Maksymalne powtórzenia w tekście i hipotezy probabilistyczne

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN
Warszawa

2 (4) Warsztat badawczy i Perypatetyczna Konferencja
„Modelowanie Systemów Poznawczych”
Zakopane, 3-5.10.2014

Teksty — między losowością a determinizmem

If a Martian scientist sitting before his radio in Mars accidentally received from Earth the broadcast of an extensive speech which he recorded perfectly through the perfection of Martian apparatus and studied at his leisure, what criteria would he have to determine whether the reception represented the effect of animate process on Earth, or merely the latest thunderstorm on Earth? It seems that the only criteria would be the arrangement of occurrences of the elements, and the only clue to the animate origin would be this: the arrangement of the occurrences would be neither of rigidly fixed regularity such as frequently found in wave emissions of purely physical origin nor yet a completely random scattering of the same.

— *George Kingsley Zipf (1965:187)*

Dwie hipotezy probabilistyczne

Hipoteza skończonej energii:

Proces generowania tekstu jest procesem o skończonej energii.

Hipoteza skończonej energii jest spełniona, jeżeli teksty są zaszumione w pewien dość ogólny sposób.

Mocna hipoteza Hilberga:

Proces generowania tekstu jest mocnym procesem Hilberga.

Zgodnie z hipotezą Hilberga tylko niewielka część możliwych tekstów podlega replikacji. ($\Longrightarrow$ ewolucja memetyczna?)

Hipoteza skończonej energii

- $(\mathbf{X_i})_{i\in\mathbb{Z}}$ — proces stochastyczny — ciąg zmiennych losowych $\mathbf{X_i : \Omega \to \mathbb{X}}$ na przestrzeni probabilistycznej $\mathbf{(\Omega, \mathcal{J}, P)}$.
- $\mathbf{X_k^l = (X_k, X_{k+1}, ..., X_l)}$ — bloki zmiennych losowych.
- $\mathbf{|u|}$ — długość napisu $\mathbf{u}$.

Definicja (Shields 1997)

Proces $(\mathbf{X_i})_{i\in\mathbb{Z}}$ nazywamy **procesem o skończonej energii**, jeżeli prawdopodobieństwo warunkowe bloków spełnia

$$\mathbf{P}\left(\mathbf{X_{i+|w|+1}^{i+|wu|}} = \mathbf{u} \middle| \mathbf{X_{i+1}^{i+|w|}} = \mathbf{w}\right) \leq \mathbf{Kc^{|u|}}$$

dla pewnych stałych $\mathbf{0 < c < 1}$ oraz $\mathbf{K > 0}$.

Mocna hipoteza Hilberga

Definicja

Dla zmiennej losowej $\mathbf{X}$, definiuje się **entropię topologiczną** jako logarytm liczby wartości o niezerowym prawdopodobieństwie,

$$\mathbf{H_{top}(X) = \log \operatorname{card} \{x : P(X = x) > 0\}}.$$

Definicja

Proces $\mathbf{(X_i)_{i\in\mathbb{Z}}}$ nazywamy **mocnym procesem Hilberga**, jeżeli

$$\mathbf{H_{top}(X_{i+1}^{i+n}) \leq An^{\beta}}$$

dla pewnych stałych $\mathbf{0 < \beta < 1}$ oraz $\mathbf{A > 0}$.

Maksymalne powtórzenie

Definicja

Dla tekstu $\mathbf{w}$ definiuje się **maksymalne powtórzenie** jako

$$\mathbf{L(w) := \max\{|s| : w = x_1 s y_1 = x_2 s y_2 \text{ i } x_1 \neq x_2\}},$$

gdzie $\mathbf{s}$, $\mathbf{x_i}$ oraz $\mathbf{y_i}$ są fragmentami tekstu $\mathbf{w}$.

Dwa twierdzenia

Twierdzenie 1 (Shields 1997)

Dla procesu o skończonej energii $(X_i)_{i\in\mathbb{Z}}$ istnieje $C > 0$ takie, że

$$L(X_1^m) \leq C \log m$$

zachodzi z prawdopodobieństwem 1 dla dostatecznie dużych m.

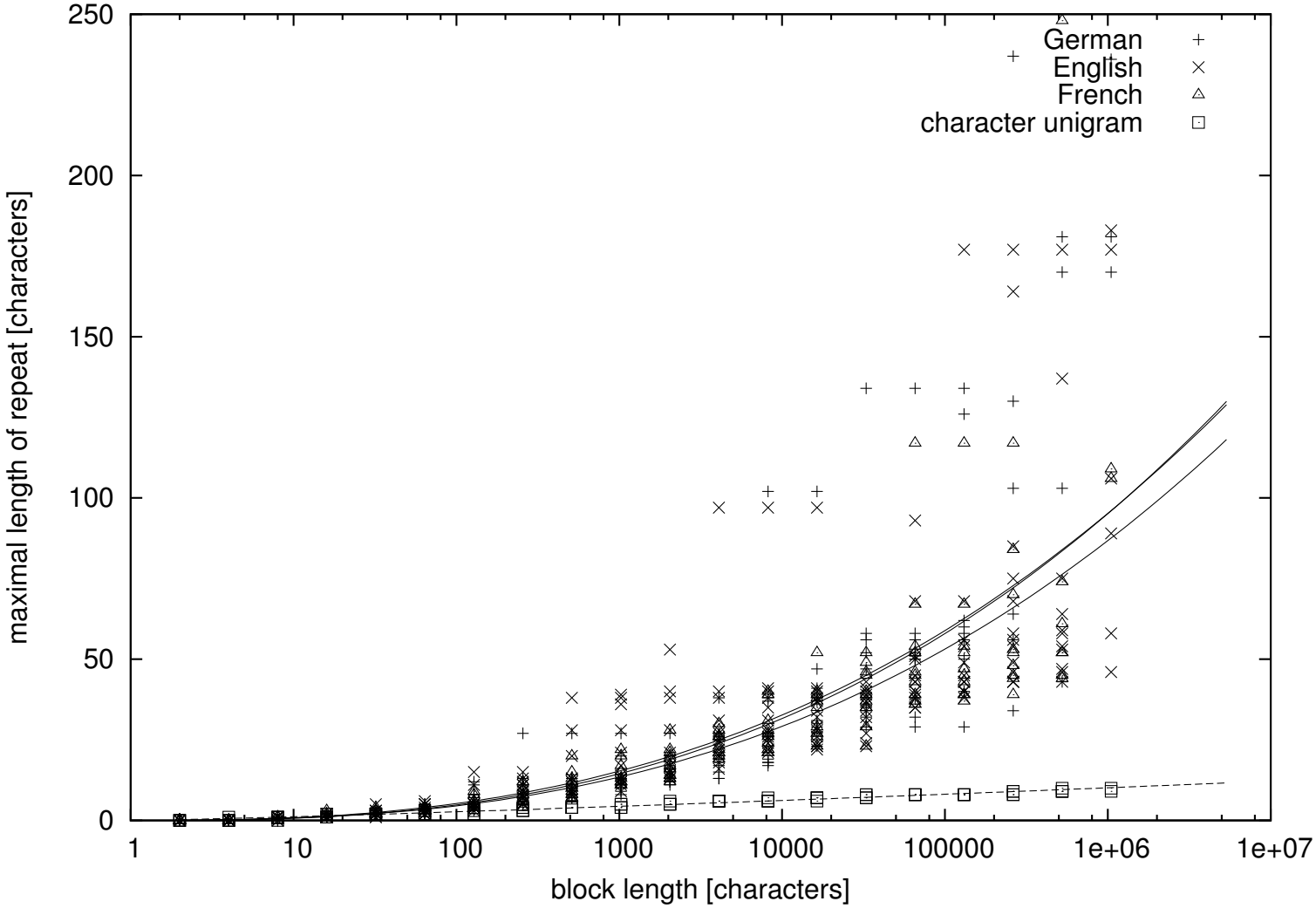
Twierdzenie 2

Dla mocnego procesu Hilberga $(X_i)_{i\in\mathbb{Z}}$ istnieje $C > 0$ takie, że

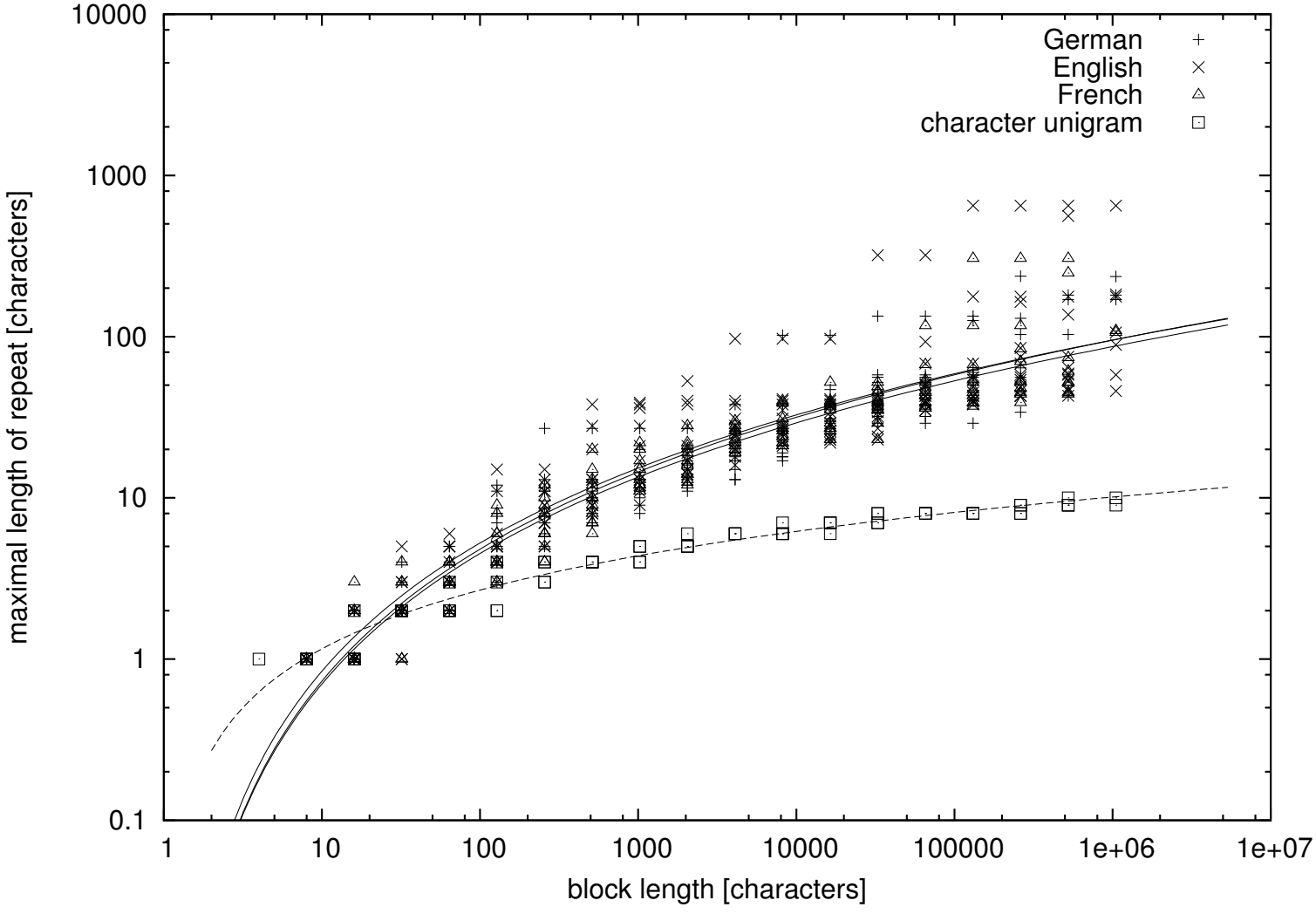
$$L(X_1^m) \geq C(\log m)^{1/\beta}$$

zachodzi z prawdopodobieństwem 1.

35 tekstów w 3 językach + tekst unigramowy



To samo w skali podwójnie logarytmicznej



Parametry modelu

$$\mathbf{L(X_1^m) \approx A(\log m)^{\alpha}}$$

klasa tekstów	**A**	α
j. niemiecki	$\mathbf{0.076 \pm 0.011}$	$\mathbf{2.71 \pm 0.07}$
j. angielski	$\mathbf{0.093 \pm 0.012}$	$\mathbf{2.64 \pm 0.06}$
j. francuski	$\mathbf{0.074 \pm 0.009}$	$\mathbf{2.69 \pm 0.06}$
unigramowe	$\mathbf{0.42 \pm 0.03}$	$\mathbf{1.21 \pm 0.03}$

Przykłady maksymalnych powtórzeń

tekst	maksymalny powtórzony napis
Thomas Mann, Buddenbrooks	“Mit raschen Schritten, die Arme ausgebreitet und den Kopf zur Seite geneigt, in der Haltung eines Mannes, welcher sagen will: Hier bin ich! Töte mich, wenn du willst!”
Jules Verne, Vingt mille lieues sous les mers	“à la partie supérieure de la coque du «Nautilus», et”
Jonathan Swift, Gulliver’s Travels	“of meat and drink sufficient for the support of 1724”
tekst unigramowy	“e t tloeu ”

Wnioski

1. W tekstach w języku naturalnym maksymalne powtórzenie rośnie znacznie szybciej niż w tekstach unigramowych.
2. Obserwacja ta falsyfikuje hipotezę skończonej energii.
3. A zatem teksty nie są zaszumione.
4. Obserwowany wzrost maksymalnego powtórzenia może być natomiast konsekwencją mocnej hipotezy Hilberga.
5. Zgodnie z hipotezą Hilberga tylko niewielka część możliwych tekstów podlega replikacji. ($\Longrightarrow$ ewolucja memetyczna?)

`www.ipipan.waw.pl/~ldebowsk`

Post scriptum: Losowość a dowolność

El universo (que otros llaman la Biblioteca) se componte de un número indefinido, y tal vez infinito, de galerías hexagonales, con vastos pozos de ventilación en el medio, cercados por barandas bajísimas. Desde cualquier hexágono se ven los pisos inferiores y superiores: interminablemente. La distribución de las galerías es invariable. Veinte anaqueles, a cinco largos anaqueles por lado, cubren todos los lados menos dos; su altura, que es la de los pisos, excede apenas la de un bibliotecario normal. Una de las caras libres da a un angosto zaguán, que desemboca en otra galería, idéntica a la primera y a todas. ...

— *La Biblioteca de Babel, Jorge Luis Borges*