Praktyczne i teoretyczne problemy statystycznego modelowania języka naturalnego

Część II: Maksymalne powtórzenie

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

XLIII Konferencja „Statystyka Matematyczna”
Będlewo, 04–08.12.2017

Statystyczne prawa językowe

Teksty w języku naturalnym spełniają przybliżone prawa ilościowe:

1. **Prawo Zipfa:** częstość słowa jest odwrotnie proporcjonalna do rangi słowa.
2. **Prawo Heapsa:** liczba różnych słów w tekście rośnie potęgowo z długością tekstu.
3. **Intensywność entropii Shannona:** jest rzędu 1 bita na literę.
4. **Hipoteza Hilberga:** entropia warunkowa litery maleje potęgowo z długością kontekstu.
5. **Prawo kodu PPM:** liczba różnych „słów" wykrywanych przez algorytm PPM w tekście rośnie potęgowo z długością tekstu.
6. **Prawo maksymalnego powtórzenia:** długość maksymalnego powtórzenia rośnie jak sześcian logarytmu długości tekstu.

Czy można coś wywnioskować o języku jako procesie stochastycznym na podstawie tych obserwacji/hipotez?

Pytania matematyka

1. Czy istnieje idealny probabilistyczny model języka?
2. Czy model ten może być modelem Markowa?
3. Czy model ten może być ukrytym modelem Markowa?
4. Czy model ten jest ergodyczny?
5. Czy model ten jest stacjonarny?
6. Czy model ten jest asymptotycznie średnio stacjonarny?
7. Czy model ten jest kodem uniwersalnym?
8. Czy model ten jest efektywnie obliczalny?

Co to jest maksymalne powtórzenie?

Maksymalne powtórzenie (maximal repetition) $L(x_1^n)$ w tekście $x_1^n = (x_1, x_2, \ldots, x_n)$ 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:

$$x_1^n = \text{“O szyby deszcz dzwoni, deszcz dzwoni jesienny.”}$$
$$L(x_1^n) = |\text{“ deszcz dzwoni”}| = 14.$$

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

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

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

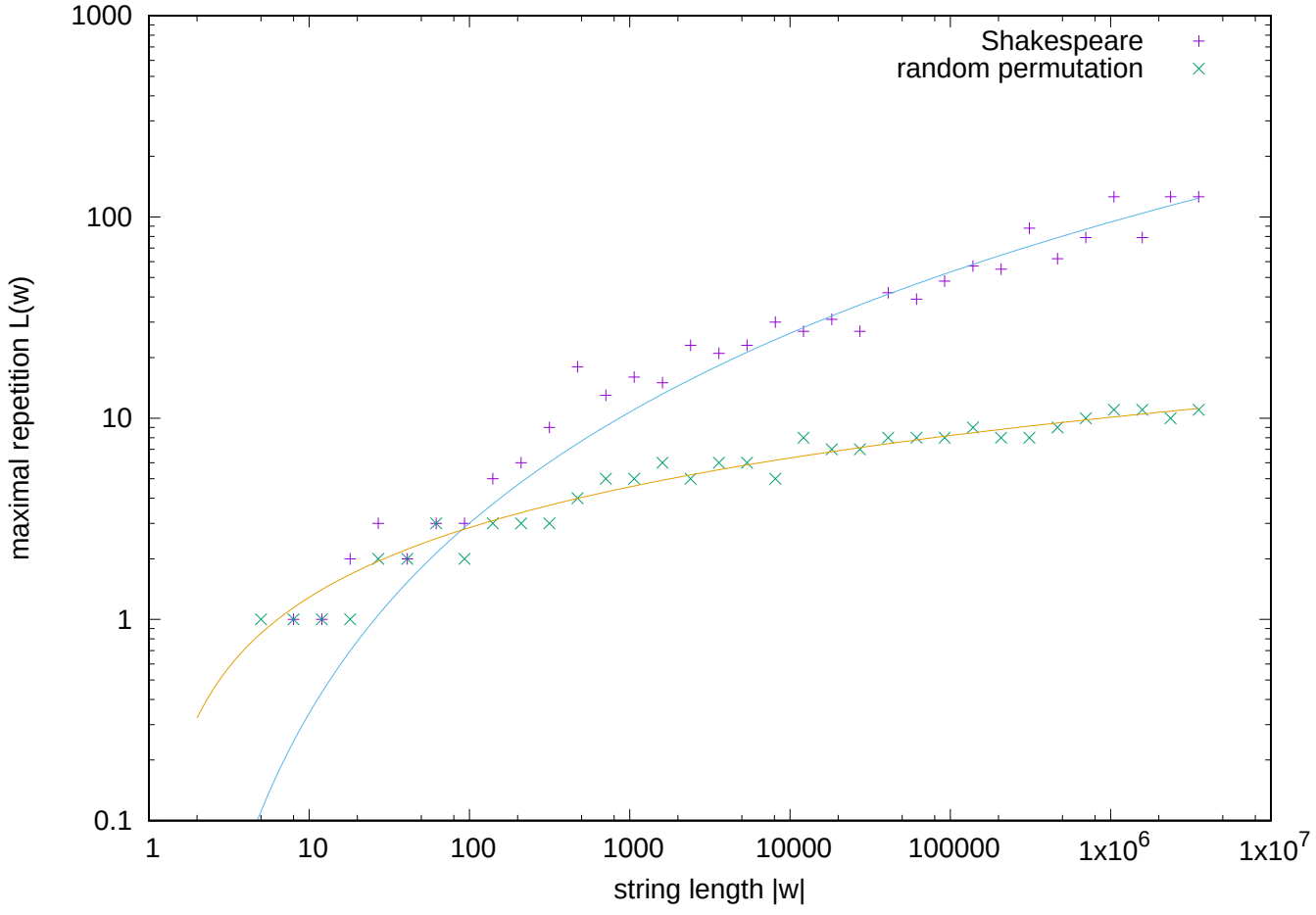
$$L(X_1^n) \leq A \log n$$

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 Rényiego

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,

$$H_0(n) := \log \operatorname{card} \left\{ x_1^n : P(X_{m+1}^{m+n} = x_1^n) > 0 \text{ dla pewnego } m \geq 0 \right\}.$$

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

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

dla pewnego $B > 0$ 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 $A \approx B^{-1/\beta}$.

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

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

dla pewnego $B > 0$, to dla $A > 3B^{-1}$ mamy

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

dla dostatecznie dużych n z prawdopodobieństwem 1.

Szkic dowodu

Mamy

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

A zatem teza wynika z lematu Borela-Cantellego.

Maksymalne powtórzenie i intensywności entropii

Mamy

$$H_\infty^{cond}(m+n) \geq H_\infty^{cond}(m) + H_\infty^{cond}(n),$$

więc z lematu Feketego dla funkcji superaddytywnej wynika

$$\lim_{n\to\infty} \frac{H_\infty^{cond}(n)}{n} = \sup_{n\geq 0} \frac{H_\infty^{cond}(n)}{n}.$$

Zatem

$$\limsup_{n\to\infty} \frac{H_0(n)}{n} = 0 \implies \liminf_{n\to\infty} \frac{L(X_1^n)}{\log n} = \infty \text{ a.s.},$$

$$\limsup_{n\to\infty} \frac{L(X_1^n)}{\log n} = \infty \text{ a.s.} \implies H_\infty^{cond}(n) = 0, \quad n \geq 1.$$

„Wzmocnione” ograniczenie dolne

Zdefiniujmy entropię Shannona bloku długości n,

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

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

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

dla pewnego $B > 0$ i $0 < \beta \leq 1$, to dla $\alpha < 1/\beta$ mamy

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

dla dostatecznie dużych n z prawdopodobieństwem 1.

Dowód wykorzystuje pojęcie czasu powrotu.

Wzmocnione ograniczenie górne

Zdefiniujmy warunkową entropię Rényiego bloku długości n,

$$H_\gamma^{cond}(n) := \frac{1}{1-\gamma} \log \mathbb{E} \sum_{x_1^n} P(X_1^n = x_1^n | X_{-\infty}^0)^\gamma \geq H_\infty^{cond}(n).$$

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

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

dla pewnego $\gamma > 1$, $B > 0$ i $0 < \beta \leq 1$, to dla $A > \left[\gamma \cdot \frac{\gamma+1}{\gamma-1}\right]^\alpha B^{-\alpha}$ i $\alpha = 1/\beta$ mamy

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

dla dostatecznie dużych n z prawdopodobieństwem 1.

Dowód wykorzystuje pojęcie czasu powrotu.

Maksymalne powtórzenie i intensywności entropii II

Intensywności entropii Shannona i Rényiego:

$$h_1 := \lim_{n\to\infty} \frac{H_1(n)}{n}, \qquad h_\gamma^{cond} := \liminf_{n\to\infty} \frac{H_\gamma^{cond}(n)}{n}.$$

Mamy

$$h_1 = 0 \implies \liminf_{n\to\infty} \frac{L(X_1^n)}{(\log n)^\alpha} = \infty \text{ a.s.}, \quad \alpha < 1,$$

$$\limsup_{n\to\infty} \frac{L(X_1^n)}{\log n} = \infty \text{ a.s.} \implies h_\gamma^{cond} = 0, \quad \gamma > 1.$$

Dla języka naturalnego prawdopodobnie $h_1 > 0$ i $h_\gamma^{cond} = 0$.

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

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

Dowód

Niech $(Y_i)_{i=1}^{\infty}$ będzie ukrytym procesem Markowa. Z warunkowej niezależności Y_{m+1} i Y_1^m względem X_m wynika, że

$$\begin{aligned}
&P\left(Y_{m+1} = y_{m+1} | Y_1^m = y_1^m\right) \\
&= \sum_{x_m} P\left(Y_{m+1} = y_{m+1} | X_m = x_m\right) P\left(X_m = x_m | Y_1^m = y_1^m\right) \\
&\leq \sum_{x_m} cP\left(X_m = x_m | Y_1^m = y_1^m\right) = c.
\end{aligned}$$

Zatem $(Y_i)_{i=-\infty}^{\infty}$ jest procesem o skończonej energii, gdy $c < 1$.

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.

Wynik Shieldsa (1992)

Dla dowolnego stacjonarnego procesu ergodycznego $(X_i)_{i=-\infty}^{\infty}$ i funkcji $\lambda(n) = o(n)$, istnieje funkcja mierzalna $f : \mathbb{X}^{\mathbb{Z}} \to \mathbb{X}$, że proces $(Y_i)_{i=-\infty}^{\infty}$, gdzie $Y_i := f((X_{i+j})_{j=-\infty}^{\infty})$, spełnia

$$\limsup_{n\to\infty} \frac{L(Y_1^n)}{\lambda(n)} \geq 1 \text{ prawie na pewno,}$$

a intensywność entropii Shannona procesu $(Y_i)_{i=-\infty}^{\infty}$ jest dowolnie bliska intensywności entropii Shannona procesu $(X_i)_{i=-\infty}^{\infty}$. Ponadto, jeżeli $(X_i)_{i=-\infty}^{\infty}$ jest procesem IID, to proces $(Y_i)_{i=-\infty}^{\infty}$ jest mieszający.

Konkluzje

1. Ukryte procesy Markowa, klasa modeli powszechnie stosowanych w inżynierii lingwistycznej, są procesami o skończonej energii.
2. Język naturalny nie jest procesem o skończonej energii, gdyż maksymalne powtórzenie w tekstach rośnie szybciej niż logarytmicznie.
3. Lepsze modele języka to prawdopodobnie procesy o dodatniej intensywności entropii Shannona i zerowej intensywności warunkowej entropii Rényiego rzędu $\gamma > \mathbf{1}$.