Maksymalne powtórzenia w tekstach i zerowa intensywność entropii

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN
Warszawa

Co to jest maksymalne powtórzenie?

Maksymalne powtórzenie (maximal repetition) $L(x_1^n)$ w tekście $x_1^n = (x_1, x_2, ..., x_n)$ to maksymalna długość powtarzającego się podsłowa.

Formalnie,

$$L(x_1^n) := \max \left\{ k : x_{i+1}^{i+k} = x_{j+1}^{j+k} \text{ dla pewnych } 0 \leq i < j \leq n - k \right\}.$$

Przykład:

$$x_1^n = \text{“O szyby deszcz dzwoni, deszcz dzwoni jesienny.”}$$
$$L(x_1^n) = |\text{“deszcz dzwoni”}| = 13.$$

Maksymalne powtórzenie $L(x_1^n)$ można policzyć w czasie $O(n)$ sortując drzewo sufiksów (Kolpakov & Kucherov, 1999).

Z punktu widzenia informatyków... (de Luca, 1999)

Złożoność podsłowna (subword complexity) $f(k|x_1^n)$ to liczba różnych podsłów długości k pojawiających się w tekście x_1^n.

Formalnie,

$$f(k|x_1^n) := \mathbf{card}\left\{y_1^k : x_{i+1}^{i+k} = y_1^k \text{ dla pewnego } 0 \leq i \leq n-k\right\}.$$

Mamy

$$f(k|x_1^n) \text{ jest ściśle rosnące dla } k \leq L(x_1^n),$$
$$f(k|x_1^n) = n - k + 1 \text{ dla } k > L(x_1^n).$$

Złożoność podsłowna $f(k|x_1^n)$ osiąga maksimum dla $k = L(x_1^n)$.

Z punktu widzenia probabilistów... (Erdős & Rényi, 1970)

Niech $(X_i)_{i=1}^{\infty}$ będzie procesem IID, tzn. nieskończonym ciągiem niezależnych zmiennych losowych o identycznym rozkładzie,

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} p(x_i).$$

Można wówczas udowodnić, że istnieje taka stała $A > 0$, że

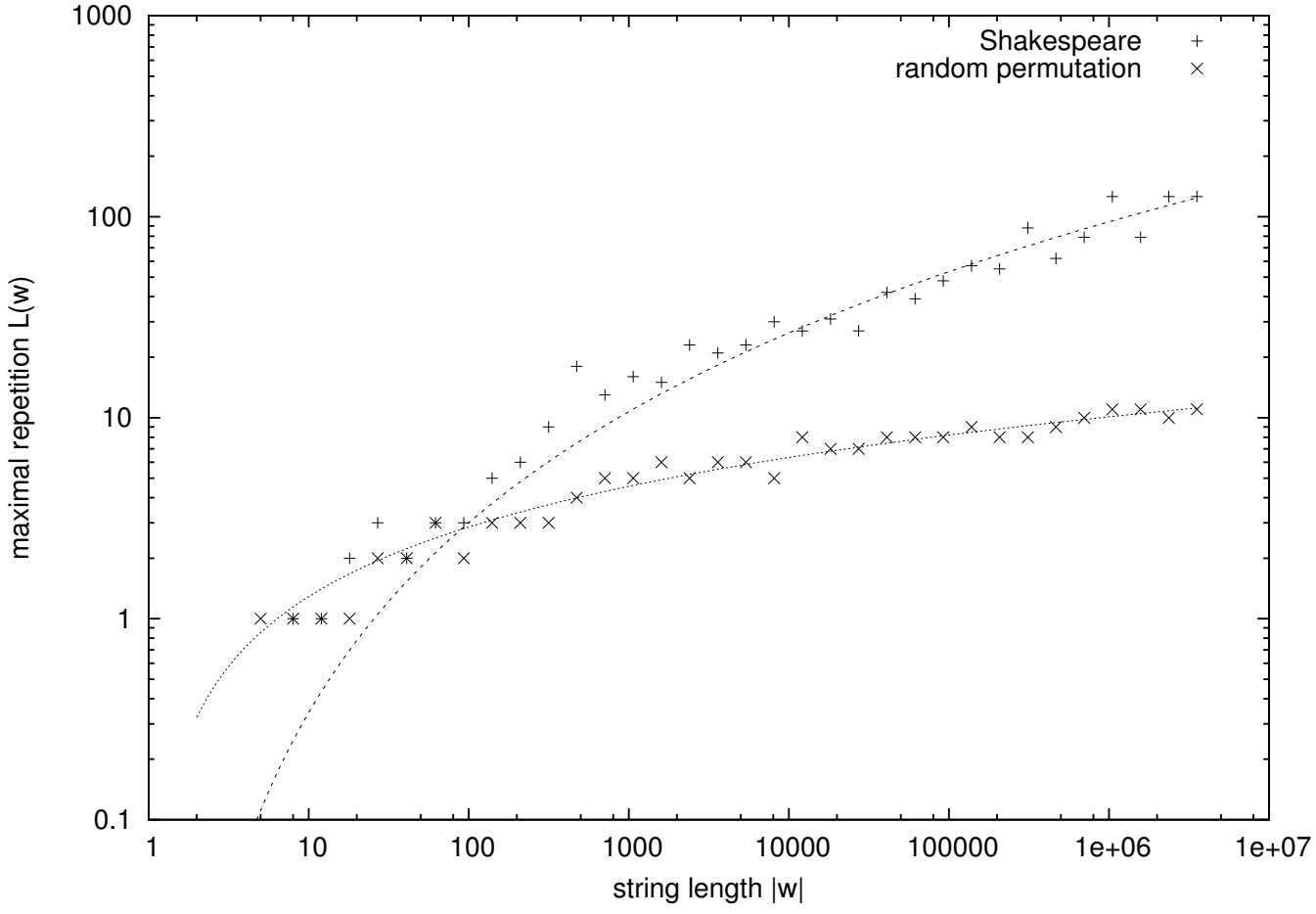
$$L(X_1^n) \leq A \log n$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Inaczej pisząc,

$$P\left(\limsup_{n\to\infty} \frac{L(X_1^n)}{\log n} \leq A\right) = 1.$$

A w odniesieniu do języka... (Dębowski, 2015)



$L(x_1^n) \approx 0.02498\,(\log n)^{3.136}$ dla tekstu w języku angielskim.

$L(x_1^n) \approx 0.4936\,(\log n)^{1.150}$ dla losowej permutacji znaków.

Dwa pytania otwarte

1. Jak szeroka jest klasa procesów stochastycznych, dla których

$$L(X_1^n) \approx A(\log n)^\alpha$$

zachodzi dla dostatecznie dużych n z prawdopodobieństwem 1 dla zadanego $\alpha \geq 1$?

2. Czy istnieją wśród nich procesy mogące służyć jako lepsze statystyczne modele języka naturalnego niż np. ukryte procesy Markowa, używane obecnie?

Rodzina entropii Renyiego

Wzór dla $\gamma \in (0,1) \cup (1,\infty)$:

$$H_\gamma(X) := \frac{1}{1-\gamma} \log \sum_x P(X=x)^\gamma.$$

Przypadki szczególne:

$$H_0(X) := \log \operatorname{card} \{x : P(X=x) > 0\} \quad \text{(entropia Hartleya)},$$

$$H_1(X) := -\sum_x P(X=x) \log P(X=x) \quad \text{(entropia Shannona)},$$

$$H_2(X) := -\log \sum_x P(X=x)^2 \quad \text{(entropia kolizji)},$$

$$H_\infty(X) := -\log \max_x P(X=x) \quad \text{(min-entropia)}.$$

$H_\gamma(X) \geq H_\delta(X)$ dla $\gamma < \delta$; $H_\gamma(X) \leq \frac{\gamma}{\gamma-1} H_\infty(X)$ dla $\gamma > 1$.

Ograniczenie dolne (Dębowski, 2015)

Intuicja:
Jeżeli dopuszczalnych kombinacji symboli, z których składają się teksty, jest mało, to w tekstach tych powtórzenia są długie.

Formalnie, zdefiniujmy entropię Hartleya bloku długości n,

$$H_0(n) := \log \operatorname{card} \left\{ x_1^n : P(X_1^n = x_1^n) > 0 \right\}.$$

Jeżeli dla stacjonarego procesu stochastycznego $(X_i)_{i=1}^{\infty}$ zachodzi

$$H_0(n) \leq Bn^{\beta}$$

dla $B > 0$ i $0 < \beta \leq 1$, to dla $A < B^{-\alpha}$ i $\alpha = 1/\beta$ mamy

$$L(X_1^n) \geq A\,(\log n)^{\alpha}$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Szkic dowodu

1. Tekst X_1^n zawiera $n - k + 1$ podsłów długości k. Z prawdopodobieństwem 1, wsród nich może być co najwyżej $\exp H_0(k)$ różnych podsłów. W rezultacie, jeżeli $\exp H_0(k) < n - k + 1$, to $L(X_1^n) \geq k$ zachodzi z prawdopodobieństwem 1.
2. Załóżmy teraz, że $H_0(k) \leq Bk^\beta$. Wówczas $L(X_1^n) \geq k$ dla $Bk^\beta < \log(n - k + 1) \approx \log n$. Warunek ten zachodzi dla $k \approx A (\log n)^{1/\beta}$, gdzie $A \approx B^{-1/\beta}$.

Wzmocnione ograniczenie dolne

Zdefiniujmy entropię Shannona bloku długości n,

$$H_1(n) := -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n) \leq H_0(n).$$

Jeżeli dla stacjonarego procesu stochastycznego $(X_i)_{i=1}^{\infty}$ zachodzi

$$H_1(n) \leq Bn^{\beta}$$

dla $B > 0$ i $0 < \beta \leq 1$, to dla $A < B^{-\alpha}/2$ i $\alpha < 1/\beta$ mamy

$$L(X_1^n) \geq A\,(\log n)^{\alpha}$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Dowód wykorzystuje pojęcie złożoności podsłownej.

Ograniczenie górne (Shields, 1997)

Intuicja:
Jeżeli dopuszczalnych kombinacji symboli, z których składają się teksty, jest dużo, to w tekstach tych powtórzenia są krótkie.

Formalnie, zdefiniujmy warunkową min-entropię bloku długości n,

$$H_\infty^{cond}(n) := -\log \sup_{x_1^{m+n}} P(X_{m+1}^{m+n} = x_{m+1}^{m+n} | X_1^m = x_1^m).$$

Jeśli proces stochastyczny $(X_i)_{i=1}^\infty$ spełnia

$$H_\infty^{cond}(n) \geq Bn^\beta$$

dla $B > 0$ i $0 < \beta \leq 1$, to dla $A > 3^\alpha B^{-\alpha}$ i $\alpha = 1/\beta$ mamy

$$L(X_1^n) < A\,(\log n)^\alpha$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Szkic dowodu

Mamy

$$
\begin{aligned}
&P(L(X_1^n) \geq k) \\
&= P\left(X_{i+1}^{i+k} = X_{j+1}^{j+k} \text{ dla pewnych } 0 \leq i < j \leq n-k\right) \\
&\leq \sum_{0 \leq i < j \leq n-k} P(X_{i+1}^{i+k} = X_{j+1}^{j+k}) \\
&= \sum_{0 \leq i < j \leq n-k} \sum_{x_1^{j-1}} P(X_1^{j-1} = x_1^{j-1}) P(X_{j+1}^{j+k} = x_{i+1}^{i+k} | X_1^{j-1} = x_1^{j-1}) \\
&\leq \sum_{0 \leq i < j \leq n-k} \exp(-H_\infty^{cond}(k)) \\
&\leq n^2 \exp(-Bk^\beta).
\end{aligned}
$$

A zatem teza wynika z lematu Borela-Cantellego.

Odpowiednik reguły łańcuchowej dla min-entropii

Zdefinujmy

- bezwarunkową min-entropię bloku

$$H_\infty(n) := -\log \sup_{x_{m+1}^{m+n}} P(X_{m+1}^{m+n} = x_{m+1}^{m+n}),$$

- warunkową min-entropię bloku

$$H_\infty^{cond}(n) := -\log \sup_{x_1^{m+n}} P(X_{m+1}^{m+n} = x_{m+1}^{m+n} | X_1^m = x_1^m),$$

- min-informację wzajemną bloku

$$I_\infty(n) := \log \sup_{x_1^{m+n}} \frac{P(X_{m+1}^{m+n} = x_{m+1}^{m+n})}{P(X_{m+1}^{m+n} = x_{m+1}^{m+n}) P(X_1^m = x_1^m)}.$$

Mamy

$$H_\infty(n) \leq H_\infty^{cond}(n) + I_\infty(n).$$

Intensywności entropii i entropia nadwyżkowa

Zdefiniujmy

- intensywność bezwarunkowej min-entropii

$$h_\infty := \liminf_{n\to\infty} \frac{H_\infty(n)}{n},$$

- intensywność warunkowej min-entropii

$$h_\infty^{cond} := \liminf_{n\to\infty} \frac{H_\infty^{cond}(n)}{n},$$

- min-entropię nadwyżkową

$$E_\infty := \limsup_{n\to\infty} I_\infty(n).$$

Mamy

$$h_\infty > 0 \text{ i } h_\infty^{cond} = 0 \implies E_\infty = \infty.$$

Maksymalne powtórzenie a globalne miary zależności

Dla języka naturalnego mamy

$$L(X_1^n) \geq A\,(\log n)^{\alpha}\,,\ \alpha \approx 3 \implies H_{\infty}^{cond}(n) \leq Bn^{\beta},\ \beta \approx 1/3$$
$$\implies h_{\infty}^{cond} = 0.$$

Ponadto dla języka naturalnego prawdopodobnie $h_{\infty} > 0$.
(Intensywność entropii Shannona to około jeden bit na literę.)

Stąd

$$h_{\infty} > 0 \text{ i } h_{\infty}^{cond} = 0 \implies E_{\infty} = \infty.$$

Język miałby zatem nieskończoną min-entropię nadwyżkową.
(Hilberg 1990: ma nieskończoną entropię nadwyżkową Shannona.)

Procesy o skończonej energii

Proces $(X_i)_{i=1}^{\infty}$ nazywa się procesem o skończonej energii, gdy

$$P(X_{m+1}^{m+n} = x_{m+1}^{m+n} | X_1^m = x_1^m) \leq c^n, \quad c < 1.$$

Procesy takie spełniają $h_{\infty}^{cond} > 0$, a zatem

$$L(X_1^n) \leq A \log n$$

dla dostatecznie dużych n z prawdopodobieństwem 1.

Stacjonarne procesy o skończonej energii mają dodatnią intensywność entropii Shannona

$$h_1 := \lim_{n \to \infty} \frac{H_1(n)}{n} > 0.$$

Przykłady procesów o skończonej energii (I)

Proces $(Y_i)_{i=1}^{\infty}$ nazywa się ukrytym procesem Markowa, jeżeli

$$Y_i = f(X_i)$$

dla pewnej funkcji f oraz dyskretnego procesu Markowa $(X_i)_{i=1}^{\infty}$,

$$P(X_1^n = x_1^n) = \pi(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1}).$$

Ukrytymi procesami Markowa są m.in. procesy Markowa i IID.

Ukryty proces Markowa $(Y_i)_{i=1}^{\infty}$ jest procesem o skończonej energii, jeśli

$$c := \sup_{y,x} P(Y_i = y|X_{i-1} = x) < 1.$$

Przykłady procesów o skończonej energii (II)

Niech $(\mathbb{X}, *)$ będzie grupą. Proces $(X_i)_{i=-\infty}^{\infty}$ nad alfabetem $\mathbb{X}$ nazywa się jednostajnie zaszumionym, jeżeli

$$X_i = W_i * Z_i,$$

gdzie $(W_i)_{i=-\infty}^{\infty}$ jest dowolnym procesem nad alfabetem $\mathbb{X}$, zaś $(Z_i)_{i=-\infty}^{\infty}$ jest niezależnym procesem IID spelniającym

$$c := \max_a P(Z_i = a) < 1.$$

Procesy jednostajnie zaszumione są procesami o skończonej energii.

Regularne procesy Hilberga

Proces $(X_i)_{i=1}^{\infty}$ nazywam regularnym procesem Hilberga, gdy

$$B_1 n^{\beta} \leq \log f(n|X_1^{\infty}) \leq B_2 n^{\beta},$$
$$A_2 (\log n)^{1/\beta} \leq L(X_1^n) \leq A_1 (\log n)^{1/\beta}$$

dla dostatecznie dużych n i pewnych $B_i, A_i > 0$ oraz $0 < \beta < 1$ z prawdopobieństwem 1.

Stacjonarne regularne procesy Hilberga mają zerową intensywność entropii Shannona

$$h_1 := \lim_{n\to\infty} \frac{H_1(n)}{n} = 0.$$

Procesy RHA (random hierachical association)

1. Weźmy liczby naturalne $k_0, k_1, \ldots$, gdzie $k_{n-1} \leq k_n \leq k_{n-1}^2$.
2. Dla każdego n, niech $(L_{nj}, R_{nj})_{j \in \{1,\ldots,k_n\}}$ będzie niezależną losową kombinacją k_n różnych par liczb z $\{1, \ldots, k_{n-1}\}$.
3. Zdefiniujmy zmienne losowe:

$$Y_j^0 = j, \qquad j \in \{1, \ldots, k_0\},$$
$$Y_j^n = Y_{L_{nj}}^{n-1} \times Y_{R_{nj}}^{n-1}, \qquad j \in \{1, \ldots, k_n\}, n \in \mathbb{N}.$$

4. Niech $C_0, C_1, \ldots$ będą niezależnymi zmiennymi o rozkładzie:

$$P(C_n = j) = 1/k_n, \qquad j \in \{1, \ldots, k_n\}.$$

5. Zdefiniujmy proces RHA:

$$X_1 \times X_2 \times X_3 \times \ldots := Y_{C_0}^0 \times Y_{C_1}^1 \times Y_{C_2}^2 \times \ldots.$$

Przykład regularnego procesu Hilberga

Dla parametrów

$$k_n = \left\lfloor \exp\left(2^{\beta n}\right) \right\rfloor,$$

gdzie $0 < \beta < 1$, średnia stacjonarna procesu RHA jest regularnym procesem Hilberga o parametrze β, tzn.

$$B_1 n^{\beta} \leq \log f(n|X_1^{\infty}) \leq B_2 n^{\beta},$$
$$A_2 (\log n)^{1/\beta} \leq L(X_1^n) \leq A_1 (\log n)^{1/\beta}.$$

Modele języka naturalnego? — nadal wyzwaniem

1. Ukryte procesy Markowa, klasa modeli powszechnie stosowanych w lingwistyce komputerowej, są procesami o skończonej energii.
2. Język naturalny nie jest procesem o skończonej energii, gdyż maksymalne powtórzenie w tekstach rośnie szybciej niż logarytmicznie.
3. Istnieją procesy RHA, dla których maksymalne powtórzenie rośnie jak w języku naturalnym.
4. Jednak procesy RHA również nie są dobrymi modelami języka, gdyż mają zerową intensywność entropii Shannona — w przeciwieństwie do języka naturalnego.
5. Lepsze modele języka to prawdopodobnie procesy o dodatniej intensywności entropii Shannona i zerowej intensywności warunkowej min-entropii.