INFORMATION THEORETIC MODELS OF NATURAL LANGUAGE

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warszawa, POLAND, ldebowsk@ipipan.waw.pl

ABSTRACT

The relaxed Hilberg conjecture is a proposition about natural language which states that mutual information between two adjacent blocks of text grows according to a power law in function of the block length. In the paper two mathematical results connected to this conjecture are reviewed. First, we exhibit an example of a stochastic process, called the Santa Fe process, which is motivated linguistically and for which the mutual information grows according to a power law. Second, we demonstrate that a power law growth of mutual information implies a power law growth of vocabulary. The latter statement is observed for texts in natural language and called Herdan's law.

1. INTRODUCTION

It is often assumed that texts in natural language may be modeled by a stationary process and the entropy a random text can be determined [1]. More specifically, in 1990, German telecommunications engineer Wolfgang Hilberg conjectured that the entropy of a random text in natural language satisfies

$$H(X_1^n) \propto n^\beta, \tag{1}$$

where X_i are characters of the random text, $X_n^m = (X_n, X_{n+1}, ..., X_m)$ are blocks of consecutive characters, $H(X) = \mathbf{E}\,[-\log P(X)]$ is the entropy of a discrete variable X, and $\beta \in (0,1)$ [2]. Hilberg's conjecture was based on an extrapolation of Shannon's seminal experimental data [3], which contained the estimates of conditional entropy for blocks of $n \leq 100$ characters.

Statement (1) implies that the entropy rate $h = \lim_{n\to\infty} H(X_1^n)/n$ equals 0. This in turn implies asymptotic determinism of utterances, which does not sound plausible. A more plausible modification of statement (1) is

$$I(X_1^n; X_{n+1}^{2n}) \propto n^\beta, \tag{2}$$

where $I(X;Y) = H(X) + H(Y) - H(X,Y)$ is the mutual information between variables X and Y. We notice that relationship (2) arises for entropy

$$H(X_1^n) = An^\beta + hn \tag{3}$$

where h can be positive. Relationship (2) will be called the relaxed Hilberg conjecture.

In this paper, we will review some previous results of ours that concern two issues:

1. We exhibit an example of a stochastic process, called Santa Fe process, which is motivated linguistically and which satisfies relationship (2) asymptotically [4].

2. We demonstrate that relationship (2) implies that the text of length n contains at least $n^\beta/\log n$ different words, under a certain plausible definition of a word [5]. Indeed, the power-law growth of the vocabulary is empirically observed for texts in natural language and called Herdan's law [6].

In our opinion, these results shed some light on probabilistic modeling of natural language.

2. THE SANTA FE PROCESS

Processes that satisfy the relaxed Hilberg conjecture arise in a very simple setting that resembles what may actually happen in natural language. Suppose that each statement X_i of a text in natural language can be represented as a pair $X_i = (k, z)$ which states that the k-th proposition in some abstract enumeration assumes Boolean value z. Moreover, suppose that there is a stochastic process $(K_i)_{i\in\mathbb{Z}}$ and a random field $(Z_{ik})_{i\in\mathbb{Z},k\in\mathbb{N}}$ such that if $X_i = (k, z)$ then $K_i = k$ and $Z_{ik} = z$. The process $(K_i)_{i\in\mathbb{Z}}$ will be called the selection process and the field $(Z_{ik})_{i\in\mathbb{Z},k\in\mathbb{N}}$ will be called the object described by text $(X_i)_{i\in\mathbb{Z}}$. Note that variable Z_{ik} has two indices—the first one refers to the while i at which the statement X_i is made whereas the second one refers to the proposition $K_i = k$, which is either asserted or negated. Observe that statements that are made in texts fall under two types:

1. Statements about objects $Z_{ik} = Z_k$, which do not change in time (like mathematical or physical constants).

2. Statements about objects $Z_{ik} \neq Z_{i+1,k}$, which evolve with a varied speed (like culture, language, or geography).

We will obtain a power-law growth of mutual information for an appropriate choice of the selection process and the described object, namely, when the bits of the described object do not evolve too fast in comparison to their selection by the selection process.

In particular, the Santa Fe process $(X_i)_{i\in\mathbb{Z}}$ will be defined as a sequence of random statements

$$X_i = (K_i, Z_{i,K_i}), \tag{4}$$

where processes $(K_i)_{i\in\mathbb{Z}}$ and $(Z_{ik})_{i\in\mathbb{Z}}$ with $k \in \mathbb{N}$ are independent and distributed as follows. First, variables K_i are distributed according to the power law

$$P(K_i = k) = k^{-1/\beta}/\zeta(\beta^{-1}), \quad (K_i)_{i\in\mathbb{Z}} \sim \text{IID}, \tag{5}$$

where $\beta \in (0,1)$ and $\zeta(x) = \sum_{k=1}^{\infty} k^{-x}$ is the zeta function. Second, each process $(Z_{ik})_{i\in\mathbb{Z}}$ is a Markov chain with the marginal distribution

$$P(Z_{ik} = 0) = P(Z_{ik} = 1) = 1/2 \tag{6}$$

and the cross-over probabilities

$$P(Z_{ik} = 0|Z_{i-1,k} = 1) = P(Z_{ik} = 1|Z_{i-1,k} = 0) = p_k. \tag{7}$$

The name "Santa Fe process" has been chosen since the author discovered this process during a stay at the Santa Fe Institute.

Observe that the description given by the Santa Fe process is strictly repetitive for $p_k = 0$: if two statements $X_i = (k, z)$ and $X_j = (k', z')$ describe bits of the same address ($k = k'$) then they always assert the same bit value ($z = z'$). In this case the Santa Fe process is nonergodic. For strictly positive p_k the description is no longer strictly repetitive and the Santa Fe process is mixing [4].

By the following result, the Santa Fe process satisfies relationship (2) asymptotically:

Theorem 1 ([4]) *Suppose* $\lim_{k\to\infty} p_k/P(K_i = k) = 0$. *Then the mutual information for the Santa Fe process obeys*

$$\lim_{n\to\infty} \frac{I(X_1^n; X_{n+1}^{2n})}{n^\beta} = \frac{(2-2^\beta)\Gamma(1-\beta)}{[\zeta(\beta^{-1})]^\beta}. \tag{8}$$

Some processes over a finite alphabet which also satisfy relationship (2) asymptotically can be constructed by stationary coding of the Santa Fe process [4].

3. VOCABULARY GROWTH

In the second turn, we will show that the relaxed Hilberg conjecture can be related to the number of distinct words appearing in texts. It has been observed that words in natural language texts correspond in a good approximation to nonterminal symbols in the shortest grammar-based encoding of those texts [7, 8, 9]. Complementing this observation, we will demonstrate that relationship (2) constrains the number of distinct nonterminal symbols in the shortest grammar-based encoding of the random text.

A short introduction to grammar-based coding is in need. Briefly speaking, grammar-based codes compress strings by transforming them first into special grammars, called admissible grammars [10], and then encoding the grammars back into strings according to a fixed simple method. An admissible grammar is a context-free grammar that generates a singleton language $\{w\}$ for some string $w \in \mathbb{X}^*$ [10]. In an admissible grammar, there is exactly one rule per nonterminal symbol and the nonterminals can be ordered so that the symbols are rewritten onto strings of strictly succeeding symbols [10]. Hence, such a grammar is given by its set of production rules

$$\left\{ \begin{array}{l} A_1 \to \alpha_1, \\ A_2 \to \alpha_2, \\ ..., \\ A_n \to \alpha_n \end{array} \right\}, \tag{9}$$

where A_1 is the start symbol, other A_i are secondary nonterminals, and the right-hand sides of rules satisfy $\alpha_i \in (\{A_{i+1}, A_{i+2}, ..., A_n\} \cup \mathbb{X})^*$.

An example of an admissible grammar is

$$\left\{ \begin{array}{l} A_1 \mapsto A_2A_2A_4A_5\texttt{dear\_children}A_5A_3\texttt{all.} \\ A_2 \mapsto A_3\texttt{you}A_5 \\ A_3 \mapsto A_4\texttt{\_to\_} \\ A_4 \mapsto \texttt{Good\_morning} \\ A_5 \mapsto \texttt{,\_} \end{array} \right\},$$

with the start symbol A_1, which produces the song

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

For the shortest grammar-based encoding of a longer text in natural language, secondary nonterminals A_i often match the word boundaries, especially if it is required that these nonterminals are defined using only terminal symbols [9].

In the following, $\mathbf{V}(w)$ will denote the number of distinct nonterminal symbols in the shortest grammar-based encoding of a text w. (The exact definition of the shortest grammar-based encoding, called admissibly minimal, is given in [5].) To connect the mutual information with $\mathbf{V}(w)$, we introduce another quantity, namely the length of the longest nonoverlapping repeat in a text w:

$$\mathbf{L}(w) := \max\{|s| : w = x_1sy_1 = x_2sy_2 \wedge x_1 \neq x_2\}, \tag{10}$$

where $s, x_i, y_i \in \mathbb{X}^*$. Using this concept, for processes over a finite alphabet we obtain this proposition.

Theorem 2 ([5]) *Let* $(X_i)_{i\in\mathbb{Z}}$ *be a stationary process over a finite alphabet. Assume that inequality*

$$\liminf_{n\to\infty} \frac{I(X_1^n; X_{n+1}^{2n})}{n^\beta} > 0 \tag{11}$$

holds for some $\beta \in (0,1)$ *and*

$$\sup_{n\geq 2} \boldsymbol{E}\left(\frac{\mathbf{L}(X_1^n)}{f(n)}\right)^q < \infty, \quad \forall q > 0, \tag{12}$$

holds for some function $f(n)$. *Then we have*

$$\limsup_{n\to\infty} \boldsymbol{E}\left(\frac{\mathbf{V}(X_1^n)}{n^\beta f(n)^{-1}}\right)^p > 0, \quad \forall p > 1. \tag{13}$$

An example of a process that satisfies the hypothesis of Theorem 2 with $f(n) = \log n$ can be constructed by stationary coding of the Santa Fe process [11, 4]. However, for texts in natural language we have checked that there holds an empirical law $\mathbf{L}(X_1^n) \approx \log^\alpha n$, where $\alpha \approx 2 \div 3$ [12]. It is an interesting open question how to construct processes which satisfy both (11) and $\mathbf{L}(X_1^n) \approx \log^\alpha n$.

4. CONCLUSION

We have discussed some constructions and theorems for discrete-valued processes with long memory. Our results have very natural linguistic interpretations. We believe that the Santa Fe process deserves further investigation.

5. REFERENCES

[1] Thomas M. Cover and Roger C. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inform. Theor.*, vol. 24, pp. 413–421, 1978.

[2] Wolfgang Hilberg, "Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?," *Frequenz*, vol. 44, pp. 243–248, 1990.

[3] Claude Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1951.

[4] Łukasz Dębowski, "Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks," *IEEE Trans. Inform. Theor.*, vol. 58, pp. 3392–3401, 2012.

[5] Łukasz Dębowski, "On the vocabulary of grammar-based codes and the logical consistency of texts," *IEEE Trans. Inform. Theor.*, vol. 57, pp. 4589–4599, 2011.

[6] Gustav Herdan, *Quantitative Linguistics*, London: Butterworths, 1964.

[7] J. Gerard Wolff, "Language acquisition and the discovery of phrase structure," *Lang. Speech*, vol. 23, pp. 255–269, 1980.

[8] Carl G. de Marcken, *Unsupervised Language Acquisition*, Ph.D. thesis, Massachussetts Institute of Technology, 1996.

[9] Chunyu Kit and Yorick Wilks, "Unsupervised learning of word boundary with description length gain," in *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*, M. Osborne and E. T. K. Sang, Eds., pp. 1–6. 1999.

[10] John C. Kieffer and Enhui Yang, "Grammar-based codes: A new class of universal lossless source codes," *IEEE Trans. Inform. Theor.*, vol. 46, pp. 737–754, 2000.

[11] Łukasz Dębowski, "Variable-length coding of two-sided asymptotically mean stationary measures," *J. Theor. Probab.*, vol. 23, pp. 237–256, 2010.

[12] Łukasz Dębowski, "Maximal lengths of repeat in English prose," in *Synergetic Linguistics. Text and Language as Dynamic System*, Sven Naumann, Peter Grzybek, Relja Vulanović, and Gabriel Altmann, Eds., pp. 23–30. Wien: Praesens Verlag, 2012.