Faktów jest mniej niż słów: Prawa potęgowe dla procesów stacjonarnych

Łukasz Dębowski
Instytut Podstaw Informatyki PAN
e-mail: ldebowsk@ipipan.waw.pl

W eksperymentach obliczeniowych w statystycznym modelowaniu języka naturalnego obserwuje się wielkoskalowe ogony potęgowe krzywych uczenia [5, 7, 8, 9]. Różnica między intensywnością entropii krzyżowej modelu statystycznego a intensywnością entropii języka maleje jak potęga ilości danych uczących. Obserwacja ta jest równoważna wzrostowi potęgowemu informacji wzajemnej między rosnącymi blokami tekstu — pierwsza wzmianka tej prawidłowości pochodzi z pracy [6], por. [1]. Wzrost potęgowy występuje w przypadku języków tak różnorodnych jak angielski, francuski, rosyjski, chiński, koreański i japoński. Co więcej, obserwuje się niezależną od języka wartość wykładnika: informacja wzajemna między dwoma blokami długości n jest proporcjonalna do $n^{0,8}$ [8, 9].

Anonsujemy matematyczną teorię tego zjawiska, którą rozwijamy od pewnego czasu. Większość wyników badań podsumowaliśmy w monografii [3] i artykułach [2, 4]. Centralnym elementem teorii krzywych uczenia o zaniku potęgowym jest twierdzenie postaci:

Liczba niezależnych faktów opisanych w skończonym tekście jest w przybliżeniu mniejsza niż liczba różnych słów użytych w tym tekście.

Tego rodzaju stwierdzenie nazywamy twierdzeniem o faktach i słowach. Twierdzenie o faktach i słowach jest wynikiem, który dotyczy dowolnego dyskretnego procesu stacjonarnego i wiąże dekompozycję ergodyczną z semantyką i statystyką.

Wynik ten wydaje się paradoksalny, gdyż można by sądzić, że łącząc słowa, daje się wyrazić znacząco więcej niezależnych faktów. Twierdzenie to jednak łatwo jest udowodnić, przyjmując dość naturalne definicje faktów i słów. Idee są następujące:

Preliminaria: Rozpatrujemy dyskretny proces stacjonarny $(X_i)_{i\in\mathbb{Z}}$. Bloki zmiennych losowych oznaczamy $X_j^k := (X_j, X_{j+1}, ..., X_k)$. *Entropia Shannona* zmiennej dyskretnej X to $H(X) := \mathbf{E}\,[-\log P(X)]$, gdzie $\mathbf{E}\,Y$ to wartość oczekiwana Y. Istnieje granica nazywana *intensywnością entropii* $h := \lim_{n\to\infty} H(X_1^n)/n$. Efekty potęgowe opisuje *wykładnik Hilberga* $\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+$. Na przykład *wykładnik redundancji* to

$$\beta := \operatorname*{hilb}_{n\to\infty} \left[H(X_1^n) - hn\right]. \tag{1}$$

Warunek $\beta > 0$ nazywamy *prawem Hilberga*, na cześć autora pracy [6].

Fakty i wykładnik redundancji: Proces stacjonarny $(X_i)_{i\in\mathbb{Z}}$ o bezatomowym σ-ciele niezmienniczym nazywamy *mocno nieergodycznym*. W tym przypadku, niech $(Z_k)_{k\in\mathbb{N}}$ będzie procesem Bernoulli($\frac{1}{2}$) mierzalnym względem tego σ-ciała niezmienniczego. Zmienne Z_k nazywamy *faktami*, bo nie zależą od czasu. Mówimy, że skończony tekst x_1^n opisuje l początkowych faktów według funkcji g, jeżeli $l+1 = U_g(x_1^n) := \min\{k \in \mathbb{N} : g(k, x_1^n) \neq Z_k\}$. Dla mocno nieergodycznego procesu $(X_i)_{i\in\mathbb{Z}}$ i dowolnej funkcji g, zachodzi dolne ograniczenie wykładnika redundancji przez liczbę opisanych faktów

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n) \leqslant \beta. \tag{2}$$

Proces Santa Fe: Pojęcie faktów warto zilustrować prostym przykładem mocno nieergodycznego procesu nazwanego procesem Santa Fe [2]. Niech $(K_i)_{i\in\mathbb{Z}}$ będzie procesem IID o wartościach w liczbach naturalnych i *rozkładzie Zipfa* $P(K_i = k) \sim k^{-\alpha}$, gdzie $\alpha > 1$. Niech proces $(Z_k)_{k\in\mathbb{N}}$ będzie Bernoulli($\frac{1}{2}$). *Proces Santa Fe* $(X_i)_{i\in\mathbb{Z}}$ to ciąg par

$$X_i = (K_i, Z_{K_i}). \tag{3}$$

Proces Santa Fe jest uproszczonym modelem tekstu, który składa się z losowych stwierdzeń postaci „k-ty fakt ma wartość Z_k". Stwierdzenia te są *niesprzeczne*. To znaczy, jeżeli stwierdzenia X_i oraz X_j opisują ten sam fakt ($K_i = K_j$), to przypisują mu tę samą wartość ($Z_{K_i} = Z_{K_j}$). Kładąc $g(k, x_1^n) := z$, jeśli $(k, z) \in x_1^n$ i $(k, 1-z) \notin x_1^n$, oraz $g(k, x_1^n) := 2$ dla innych (k, x_1^n), otrzymujemy prawo potęgowe $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n) = 1/\alpha \in (0, 1)$.

Słowa i informacja wzajemna: Rozpatrujemy procesy stacjonarne $(X_i)_{i\in\mathbb{Z}}$ nad alfabetem D-arnym. *Maksimum wiarogodności* (ML) w klasie procesów Markowa to

$$\hat{\mathbb{P}}(k|x_1^n) := \max_Q \prod_{i=k+1}^{n} Q(x_i|x_{i-k}^{i-1}), \qquad Q(x_i|x_{i-k}^{i-1}) \geqslant 0, \qquad \sum_{x_i} Q(x_i|x_{i-k}^{i-1}) = 1. \tag{4}$$

Rozważamy też *liczbę podsłów* $V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \leqslant i \leqslant n-k\right\}$. *Penalizowane maksimum wiarogodności* (PML) to $\mathbb{P}(x_1^n) := w_n \max_{k\geqslant 0} \dfrac{w_k \hat{\mathbb{P}}(k|x_1^n)}{Z(k|x_1^n)}$, gdzie $w_k := \dfrac{1}{k+1} - \dfrac{1}{k+2}$ oraz $\log Z(k|x_1^n) := k \log D + DV(k|x_1^n)(\log n + 6)$. Ponieważ $\sum_{n\geqslant 0} \sum_{x_1^n} \mathbb{P}(x_1^n) \leqslant 1$, zachodzi $\mathbf{E}\, K(X_1^n) \geqslant H(X_1^n)$ dla *entropii PML* $K(u) := -\log \mathbb{P}(u)$. Mamy też mocną i słabą uniwersalność, $\lim_{n\to\infty} K(X_1^n)/n = h$ pnp. i w L^1. Stąd wykładnik redundancji jest ograniczony przez *informację wzajemną PML* $J(u, v) := K(u) + K(v) - K(u, v)$ jako

$$\beta \leqslant \operatorname*{hilb}_{n\to\infty} [\mathbf{E}\, K(X_1^n) - hn] = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^{2n}). \tag{5}$$

Statystyka $M(x_1^n) := \min\left\{k \geqslant 0 : \hat{\mathbb{P}}(k|x_1^n) \geqslant \mathbb{P}(x_1^n)\right\}$ jest zgodnym i asymptotycznie nieobciążonym estymatorem *rzędu Markowa*. Mamy $\lim_{n\to\infty} M(X_1^n) = M$ pnp. i w L^1, gdzie $M := \inf\left\{k \geqslant 0 : H(X_i|X_{i-k}^{i-1}) = h\right\}$ oraz $H(X|Y) := H(X, Y) - H(Y)$. Ponieważ zachodzi nierówność $M(x_1^n)K(x_1^n) \leqslant n \log n$, zachodzi też ograniczenie

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^{2n}) \leqslant \operatorname*{hilb}_{n\to\infty} \mathbf{E}\left[DV(X_1^n) + \frac{n \log D}{K(X_1^n)}\right], \tag{6}$$

gdzie *liczba podsłów Markowa* jest zdefiniowana jako $V(x_1^n) := V(M(x_1^n)|x_1^n)$. Potęgowy wzrost statystyki $V(x_1^n) \sim n^{0,8}$ obserwuje się dla języka naturalnego.

Łącząc nierówności (2), (5) i (6), otrzymujemy twierdzenie o faktach i słowach.

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

[2] Ł. Dębowski, On the vocabulary of grammar-based codes and the logical consistency of texts, *IEEE Transactions on Information Theory*, 57, 4589–4599, 2011

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

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

[5] 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, *preprint*, `https://arxiv.org/abs/1712.00409`, 2017

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

[7] 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, *preprint*, `https://arxiv.org/abs/2001.08361`, 2020

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

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