# Hilberg's Conjecture: a Challenge for Machine Learning

Łukasz Dębowski
ldebowsk@ipipan.waw.pl


Institute of Computer Science
Polish Academy of Sciences

# Probabilistic model of texts?

May generation of texts in natural language be described by a probabilistic model?

- This interdisciplinary question inspired a few concepts in applied mathematics:
  - Markov chains (Markov),
  - entropy (Shannon),
  - fractals (Mandelbrot),
  - algorithmic complexity (Kolmogorov).
- Common intuition: Texts are a result of a process that is neither purely deterministic nor purely random (Zipf).
- Some empirical statistical laws of language: Zipf-Mandelbrot's law, Herdan's law, Menzerath's law.
- Some well-defined stochastic process may describe the text generation in spite of great problems with its identification.

# Practical importance of language modeling

- Statistical language modeling is highly relevant for practical applications (e.g. speech recognition, machine translation).
- Bayes theorem in speech recognition:
  - **P(A|W)**—probability of speech **A** corresponding to text **W**,
  - **P(W)**—probability of text **W**.

$$\max_{\mathbf{W}} \mathbf{P(W|A)} = \max_{\mathbf{W}} \mathbf{P(A|W)P(W)}.$$

- Quality of speech recognition system depends on both models **P(A|W)** and **P(W)**.
- State-of-the-art modeling of **P(W)** consists in approximating the hypothetical stochastic process by Markov chains

— far from optimal.

# Need for fundamental research?

- Can theoretical research provide some insight into the practical task of statistical language modeling?
- We suppose that randomness of texts is constrained by the existence or the search for meaning...
- ...but the meaning itself may manifest as a form of apparent randomness. (Just recall halting probability, the number $\mathbf{\Omega}$, which is apparently random but stores an infinite amount of mathematical knowledge)
- Let's not lose hope that we may get some theoretical insight into probabilistic modeling of texts. We have both to look at empirical data and to build idealized mathematical models.

There seems to be a fundamental property of natural language, called Hilberg's conjecture, which can model some observations.

# Stochastic processes

- Probability space: $(\Omega, \mathcal{J}, \mathbf{P})$.
- Alphabet: $\mathbb{X}$.
- Random variables: $\mathbf{X_i : \Omega \to \mathbb{X}}$.
- Stochastic process: $\mathbf{(X_i)_{i \in \mathbb{Z}}}$.
- Blocks: $\mathbf{X_k^l = (X_i)_{k \le i \le l}}$.

Process $\mathbf{(X_i)_{i \in \mathbb{Z}}}$ is stationary $\overset{\mathbf{def}}{\iff}$ $\mathbf{P(X_{i+1}^{i+n})}$ doesn't depend on $\mathbf{i}$.

# Entropy

- Entropy of a random variable:

$$\mathbf{H}(\mathbf{X}_{\mathbf{k}}^{\mathbf{l}}) := \mathbb{E}\left[-\log \mathbf{P}(\mathbf{X}_{\mathbf{k}}^{\mathbf{l}})\right].$$

- It measures uncertainty of a random variable.
  - $\mathbf{H}(\mathbf{X}_1^{\mathbf{n}}) = \mathbf{n} \log \operatorname{card} \mathbb{X}$ — all values are equally probable.
  - $\mathbf{H}(\mathbf{X}_1^{\mathbf{n}}) = \mathbf{0}$ — the random variable is almost surely constant.
- Block entropy of a stationary process:

$$\mathbf{H}(\mathbf{n}) := \mathbf{H}(\mathbf{X}_{\mathbf{i}+1}^{\mathbf{i}+\mathbf{n}}).$$

- Entropy rate of a stationary process:

$$\mathbf{h} = \lim_{\mathbf{n}\to\infty} \mathbf{H}(\mathbf{n})/\mathbf{n}.$$

# Hilberg's conjecture

- Shannon (1951) estimated entropy of text in English, assuming it is drawn from a stationary process.
- Hilberg (1990) replotted these estimates in the log-log scale and observed a straightish line.
- This line corresponds to

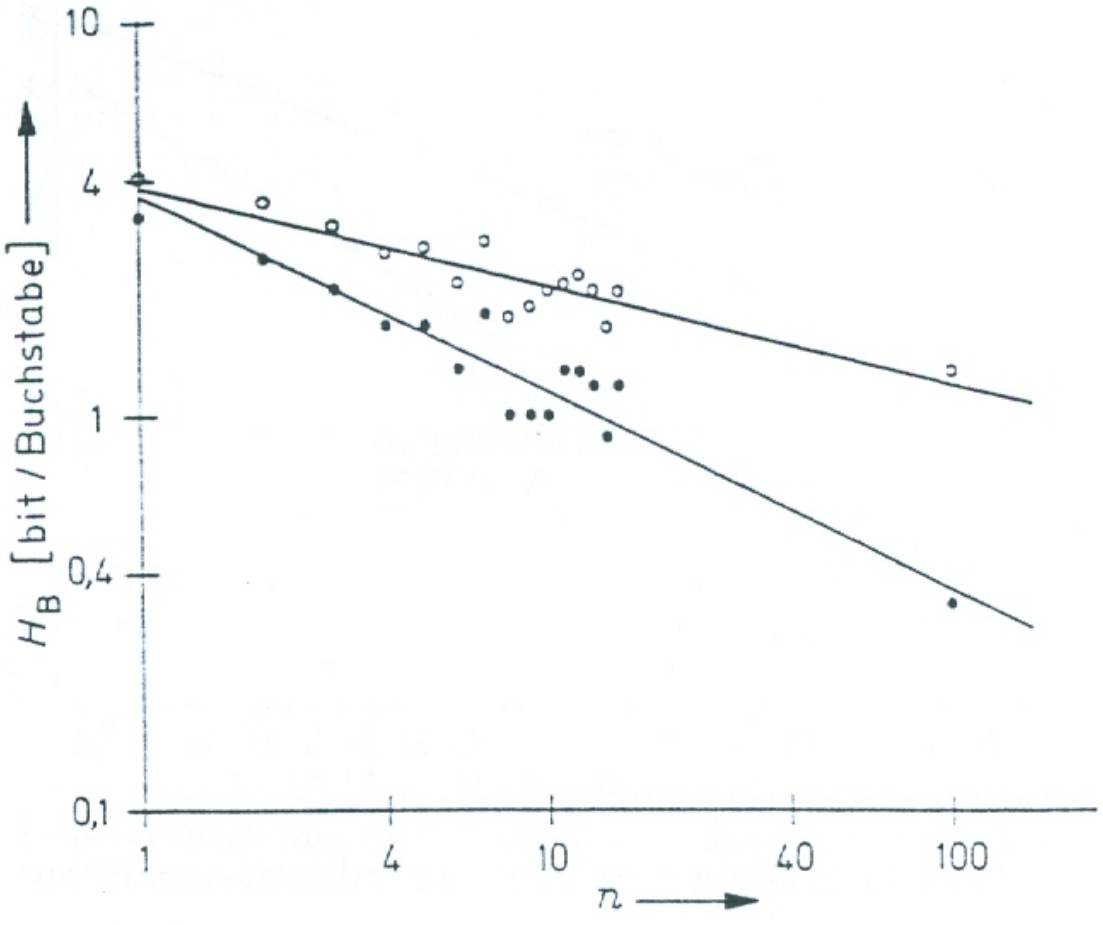
$$\mathbf{H(n) \approx Bn^{\beta} + hn}, \tag{1}$$

where $\beta \approx \mathbf{0.5}$ and $\mathbf{h \approx 0}$.

- Shannon provided estimates of $\mathbf{H(n)}$ for $\mathbf{n \leq 100}$ characters.
- Hilberg supposed that relationship (1) can be extrapolated and $\mathbf{h = 0}$ holds asymptotically.

# Shannon's data in log-log scale (Hilberg 1990)

Conditional entropy $\mathbf{H(n+1) - H(n)}$ vs. context length $\mathbf{n}$:



# The original vs. the relaxed Hilberg conjecture

Process $(\mathbf{X_i})_{i\in\mathbb{Z}}$ is asymptotically deterministic $\overset{\text{def}}{\iff}$ $\mathbf{h = 0}$.
(Each random variable is a function of infinite past.)

Consider mutual information

$$\mathbf{I(X;Y) = H(X) + H(Y) - H(X,Y).}$$

The original Hilberg conjecture is

$$\mathbf{H(n) \propto n^{\beta}}, \tag{2}$$

whereas the relaxed Hilberg conjecture is

$$\mathbf{I(X_1^n; X_{n+1}^{2n}) = 2H(n) - H(2n) \propto n^{\beta}.} \tag{3}$$

Relationship (3) follows from (2) but it does not imply $\mathbf{h = 0}$.

# Why is Hilberg's conjecture (HC) important?

- HC corroborates Zipf's insight that texts produced by humans diverge from both pure randomness and pure determinism. (In a sense, they would be both random and deterministic.)
- Relaxed HC also distinguishes natural language from **k**-parameter sources. (Some basic model of statistics.)
- HC, in its original form, implies that texts are in a sense deterministic and infinitely compressible. (We have to explain why modern text compressors cannot achieve that.)
- HC can be linked with Zipf's law and Herdan's law. (These are celebrated laws of quantitative linguistics.)
- Stochastic processes that satisfy HC are mathematical terra incognita. (Understanding their construction and properties can lead to a progress both in mathematics and applications like computational linguistics and machine learning.)

# A skeptic's remark

- Hilberg's conjecture in its original form implies that a typical text of one million letters could be theoretically compressed into a string of roughly one thousand letters. This is far beyond the power of any known text compressor!
- How is it possible? What blocks the optimal compression? How does the optimal compression look like?
- Some idea: Modern text compressors work mostly by detecting repeated strings and replacing them with shorter identifiers. They cannot compress texts beyond the maximal repetition.
- Another idea: Giving the ISBN number is sufficient to identify a printed literary text that remains in cultural circulation. Thus given enough memory, "hypercompression" is achievable.
- Is something similar possible in the world of stationary stochastic processes? We suppose that it is.

# Maximal repetition

## Definition

The maximal repetition in text $\mathbf{w}$ is defined as

$$\mathbf{L(w) := \max \{|s| : w = x_1 s y_1 = x_2 s y_2 \text{ and } x_1 \neq x_2\}},$$

where $\mathbf{s}$, $\mathbf{x_i}$, and $\mathbf{y_i}$ are substrings of text $\mathbf{w}$.

# Maximal repetition and Hilberg's conjecture

## Definition

For a random variable $\mathbf{X}$, topological entropy is

$$\mathbf{H_{top}(X) = \log \operatorname{card} \{x : P(X = x) > 0\}}.$$

## Theorem

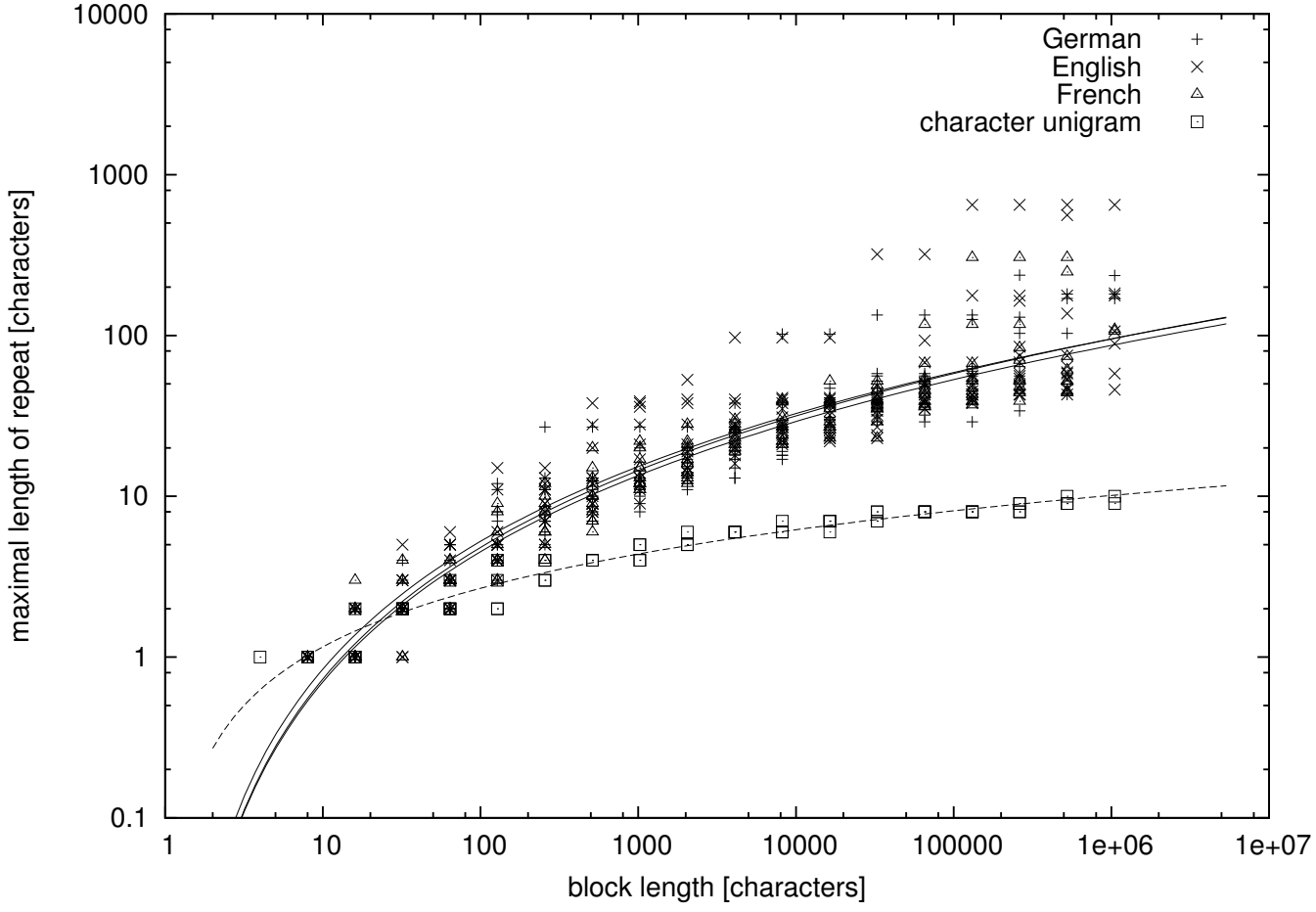
*If a stationary stochastic process* $\mathbf{(X_i)_{i\in\mathbb{Z}}}$ *satisfies*

$$\mathbf{H_{\mathit{top}}(X_{i+1}^{i+n}) \leq Bn^{\beta}}$$

*for certain constants* $\mathbf{0 < \beta < 1}$ *and* $\mathbf{B > 0}$ *then there exists* $\mathbf{A > 0}$ *such that for* $\alpha = 1/\beta$ *almost surely we have*

$$\mathbf{L(X_1^m) \geq A(\log m)^{\alpha}}.$$

# 35 texts in 3 languages + unigram text




**A ≈ 0.093** $\alpha \approx 2.64$

# Regular Hilberg processes

## Definition

A stationary process $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ is called a regular Hilberg process if

$$\mathbf{H(n)} = \mathbf{\Theta(n^{\beta})},$$
$$\mathbb{E}\,\mathbf{L(X_1^n)} = \mathbf{\Theta\left((\log n)^{\alpha}\right)}$$

for a certain $\beta \in (0,1)$ and $\alpha \geq 1/\beta$.

# Bound for the Lempel-Ziv code

- The Lempel-Ziv (LZ) code is the oldest known universal code.

**Theorem**

*The length of the LZ code satisfies*

$$|\mathbf{C(w)}| \geq \frac{|\mathbf{w}|}{\mathbf{L(w)}+1} \log \frac{|\mathbf{w}|}{\mathbf{L(w)}+1}.$$

- Similar bounds hold for a few other known universal codes.
- Hence, for regular Hilberg processes, the length of the LZ code is orders of magnitude larger than the block entropy,

$$\mathbf{H(n)} = \Theta(\mathbf{n}^{\beta}), \qquad \mathbb{E}\,\left|\mathbf{C}(\mathbf{X}_1^{\mathbf{n}})\right| = \Omega\left(\frac{\mathbf{n}}{(\log \mathbf{n})^{\alpha}}\right).$$

Hilberg's conjecture constitutes a challenge for machine learning.

# Herdan's law (an integrated version of Zipf's law)

- Consider texts in a natural language (such as English):
  - $\mathbf{V}$ — the number of different words in the text,
  - $\mathbf{n}$ — the length of the text.
- We observe Herdan's law, i.e., the relationship

$$\mathbf{V} \propto \mathbf{n}^{\gamma},$$

  where $\gamma$ is between $\mathbf{0.5}$ a $\mathbf{1}$ depending on a text.
- We will show that Hilberg's conjecture implies Herdan's law.

# A context-free grammar that generates one text

$$\left\{\begin{array}{l} \mathbf{A_1 \to A_2 A_2 A_4 A_5 dear\_children A_5 A_3 all.} \\ \mathbf{A_2 \to A_3 you A_5} \\ \mathbf{A_3 \to A_4\_to\_} \\ \mathbf{A_4 \to Good\_morning} \\ \mathbf{A_5 \to ,\_} \end{array}\right\}.$$

*Good morning to you,*
*Good morning to you,*
*Good morning, dear children,*
*Good morning to all.*

# The grammar-based codes

- A function $\mathbf{\Gamma}$ such that $\mathbf{\Gamma(w)}$ is an admissible grammar that generates text $\mathbf{w}$ is called a grammar transform.
- Certain grammar transforms can be turned into universal codes if we apply a certain encoding of an arbitrary grammar into a string.
- We may suppose that the number of distinct words in text $\mathbf{X_1^n}$ can be approximated by the number of distinct nonterminals $\mathbf{V(X_1^n)}$ in an admissibly minimal grammar-based code $\mathbf{C(X_1^n)}$ for text $\mathbf{X_1^n}$.

# The first result (non-zero entropy rate)

- Admissibly minimal grammar-based codes satisfy:

$$|\mathbf{C(u)}| + |\mathbf{C(v)}| - |\mathbf{C(uv)}| \leq \mathbf{BV(uv)(1 + L(uv))}.$$

- The left hand side is an estimate of mutual information

$$\mathbf{I(X;Y) = H(X) + H(Y) - H(X,Y)}.$$

**Theorem**

*Let* $\mathbf{V(X_1^n)}$ *be the number of distinct nonterminals in an admissibly minimal grammar-based code* $\mathbf{C(X_1^n)}$*. If for a stationary process* $\mathbf{(X_i)_{i\in\mathbb{Z}}}$ *over a finite alphabet with a strictly positive entropy rate we have* $\mathbf{I(X_1^n; X_{n+1}^{2n}) = \Omega(n^\beta)}$ *for some* $\beta \in (0,1)$ *then*

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\mathbf{V(X_1^n)(1 + L(X_1^n))}\right]}{\mathbf{n^\beta}} > 0.$$

# The second result (zero entropy rate)

- We suppose that admissibly minimal grammar-based codes are universal, i.e., they satisfy $\mathbb{E}\left|\mathbf{C}(\mathbf{X}_1^{\mathbf{n}})\right| - \mathbf{H(n)} = \mathbf{o(n)}$. This would guarantee $\mathbb{E}\left|\mathbf{C}(\mathbf{X}_1^{\mathbf{n}})\right| = \mathbf{o(n)}$ if $\mathbf{H(n)} = \mathbf{o(n)}$.

**Theorem**

*Let* $\mathbf{V}(\mathbf{X}_1^{\mathbf{n}})$ *be the number of distinct nonterminals in an admissibly minimal grammar-based code* $\mathbf{C}(\mathbf{X}_1^{\mathbf{n}})$. *If for a stationary process* $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ *over a finite alphabet the code satisfies* $\mathbb{E}\left|\mathbf{C}(\mathbf{X}_1^{\mathbf{n}})\right| = \mathbf{o(n)}$ *then*

$$\limsup_{\mathbf{n}\to\infty} \frac{\mathbb{E}\left[\mathbf{V}(\mathbf{X}_1^{\mathbf{n}})(1+\mathbf{L}(\mathbf{X}_1^{\mathbf{n}}))\right]}{\mathbf{n}/\mathbb{E}\,\mathbf{L}(\mathbf{X}_1^{\mathbf{n}})} > \mathbf{0}.$$

# The second result for regular Hilberg processes

For a regular Hilberg process we have:

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[V(X_1^n)(1+L(X_1^n))\right]}{n/(\log n)^\alpha} > 0.$$

# Processes satisfying Hilberg's conjecture?

- Hilberg's conjecture:

$$\mathbf{H(n) \approx Bn^{\beta} + hn.}$$

- There are a few processes that satisfy HC with $\mathbf{h > 0}$.
- We have some idea how to construct a process that satisfies HC with $\mathbf{h = 0}$ but have not completed the construction yet.

# The Santa Fe process

## A linguistic interpretation

Process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of random statements consistently describing the state of an "earlier drawn" random object $(Z_k)_{k\in\mathbb{N}}$. $X_i = (k, z)$ asserts that the $k$-th bit of $(Z_k)_{k\in\mathbb{N}}$ has value $Z_k = z$.

Let a process $(X_i)_{i\in\mathbb{Z}}$ have the form

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

where $(K_i)_{i\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$ are independent IID processes,

$$P(K_i = k) = k^{-1/\beta}/\zeta(\beta^{-1}), \qquad \beta \in (0,1),$$
$$P(Z_k = z) = \tfrac{1}{2}, \qquad z \in \{0,1\}.$$

We have $\lim_{n\to\infty} I(X_1^n; X_{n+1}^{2n})/n^\beta > 0$.

# A mixing Santa Fe process

## A linguistic interpretation

Object $(\mathbf{Z_{ik}})_{\mathbf{k}\in\mathbb{N}}$ described by text $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ is a function of time $\mathbf{i}$.

Let a process $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ have the form

$$\mathbf{X_i} := (\mathbf{K_i}, \mathbf{Z_{i,K_i}}),$$

where $(\mathbf{K_i})_{\mathbf{i}\in\mathbb{Z}}$ and $(\mathbf{Z_{ik}})_{\mathbf{i}\in\mathbb{Z}}$, $\mathbf{k}\in\mathbb{N}$, are independent,

$$\mathbf{P(K_i = k)} = \mathbf{k}^{-1/\beta}/\zeta(\beta^{-1}), \qquad (\mathbf{K_i})_{\mathbf{i}\in\mathbb{Z}} \sim \mathrm{IID},$$

whereas $(\mathbf{Z_{ik}})_{\mathbf{i}\in\mathbb{Z}}$ are Markov chains with

$$\mathbf{P(Z_{ik} = z)} = \tfrac{1}{2},$$
$$\mathbf{P(Z_{ik} = z | Z_{i-1,k} = z)} = \mathbf{1 - p_k}.$$

We have $\mathbf{\lim_{n\to\infty} I(X_1^n; X_{n+1}^{2n})/n^\beta > 0}$ for $\mathbf{p_k \le P(K_i = k)}$.

# Conclusions

- Hilberg's conjecture is a hypothesis about a power law growth of block entropy for texts in natural language.
- It may have profound implications for text compression, statistical natural language modeling, and machine learning.
- Further fundamental mathematical research is needed (models of processes, entropy estimation methods).

www.ipipan.waw.pl/~ldebowsk