W analogii do języka naturalnego... czyli twierdzenie o faktach i słowach dla procesów stacjonarnych

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

LVIII Szkoła Matematyki Poglądowej
24–28.08.2018

Punkt wyjścia

Teoria informacji:

- **Podstawowy problem:**
  Jak zakodować dany napis (ciąg znaków) przy pomocy jak najmniejszej liczby cyfr binarnych (zer lub jedynek)?
- **Zastosowania:**
  — kompresja i przesyłanie danych, teoretyczne podstawy informatyki, statystyki, a nawet matematyki jako takiej.

Statystyczne modelowanie języka naturalnego:

- **Podstawowy problem:**
  Jakie przypisać prawdopodobieństwo dowolnym wypowiedziom w danym języku naturalnym (angielskim, polskim, chińskim)?
- **Zastosowania:**
  — automatyczne rozpoznawanie mowy, klawiatury telefonów komórkowych, maszynowe tłumaczenie.

Temat wystąpienia

Twierdzenie o faktach i słowach (sformułowanie nieformalne)

Liczba niezależnych faktów opisywanych przez skończony tekst jest z grubsza mniejsza niż liczba różnych słów w tymże tekście.

W analogii do języka naturalnego, pojęcie faktu i słowa można zdefiniować dla dość dowolnego procesu stochastycznego.

Ku sformułowaniu formalnemu

Na dalszych slajdach zdefiniujemy:

- $\mathbb{U}(x_1^n)$ — zbiór faktów przewidywalnych z napisu x_1^n.
- $G_{\mathrm{PPM}}(x_1^n)$ — rząd PPM napisu x_1^n.
- $V_{\mathrm{PPM}}(x_1^n)$ — zbiór słów PPM napisu x_1^n.
- Wykładnik Hilberga: $\operatorname{hilb}_{n\to\infty} n^{\beta} = \beta$.

Twierdzenie o faktach i słowach

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończonym alfabecie zachodzi nierówność

$$\operatorname{hilb}_{n\to\infty} \mathbb{E}\,\operatorname{card}\mathbb{U}(X_1^n) \leq \operatorname{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n)\right].$$

Jak zdefiniować słowo w dowolnym tekście?

- W przypadku tekstów w wielu językach naturalnych, słowo to ciąg liter od spacji do spacji:

  Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

- Co począć z japońskim?

  すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心とを授けられており、互いに同 胞の精神をもって行動しなければならない。

- Albo z ciągiem zer i jedynek?

  $\Omega = 0.000000100000010000011000100001101000111\ldots$

Podejście toporne, acz skuteczne

- **Notacja:** napis $x_j^k = x_j x_{j+1} \ldots x_k$, gdzie x_i to symbole.
- Zbiór podsłów długości m w napisie x_1^n to

$$V(m|x_1^n) := \left\{ x_{t+1}^{t+m} : 0 \leq t \leq n - m \right\}.$$

- Niech $G_{opt}(x_1^n)$ będzie optymalną długością podsłowa dla x_1^n.
- Zbiór optymalnych słów w napisie x_1^n to

$$V_{opt}(x_1^n) := V(G_{opt}(x_1^n)|x_1^n).$$

Jak wybrać sensowne $G_{opt}(x_1^n)$?

Rozkłady prawdopodobieństw PPM

Sugestii, jak wybrać sensowną optymalną długość podsłowa, udziela konstrukcja rozkładów prawdopodobieństw PPM.

- Rozkłady PPM (Prediction by Partial Matching) to pewna rodzina rozkładów $\mathbf{PPM}_k(x_1^n)$, gdzie $k = -1, 0, 1, \ldots$
- Każdy rozkład $\mathbf{PPM}_k(x_1^n)$ to rozkład procesu Markowa rzędu k estymowany sukcesywnie na danym napisie x_1^n.
- Dla każdego napisu x_1^n istnieje rząd k, dla którego prawdopodobieństwo $\mathbf{PPM}_k(x_1^n)$ jest największe.

Rozkłady prawdopodobieństw PPM — definicja

Częstość podsłowa w napisie to

$$N(w_1^k|x_1^n) := \sum_{i=1}^{n-k+1} 1\left\{x_i^{i+k-1} = w_1^k\right\},$$

gdzie $x_j^k := \lambda$ oraz $\sum_{i=j}^{k} f(i) := 0$ dla $k < j$.

Dla $x_i \in \{a_1, a_2, ..., a_D\}$ definiujemy

$$\mathrm{PPM}_k(x_i|x_1^{i-1}) := \begin{cases} \dfrac{1}{D}, & k = -1, \\ \dfrac{N(x_{i-k}^{i}|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D}, & k \geq 0, \end{cases}$$

$$\mathrm{PPM}_k(x_1^n) := \prod_{i=1}^{n} \mathrm{PPM}_k(x_i|x_1^{i-1}).$$

Maksymalne powtórzenie, rząd PPM i słownik PPM

- Długość maksymalnego powtórzenia w napisie x_1^n to

$$L(x_1^n) := \max\left\{k : N(w_1^k|x_1^n) \geq 2 \text{ dla pewnego } w_1^k\right\}.$$

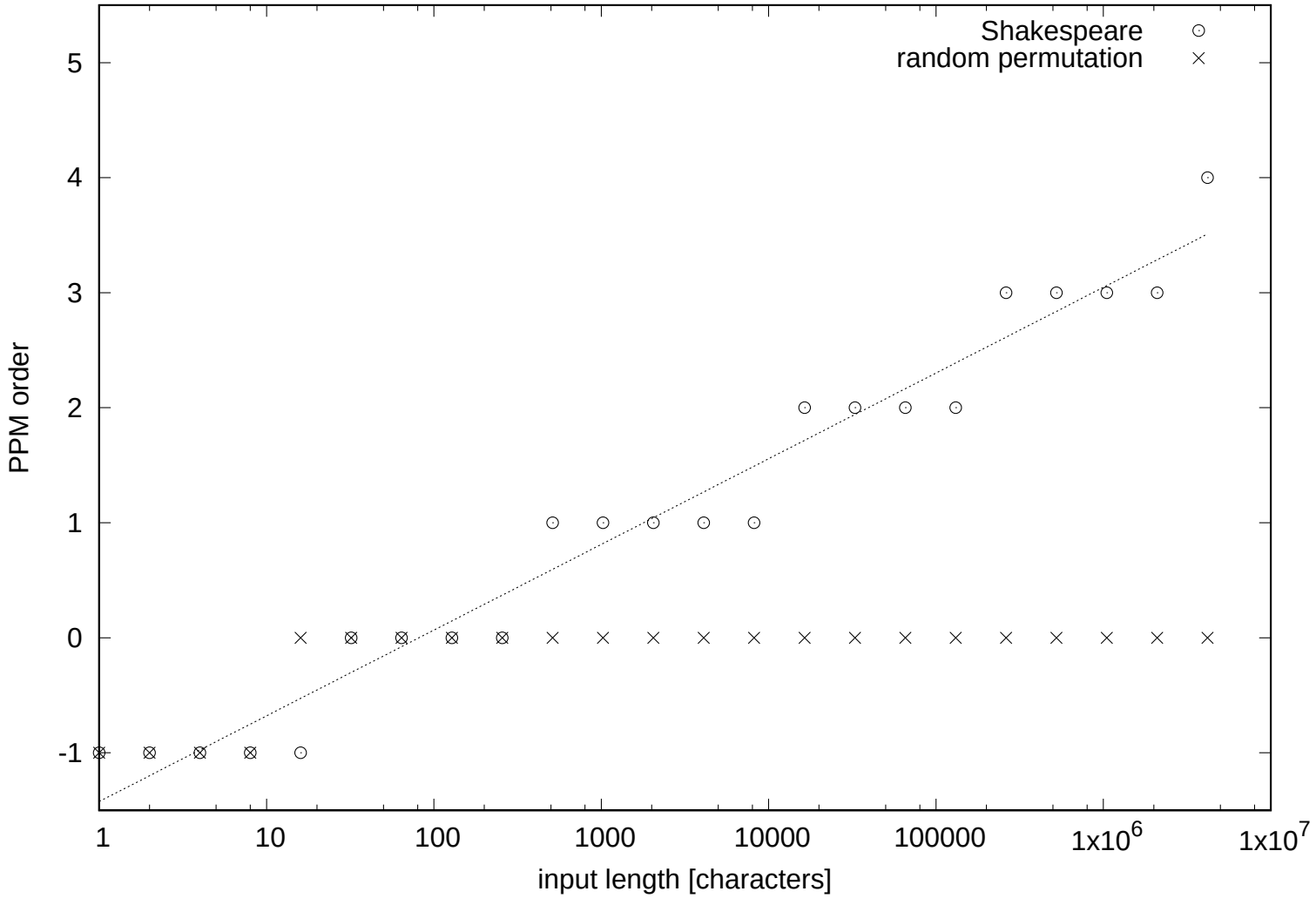
- Zauważmy, że $\mathbf{PPM}_k(x_1^n) = D^{-n}$ dla $k > L(x_1^n)$.
- Rząd PPM $G_{\mathbf{PPM}}(x_1^n)$ to najmniejsza liczba G taka, że

$$\mathbf{PPM}_G(x_1^n) \geq \mathbf{PPM}_k(x_1^n) \text{ dla każdego } k \geq -1.$$

- Mamy $G_{\mathbf{PPM}}(x_1^n) \leq L(x_1^n) \leq n$.
- Zbiór słów PPM w napisie x_1^n to

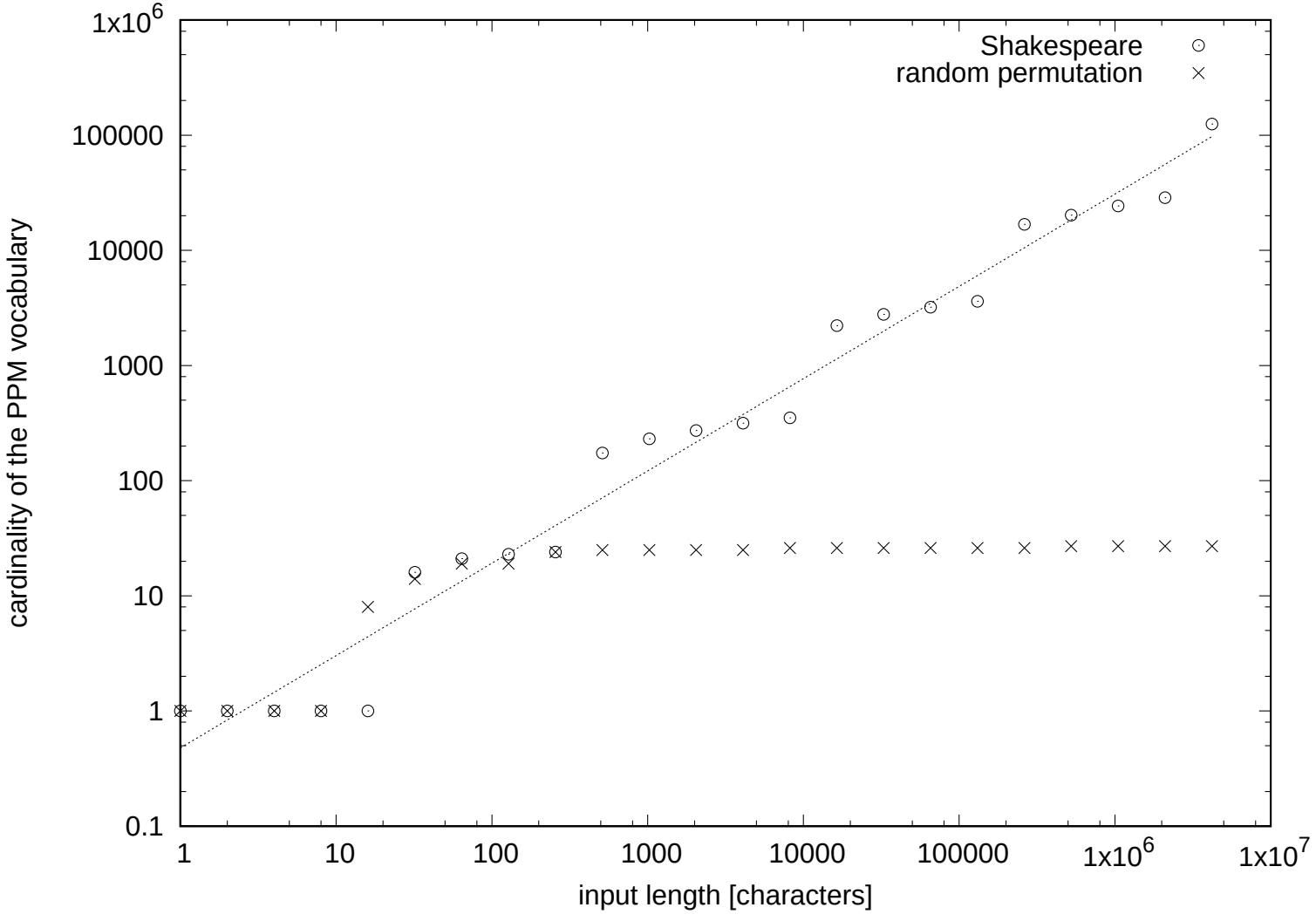
$$V_{\mathbf{PPM}}(x_1^n) := V(G_{\mathbf{PPM}}(x_1^n)|x_1^n).$$

Rząd PPM na wykresie



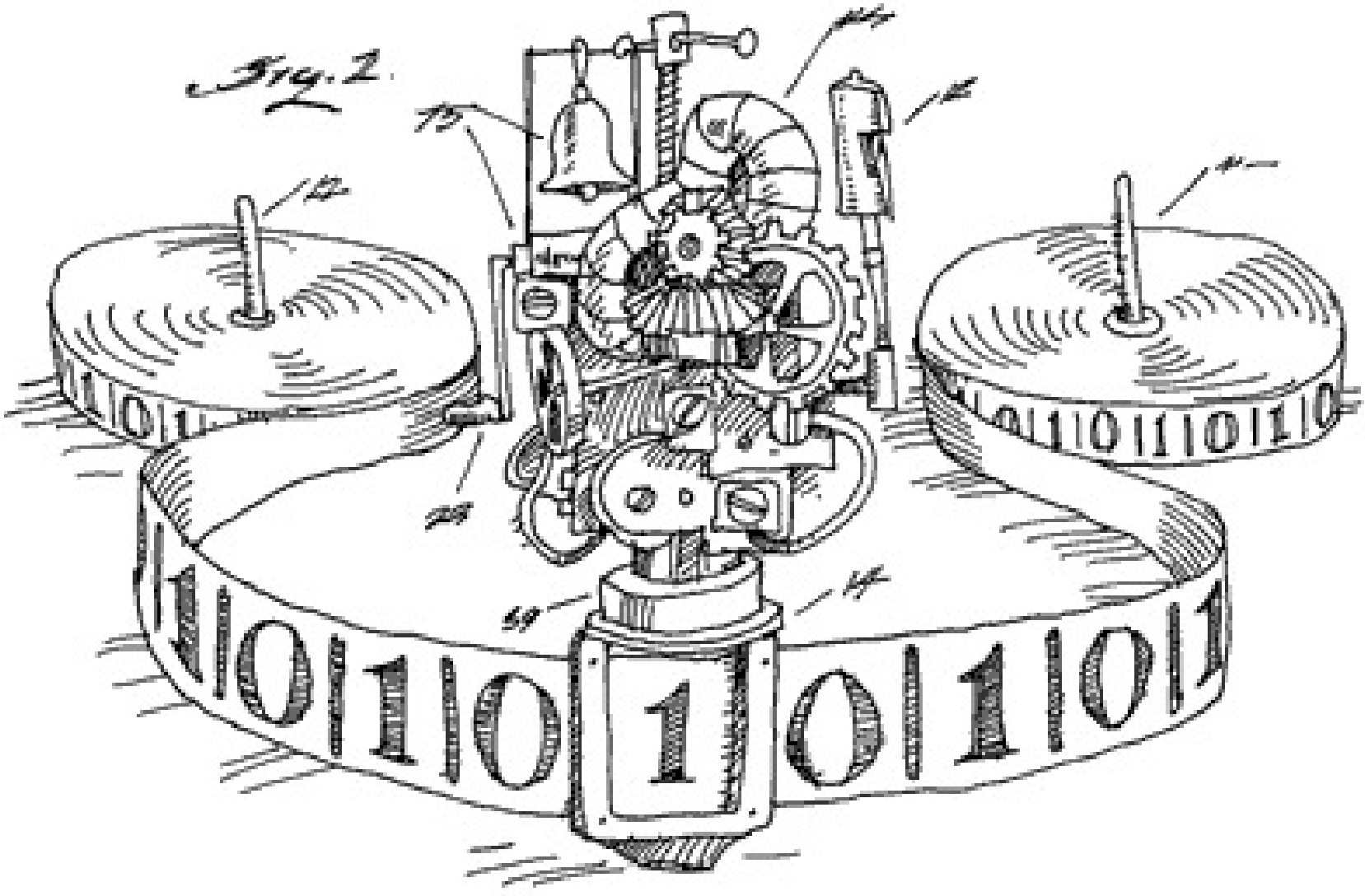
$$G_{\mathrm{PPM}}(x_1^n) \approx -1.423 + 0.323 \ln n$$

Liczba różnych słów PPM na wykresie



$$\operatorname{card} V_{\mathrm{PPM}}(x_1^n) \approx 1.300 n^{0.801}$$

Maszyna Turinga (z dzwonkiem)



Algorytmiczna teoria informacji

- Mamy ustaloną bezprzedrostkową maszynę Turinga.
- Złożoność Kołmogorowa $\mathbb{H}(x_1^n)$ to długość najkrótszego programu generującego napis x_1^n:

$$\mathbb{H}(x_1^n) := \min\left\{|p| : S(p) = x_1^n\right\},$$

gdzie $S(p)$ to wynik programu p.

(Funkcja $x_1^n \mapsto \mathbb{H}(x_1^n)$ nie jest obliczalna.)

- Nieskończony ciąg $(x_i)_{i=1}^{\infty} = (x_1, x_2, x_3, \ldots)$ jest nazywany algorytmicznie losowym, gdy istnieje stała $c > 0$ taka, że

$$\mathbb{H}(x_1^n) \geq n - c.$$

(Zachodzi to, gdy najkrótszy program ma postać **print** x_1^n.)

- Ciąg rzutów uczciwą monetą jest algorytmicznie losowy pnp.

Prawdopodobieństwo stopu

- Prawdopodobieństwo stopu Ω to liczba

$$\Omega = \sum_{p:S(p)\neq\perp} 2^{-|p|} \in (0,1).$$

- Zdefiniujmy rozwinięcie binarne $(\Omega_k)_{k=1}^{\infty} = (\Omega_1, \Omega_2, \Omega_3, \ldots)$, gdzie $\Omega_k \in \{0,1\}$ oraz $\sum_{k=1}^{\infty} 2^{-k}\Omega_k = \Omega$.
- Ciąg $(\Omega_k)_{k=1}^{\infty}$ jest ciągiem algorytmicznie losowym.
- Ponadto, gdybyśmy znali napis Ω_1^n, moglibyśmy odpowiedzieć na pytanie, które stwierdzenia matematyczne długości mniejszej od n są prawdziwe, a które nie.

Prawdopodobieństwo stopu Ω jest „kamieniem filozoficznym”.
Cyfry Ω_k są niezależnymi faktami matematycznymi.

Proces Santa Fe

1. Rozpatrzmy przestrzeń probabilistyczną $(\mathbb{J}, \mathcal{J}, P)$.
2. Niech $(K_i)_{i=1}^{\infty}$ będzie ciągiem niezależnych zmiennych losowych o wartościach w liczbach naturalnych, $K_i : \mathbb{J} \to \mathbb{N} = \{1, 2, 3, ...\}$, i o rozkładzie

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

gdzie $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$.

3. Proces Santa Fe to ciąg zmiennych $(X_i)_{i=1}^{\infty}$ złożonych z par

$$X_i = (K_i, \Omega_{K_i}),$$

gdzie Ω_k to cyfry prawdopodobieństwa stopu.

Od procesu Santa Fe do prawdopodobieństwa stopu

- Zdefiniujmy funkcję:

$$g(k, x_1^n) = \begin{cases} 0, & \text{jeśli } \forall_{i\in\{1,\dots n\}}\,(x_i = (k,z) \implies x_i = (k,0)), \\ 1, & \text{jeśli } \forall_{i\in\{1,\dots n\}}\,(x_i = (k,z) \implies x_i = (k,1)), \\ 2, & \text{inaczej.} \end{cases}$$

- Funkcja ta jest obliczalna i dla procesu Santa Fe spełnia

$$\lim_{n\to\infty} g(k, X_{i+1}^{i+n}) = \Omega_k \text{ prawie na pewno.}$$

- Określmy zbiór niezależnych faktów przewidywalnych z x_1^n:

$$\mathbb{U}(x_1^n) := \left\{ l \in \mathbb{N} : g(k, x_1^n) = \Omega_k \text{ dla wszystkich } k \leq l \right\}.$$

- Dla procesu Santa Fe mamy wzrost potęgowy

$$\limsup_{n\to\infty} \frac{\operatorname{card} \mathbb{U}(X_1^n)}{n^{1/\alpha}} \in (0,\infty) \text{ prawie na pewno.}$$

Zbiór przewidywalnych faktów — w ogólności

- Niech $(z_k)_{k=1}^{\infty} = (z_1, z_2, z_3, \ldots)$ będzie ustalonym ciągiem algorytmicznie losowym, niekoniecznie równym rozwinięciu binarnemu prawdopodobieństwa stopu.
- Symbole z_k nazywać będziemy niezależnymi faktami.
- Niech $g(k, x_1^n)$ będzie ustaloną funkcją obliczalną.
- Niech x_1^n będzie dowolnym napisem.
- Określmy zbiór niezależnych faktów przewidywalnych z x_1^n:

$$\mathbb{U}(x_1^n) := \left\{ l \in \mathbb{N} : g(k, x_1^n) = z_k \text{ dla wszystkich } k \leq l \right\}.$$

Wykładnik Hilberga

- Wykładnik Hilberga definiujemy jako

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log^+ s(n)}{\log n}, \quad \log^+ x = \begin{cases} \log(x+1), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

  Mamy $\operatorname*{hilb}_{n\to\infty} n^\beta = \beta$ dla $\beta \geq 0$.

- Dla procesu Santa Fe mamy prawie na pewno

$$\operatorname*{hilb}_{n\to\infty} \operatorname{card} \mathbb{U}(X_1^n) = 1/\alpha \in (0,1).$$

- Dla języka naturalnego mamy prawdopodobnie

$$\operatorname*{hilb}_{n\to\infty} \operatorname{card} V_{\mathrm{PPM}}(x_1^n) \approx 0.8.$$

Twierdzenie o faktach i słowach

Na poprzednich slajdach zdefiniowaliśmy:

- $\mathbb{U}(x_1^n)$ — zbiór faktów przewidywalnych z napisu x_1^n.
- $G_{\mathrm{PPM}}(x_1^n)$ — rząd PPM napisu x_1^n.
- $V_{\mathrm{PPM}}(x_1^n)$ — zbiór słów PPM napisu x_1^n.
- Wykładnik Hilberga: $\operatorname{hilb}_{n\to\infty} n^{\beta} = \beta$.

Twierdzenie o faktach i słowach

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończonym alfabecie zachodzi nierówność

$$\operatorname{hilb}_{n\to\infty} \mathbb{E} \operatorname{card} \mathbb{U}(X_1^n) \leq \operatorname{hilb}_{n\to\infty} \mathbb{E} \left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n) \right].$$

Twierdzenie o wykładnikach Hilberga

Twierdzenie

Niech $J(n) := 2G(n) - G(2n)$. Jeżeli $\lim_{n\to\infty} G(n)/n = g$, to

$$\operatorname*{hilb}_{n\to\infty} [G(n) - ng] \leq \operatorname*{hilb}_{n\to\infty} J(n),$$

gdzie równość zachodzi, gdy $J(n) \geq 0$.

- Entropia Shannona $H(X_1^n) := \mathbb{E}\left[-\log P(X_1^n)\right]$.
- Intensywność entropii Shannona $h := \lim_{n\to\infty} H(X_1^n)/n$.
- Informacja wzajemna Shannona

$$\begin{aligned} I(X_1^n; X_{n+1}^{2n}) &:= H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}) \\ &= 2H(X_1^n) - H(X_1^{2n}) \geq 0. \end{aligned}$$

- Stąd $\operatorname{hilb}_{n\to\infty} \left[H(X_1^n) - nh\right] = \operatorname{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n})$.

Zarys dowodu twierdzenia o faktach i słowach

- Niech $\mathbb{H}_{\mathbf{PPM}}(x_1^n) := -\log\left[\sum_{k=-1}^{\infty} \mathbf{PPM}_k(x_1^n)/(k+2)^2\right]$.
- Mamy równość intensywności entropii

$$h = \lim_{n\to\infty} \frac{\mathbb{E}\,\mathbb{H}(X_1^n)}{n} = \lim_{n\to\infty} \frac{\mathbb{E}\,\mathbb{H}_{\mathbf{PPM}}(X_1^n)}{n}.$$

- Oznaczmy informację wzajemną PPM

$$\mathbb{I}_{\mathbf{PPM}}(x_1^n; x_{n+1}^{2n}) := \mathbb{H}_{\mathbf{PPM}}(x_1^n) + \mathbb{H}_{\mathbf{PPM}}(x_{n+1}^{2n}) - \mathbb{H}_{\mathbf{PPM}}(x_1^{2n}).$$

- Twierdzenie o faktach i słowach jest konsekwencją nierówności

$$\begin{aligned}
&\operatorname*{hilb}_{n\to\infty} \mathbb{E}\ \mathbf{card}\,\mathbb{U}(X_1^n) \le \operatorname*{hilb}_{n\to\infty} \left[\mathbb{E}\,\mathbb{H}(X_1^n) - hn\right] \\
&\le \operatorname*{hilb}_{n\to\infty} \left[\mathbb{E}\,\mathbb{H}_{\mathbf{PPM}}(X_1^n) - hn\right] \le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\,\mathbb{I}_{\mathbf{PPM}}(X_1^n; X_{n+1}^{2n}) \\
&\le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\,\left[G_{\mathbf{PPM}}(X_1^n) + \mathbf{card}\, V_{\mathbf{PPM}}(X_1^n)\right].
\end{aligned}$$

Konkluzje

- Przedstawiliśmy twierdzenie o faktach i słowach, które orzeka, że liczba niezależnych faktów opisywanych przez skończony tekst jest mniejsza niż liczba różnych słów w tymże tekście.
- Dla tekstów w języku naturalnym, liczba różnych słów zdaje się rosnąć potęgowo z długością tekstu.
- Czy zatem liczba niezależnych faktów opisywanych przez teksty w języku naturalnym również rośnie potęgowo z długością tekstu?

Literatura

1. G. J. Chaitin, (1975). *A theory of program size formally identical to information theory.* Journal of the ACM, 22:329–340.
2. T. M. Cover and J. A. Thomas, (2006). *Elements of Information Theory, 2nd ed.* New York: John Wiley.
3. Ł. Dębowski, (2018). *Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited.* Entropy, 20(2):85.