := \sum_{a_1^k} \frac{N(a_1^k|x_1^{n-1})}{n-k} \left[-\sum_{a_{k+1}} \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})} \log \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})} \right].$$

We have

$$\hat{\tau}(a_{k+1}|a_1^k, x_1^n) = \frac{N(a_1^{k+1}|x_1^n)}{N(a_1^k|x_1^{n-1})}, \quad -\log \hat{Q}(k|x_1^n) = (n-k)\mathcal{H}(k|x_1^n).$$

Penalized maximum likelihood (PML)

Consider the subword complexity

$$V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \le i \le n-k\right\} \le \min\left\{D^k, n-k+1\right\}.$$

We define the penalized maximum likelihood (PML)

$$Q(k|x_1^n) := \frac{\hat{Q}(k|x_1^n)}{Z(k|x_1^n)}, \quad Z(k|x_1^n) := D^k(n-k+1)^{V(k+1|x_1^n)+1},$$

$$Q(x_1^n) := w_n \max_{k\ge 0} w_k Q(x_1^n|k), \quad w_k := \frac{1}{k+1} - \frac{1}{k+2}.$$

Theorem

The penalized maximum likelihood Q is an incomplete measure and it satisfies the conditional universality criterion.

Markov order estimation

Let $(X_i)_{i\in\mathbb{N}}$ be stationary ergodic over $\mathbb{X} = \{1, 2, ...D\}$.

The Markov order of the process is defined as

$$M := \inf \left\{ k \geq 0 : H(X_i | X_{i-k}^{i-1}) = h \right\}.$$

In the above, IID processes are 0-th order Markov processes.

The Markov order estimator is defined as

$$M(x_1^n) := \inf \left\{ k \geq 0 : \hat{Q}(x_1^n | k) \geq Q(x_1^n) \right\}.$$

For $M \in [0, \infty]$, we have consistent estimation

$$\lim_{n\to\infty} M(X_1^n) = M \text{ a.s.},$$
$$\lim_{n\to\infty} \mathbb{E}\, M(X_1^n) = M.$$

The number of Markov subwords and the PML MI

- Let us denote the PML entropy

$$K(u) := -\log Q(u)$$

and the PML mutual information (PML MI)

$$J(u,v) := K(u) + K(v) - K(u,v).$$

- The number of Markov subwords is

$$V(x_1^n) := V(M(x_1^n) + 1|x_1^n).$$

- Since $M(x_1^n)K(x_1^n) \le n \log n$, we may bound the PML MI

$$J(X_1^n; X_{n+1}^{2n}) \le 2\left(V(X_1^{2n}) + \frac{2n \log D}{K(X_1^{2n})} + 3\right) \log(2n+2).$$

The telescope sum for excess entropy

Theorem

For a function $K : \mathbb{N} \to \mathbb{R}$*, define* $J(n) := 2K(n) - K(2n)$*. If there exists limit* $\lim_{n\to\infty} K(n)/n = h$ *then*

$$\sum_{k=0}^{\infty} \frac{J(2^k n)}{2^{k+1}} = K(n) - nh.$$

Proof.

We have the telescope sum

$$\sum_{k=0}^{m-1} \frac{J(2^k n)}{2^{k+1}} = K(n) - n \cdot \frac{K(2^m n)}{2^m n}.$$

For m tending to infinity, the above equality implies the claim. □

Almost the main theorem

Chaining the received inequalities yields

$$\begin{aligned}
&\mathbb{E}\, U_g(X_1^n) - 4\log\left(n\log D + 2\right) \le \mathbb{E}\, U_g(X_1^n) - 4\log\left(H(X_1^n) + 2\right) \\
&\le H(X_1^n) - hn \le \mathbb{E}\, K(X_1^n) - hn = \frac{1}{2}\sum_{k=0}^{\infty} 2^{-k}\, \mathbb{E}\, J(X_1^{2^k n}; X_{2^k n+1}^{2^{k+1} n}) \\
&\le 2\sum_{k=1}^{\infty} 2^{-k}\, \mathbb{E}\left(V(X_1^{2^k n}) + \frac{2^k n\log D}{K(X_1^{2^k n})} + 3\right)(k + \log(n+1)).
\end{aligned}$$

We will simplify the last expression using a power-law upper bound and the sums of infinite series

$$\sum_{k=1}^{\infty} z^k = \frac{z}{1-z}, \qquad \sum_{k=1}^{\infty} k z^k = \frac{z}{(1-z)^2}.$$

Theorem about facts and words

In this way, we obtain the finitary theorem about facts and words

$$\mathbb{E}\, U_g(X_1^n) \le 2 \left(\frac{2^{\beta_n}}{2 - 2^{\beta_n}} \mathbb{E}\, V(X_1^n) + \gamma_n + 5 \right) \left(\log(n \log D) + \frac{3}{2 - 2^{\beta_n}} \right),$$

where

$$\beta_n := \sup_{r>n} \log \left(\frac{\mathbb{E}\, V(X_1^r)}{\mathbb{E}\, V(X_1^n)} \right) / \log \left(\frac{r}{n} \right), \qquad \gamma_n := \sup_{r>n} \mathbb{E} \left(\frac{r \log D}{K(X_1^r)} \right).$$

We have $\mathbb{E}\, Y^{-1} \le \frac{1}{\mathbb{E}\, Y}(\alpha + \frac{\alpha^2 \operatorname{Var} Y}{(\alpha-1)^2 \mathbb{E}\, Y})$ if $Y \ge 1$ (by Paley-Zygmund).

The number of independent facts described in a finite text is roughly less than the number of distinct words used in this text.

- For $\beta_n = 0.8$, $D = 27$, and $\gamma_n = \log 27$, we obtain

$$\mathbb{E}\, U_g(X_1^n) \le (13.45\, \mathbb{E}\, V(X_1^n) + 19.51)(\log n + 13.84).$$

- For $\beta_n = 0.7$, $D = 27$, and $\gamma_n = \log 27$, we obtain

$$\mathbb{E}\, U_g(X_1^n) \le (8.652\, \mathbb{E}\, V(X_1^n) + 19.51)(\log n + 10.24).$$

Combining this bound with the trivial bound

But we also have a trivial bound

$$\mathbb{E}\, U_g(X_1^n) \le H(X_1^n) \le n \log D.$$

In particular:

- For $\mathbb{E}\, V(X_1^n) = n^{0.8}$, $D = 27$, and $\gamma_n = \log 27$, we have

  $$\mathbb{E}\, U_g(X_1^n) \le \min \left\{ 4.75n, (13.45n^{0.8} + 19.51)(\log n + 13.84) \right\}.$$

  The regime of the bound changes for $n = 5.41 \cdot 10^{10}$.
- For $\mathbb{E}\, V(X_1^n) = n^{0.7}$, $D = 27$, and $\gamma_n = \log 27$, we have

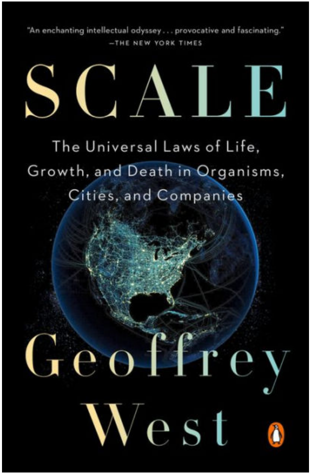
  $$\mathbb{E}\, U_g(X_1^n) \le \min \left\{ 4.75n, (8.652n^{0.7} + 19.51)(\log n + 10.24) \right\}.$$

  The regime of the bound changes for $n = 2.44 \cdot 10^{9}$.

The life expectancy of a human is around $4 \cdot 10^9$ heart beats.

A human should memorize everything, the posterity will verify it?

Allometric laws are everywhere!



Is the Hilberg exponent closer to 3/4 (biology) or 4/5 (economy)?

Further reading I

J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.

Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Inform. Theory*, 57:4589–4599, 2011.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. Wiley & Sons, 2021a.

Ł. Dębowski. A refutation of finite-state language models through Zipf's law for factual knowledge. *Entropy*, 23:1148, 2021b.

Ł. Dębowski. A short course in universal coding. Book manuscript in preparation, 2024.

T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. `https://arxiv.org/abs/2010.14701`, 2020.

D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. `https://arxiv.org/abs/2102.01293`, 2021.

Further reading II

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. https://arxiv.org/abs/1712.00409, 2017.

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

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361, 2020.

R. Takahira, K. Tanaka-Ishii, and Ł. Dębowski. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364, 2016.

K. Tanaka-Ishii. *Statistical Universals of Language: Mathematical Chance vs. Human Choice*. Springer, 2021.

G. West. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. New York: Penguin Press, 2017.