A Refutation of Finite-State Language Models through Zipf's Law for Factual Knowledge

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Tenth Peripatetic Conference, Zakopane, October 2021

The aim of the paper

We will show that finite-state statistical language models can be refuted using an argument based on semantics rather than syntax.

- This semantic argument is rooted in recent mathematical research in information theory.
- Even if some hypotheses thereof do not pertain to natural language, we suppose that our reasoning points out interesting directions of future research.
- Despite Claude Shannon's influential opinion, information theory is also a theory of semantics but a quantitative one. It deals with amounts of meaning rather than with structures thereof. Yet, amounts and structures constrain one another.

We presented similar results at previous Peripatetic Conferences. This paper improves on several mathematical details.

1 Introduction

2 Finite-state models

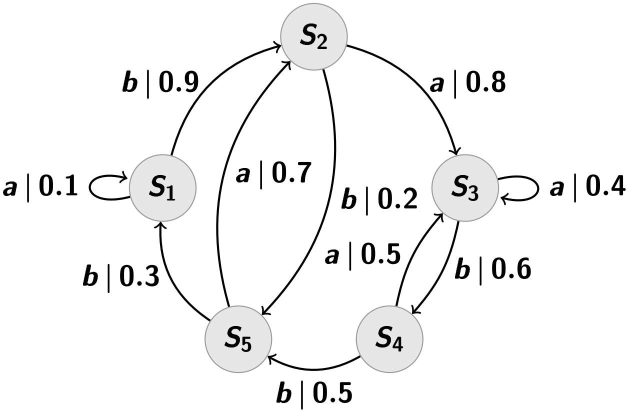
3 Zipf's law

4 Factual knowledge

5 Mutual information

6 Conclusion

Finite-state automata and processes



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

Is natural language a finite-state process?

- **Yes:** Burrhus 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:** Noam 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).

We will demonstrate a novel argument against finite-state models, which is based on a hypothetical Zipf law for factual knowledge.

1 Introduction

2 Finite-state models

3 Zipf's law

4 Factual knowledge

5 Mutual information

6 Conclusion

Rank list of words (Shakespeare's plays)

rank	frequency	word
1	21557	I
2	19059	and
3	16571	to
4	14921	of
5	14491	a
6	12077	my
7	10463	you
8	9789	in
9	8754	is
10	7428	that
...	...	...

Zipf's distribution

Zipf's law (empirical law)

If we count frequencies of words and sort them with respect to decreasing frequencies then they roughly follow Zipf's distribution.

Zipf's distribution

For a random variable K taking values in natural numbers,

$$P(K = k) = \frac{1}{\zeta(\alpha)} \cdot \frac{1}{k^\alpha}, \quad \alpha > 1,$$

where $\zeta(\alpha)$ is the famous Riemann zeta function,

$$\zeta(\alpha) := \frac{1}{1^\alpha} + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \ldots$$

The monkey-typing explanation

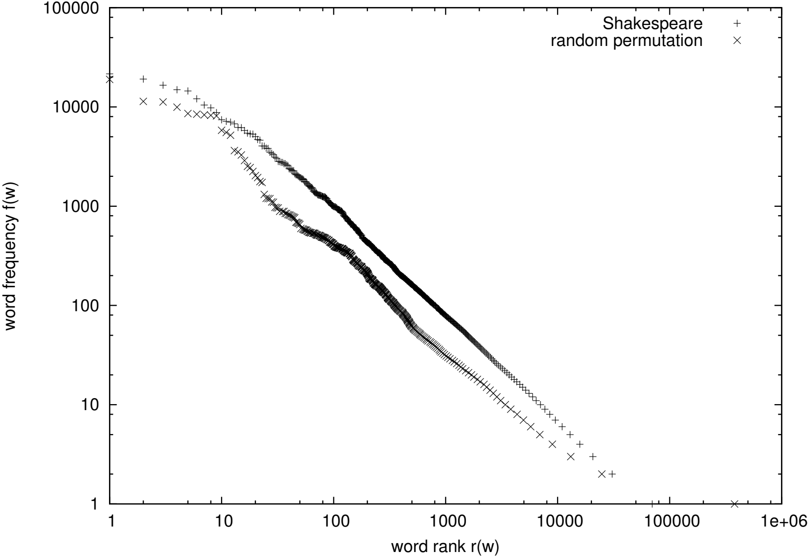
Benoît Mandelbrot (1954), George A. Miller (1957):

A simple finite-state process which exhibits Zipf's law:

If we press keys of the keyboard at random, the resulted text obeys Zipf's law for strings of letters separated by spaces.

Thus, *mere* Zipf law cannot refute finite-state models.

Zipf's law in the plot



1 Introduction

2 Finite-state models

3 Zipf's law

4 Factual knowledge

5 Mutual information

6 Conclusion

Factual knowledge cannot be ignored

- Linguists often assume that the description of language system can be delineated from factual knowledge expressed in texts.
- In statistical language modeling (speech recognition/machine translation), we cannot afford ignoring factual knowledge:

> *"Every time I fire a linguist the performance improves."*
> — *attributed to Frederick Jelinek*

- We have to model also things that are expressed in language, which come as a large number of rare events (LNRE).
- Under Zipf's law, roughly a half of the vocabulary of a text are hapax legomena, i.e., words that appear only once.
- The same may apply to mentions of facts.

Santa Fe process: a model of a random consistent text

Let $(K_i)_{i=1}^{\infty}$ be independent variables following Zipf's distribution.
Let $(Z_k)_{k=1}^{\infty}$ be a sequence of random independent bits (facts).
The Santa Fe process $(X_i)_{i=1}^{\infty}$ is an infinite sequence of pairs

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

A semantic interpretation

Process $(X_i)_{i=1}^{\infty}$ is a sequence of random propositions:

- Proposition $X_i = (k, z)$ asserts that the k-th fact has value z, in such way that one can determine both k and z.
- For $X_i = (k, z)$ and $X_j = (k', z')$ we do not know in advance which facts they describe but $k = k' \implies z = z'$.

The number $U(n)$ of facts that can be computed from both $(X_{-n}, ..., X_0)$ and $(X_1, ...X_n)$ is $\propto n^{\beta}$, $\beta = 1/\alpha$.

1 Introduction

2 Finite-state models

3 Zipf's law

4 Factual knowledge

5 Mutual information

6 Conclusion

Mutual information in the Shannon framework

- Shannon entropy: $H(X) := -\sum_x P(X = x) \log P(X = x)$
- Random strings of letters: $X_j^k := (X_j, X_{j+1}, ..., X_k)$
- Shannon mutual information:

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

- OLD RESULT: Bound for finite-state processes:

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

- Bound for Santa Fe processes:

$$I(X_{-n}^0; X_1^n) \geq \mathbf{E}\, U(n) \propto n^\beta, \quad \beta = 1/\alpha$$

Mutual information in the algorithmic framework

- Kolmogorov complexity:
  $K(x) :=$ (the length of the shortest program computing x)
- Strings of letters: $x_j^k := (x_j, x_{j+1}, ..., x_k)$
- Algorithmic mutual information:

$$J(x; y) := K(x) + K(y) - K(x, y) \geq 0$$

- NEW RESULT: Bound for finite-state processes:

$$\mathbf{E}\, J(X_{-n}^{0}; X_1^n) \leq (\#\text{ of hidden states}) \log n$$

- Bound for Santa Fe processes:

$$\mathbf{E}\, J(X_{-n}^{0}; X_1^n) \geq \mathbf{E}\, U(n) \propto n^{\beta}, \quad \beta = 1/\alpha$$

An overseen more prominent effect

For text length $\boldsymbol{n}$, the # of hidden states is greater than:

- $\boldsymbol{D(n)}$ — maximal depth of central embedding,
- $\boldsymbol{U(n)}$ — amount of factual knowledge conveyed repeatedly.

Values $\boldsymbol{D(n)}$ and $\boldsymbol{U(n)}$ are finite but we may suppose that $\boldsymbol{U(n)}$ grows like $\boldsymbol{n^{\beta}}$ whereas $\boldsymbol{D(n)}$ grows like $\boldsymbol{\log n}$.

Semantics matters more than syntax.

Hilberg's hypothesis and Big Data

- German telecom engineer Wolfgang Hilberg (1990) replotted Shannon's (1951) guessing data in log-log scale:

$$I(X_{-n}^{0}; X_1^n) \propto n^{\beta}, \quad \beta \approx 0.5, \; n \leq 100.$$

- Estimates for `bzip2` of 8GB text (Takahira et al., 2016):

$$J(x_{-n}^{0}; x_1^n) \propto n^{\beta}, \quad \beta \approx 0.8, \; n \leq 10^9.$$

  We call this relationship Hilberg condition.
- Similar estimates for neural statistical language models:
  - Hestness et al. (2017). Deep Learning Scaling Is Predictable, Empirically.
  - Hahn, Futrell (2019). Estimating Predictive Rate-Distortion Curves via Neural Variational Inference.
  - Braverman et al. (2020). Calibration, Entropy Rates, and Memory in Language Models.
  - Kaplan et al. (2020). Scaling Laws for Neural Language Models.
  - Henighan et al. (2020). Scaling Laws for Autoregressive Generative Modeling.
  - Hernandez et al. (2021). Scaling Laws for Transfer.

1 Introduction

2 Finite-state models

3 Zipf's law

4 Factual knowledge

5 Mutual information

6 Conclusion

A refutation of finite-state models

- As we have shown, finite-state language models do not satisfy the Hilberg condition neither in the Shannon framework nor in the algorithmic one.
- By contrast, Santa Fe processes satisfy the Hilberg condition.
- Santa Fe processes are stationary processes in which mentions of independent elementary facts are distributed asymptotically according to Zipf's law.

Open questions

To what extent does language resemble a Santa Fe process?
Can we estimate the amount of facts carried by a text?

References

1. Ł. Dębowski, (2021). A Refutation of Finite-State Language Models Through Zipf's Law for Factual Knowledge. *Entropy*, vol. 23, pp. 1148.
2. Ł. Dębowski, (2018). Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy*, vol. 20(2), pp. 85.
3. R. Takahira, K. Tanaka-Ishii, Ł. Dębowski, (2016). Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy*, vol. 18(10), pp. 364.
4. Ł. Dębowski, (2011). On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Transactions on Information Theory*, vol. 57, pp. 4589–4599.
5. Ł. Dębowski, (2006). On Hilberg's law and its links with Guiraud's law. *Journal of Quantitative Linguistics*, vol. 13, pp. 81–109.

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