Ograniczenie rozmiaru słownika uniwersalnych kodów gramatykowych

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Lingwistyczna motywacja problemu

Rozpatrujemy teksty w języku naturalnym (np. w j. polskim):

- **A** — liczba różnych słów w tekście,
- **N** — długość tekstu w słowach.

Wzór empiryczny (**„prawo Guirauda-Herdana-Heapsa”**):

$$\mathbf{A} \propto \mathbf{N}^{\alpha},$$

gdzie α waha się między **0.5** a **1** w zależności od zbioru tekstów.

- *Władysław Kuraszkiewicz, Józef Łukaszewicz (1951),*
- *Pierre Guiraud (1954),*
- *Gustav Herdan (1964),*
- *H. S. Heaps (1978).*

Przełożenie na problem z teorii informacji

Słowa w tekście to w znacznej części powtarzające się napisy:

- liczba różnych słów $\Rightarrow$ liczba napisów powtarzających się.

Modelowanie probabilistyczne tekstu wg hipotezy Hilberga (1990):

- tekst w języku naturalnym $\Rightarrow$ tekst wylosowany z rozkładu stacjonarnego scharakteryzowanego parametrem globalnej zależności, którym jest informacja wzajemna.

Problem

Ograniczyć w terminach informacji wzajemnej oczekiwaną liczbę powtórzeń w tekście wylosowanym z rozkładu stacjonarnego.

Metoda

Wykorzystać problem najmniejszej gramatyki (CFG) dla tekstu.

Gramatyka bezkontekstowa generująca jeden tekst

$$
\mathbf{G} = \left\{
\begin{array}{l}
\mathbf{A_1} \longrightarrow \mathbf{A_2 A_2 A_4 A_9 A_4 A_3 A_7 A_5} \\
\mathbf{A_2} \longrightarrow \mathbf{A_6 A_9 A_6 A_3} \\
\mathbf{A_3} \longrightarrow \mathbf{A_9 A_7 A_8\,, A_5} \\
\mathbf{A_4} \longrightarrow \texttt{Jeszcze\_raz} \\
\mathbf{A_5} \longrightarrow \mathbf{A_8}\texttt{\_nam.\_} \\
\mathbf{A_6} \longrightarrow \texttt{Sto\_lat} \\
\mathbf{A_7} \longrightarrow \texttt{Niech} \\
\mathbf{A_8} \longrightarrow \texttt{\_żyje} \\
\mathbf{A_9} \longrightarrow \texttt{!\_}
\end{array}
\right\}
$$

Sto lat! Sto lat! Niech żyje, żyje nam.
Sto lat! Sto lat! Niech żyje, żyje nam.
Jeszcze raz! Jeszcze raz! Niech żyje, żyje nam.
Niech żyje nam.

Rozmiar słownika i długość gramatyki

Skrócony zapis gramatyki:

$$\mathbf{G} = \left\{ \begin{array}{l} \mathbf{A_1} \rightarrow \boldsymbol{\alpha_1}, \\ \mathbf{A_2} \rightarrow \boldsymbol{\alpha_2}, \\ \ldots, \\ \mathbf{A_n} \rightarrow \boldsymbol{\alpha_n} \end{array} \right\},$$

$$\mathbb{V}[\mathbf{G}] := \mathbf{n} \quad \text{(rozmiar słownika)},$$

$$|\mathbf{G}| := \textstyle\sum_{\mathbf{i}} |\boldsymbol{\alpha_i}| \quad \text{(długość gramatyki)}.$$

Symbol $\mathbf{A_1}$ ustalony jako symbol startowy.

Najmniejsza gramatyka dla tekstu to $\mathbf{G}$ o najmniejszym $|\mathbf{G}|$, generująca ten tekst jako jedyną produkcję.

- Moses Charikar, Eric Lehman, ..., Abhi Shelat (2005),
- Wojciech Rytter (2003),
- John C. Kieffer, Enhui Yang (2000).

Przybliżenie algorytmicznej informacji wzajemnej

Minimalna transformacja gramatykowa:

$$\mathbf{\Gamma} : \mathbf{w} \in \mathbb{X}^* \longrightarrow \mathbf{\Gamma(w)},$$

gdzie $\mathbf{\Gamma(w)}$ to najmniejsza gramatyka dla $\mathbf{w}$.

$|\mathbf{\Gamma(w)}|$ — przybliżenie złożoności algorytmicznej napisu $\mathbf{w}$.

Przybliżenie algorytmicznej informacji wzajemnej między $\mathbf{u}$ i $\mathbf{v}$:

$$\underbrace{|\mathbf{\Gamma(u)}| + |\mathbf{\Gamma(v)}| - |\mathbf{\Gamma(uv)}|}_{\text{nadwyżka długości gramatyki}}.$$

Ograniczenie rozmiaru słownika (Dębowski 2006)

Dla minimalnej transformacji gramatykowej zachodzi nierówność:

$$\mathbf{0} \leq |\mathbf{\Gamma(u)}| + |\mathbf{\Gamma(v)}| - |\mathbf{\Gamma(w)}| \leq \mathbb{V}\mathbf{[\Gamma(w)]}\mathbb{L}\mathbf{(w)}, \qquad (1)$$

gdzie $\mathbf{w = uv}$ zaś

$$\mathbb{L}\mathbf{(w)} := \max_{\mathbf{s,x,y,z} \in \mathbb{X}^* : \mathbf{w=xsysz}} |\mathbf{s}|,$$

to długość najdłuższego powtórzenia w napisie $\mathbf{w}$.

W dalszej części interesować nas będą uogólnienia nierówności (1) w przypadku kompresji procesu stochastycznego.

Stacjonarny proces stochastyczny

Ciąg zmiennych losowych:

$$(..., \mathbf{X_{-1}}, \mathbf{X_0}, \mathbf{X_1}, \mathbf{X_2}, ...), \qquad \mathbf{X_i} : \Omega \to \mathbb{X}.$$

Rozkład prawdopodobieństwa niezmienniczy względem przesunięć:

$$\mathbf{P(X_{i+1} = x_1, X_{i+2} = x_2, ..., X_{i+n} = x_n) = \mathbb{P}(x_1 x_2 ... x_n)},$$

gdzie $\mathbb{P} : \mathbb{X}^* \to \mathbb{R}$.

Kody jednoznacznie dekodowalne

Kod $\mathbf{C} : \mathbb{X}^* \longrightarrow \mathbb{X}^*$ nazywa się jednoznacznie dekodowalnym, gdy jego rozszerzenie (na skończone konkatenacje)
$\mathbf{C^*} : (\mathbb{X}^*)^* \ni (\mathbf{u_1}, ..., \mathbf{u_k}) \mapsto \mathbf{C(u_1)...C(u_n)} \in \mathbb{X}^*$
jest przekształceniem wzajemnie jednoznacznym.

Kod jest j.d., jeżeli jest bezprzedrostkowy, tzn. żaden napis z jego obrazu nie jest przedrostkiem innego napisu z obrazu.

Dla dowolnego j.d. kodu $\mathbf{C}$ i każdego rozkładu p-stwa $\mathbb{P}$ zachodzi

$$\sum_{\mathbf{u} \in \mathbb{X}^* \mathbf{n}} \mathbb{P}(\mathbf{u}) \, |\mathbf{C(u)}| \geq \mathbf{H(n)} := - \sum_{\mathbf{u} \in \mathbb{X}^{\mathbf{n}}} \mathbb{P}(\mathbf{u}) \log_{|\mathbb{X}|} \mathbb{P}(\mathbf{u}) \quad \text{(entropia)}.$$

Entropia a kodowanie arytmetyczne

Dla każdego dyskretnego rozkładu p-stwa $\mathbb{P} : \mathbb{X}^{\mathbf{n}} \longrightarrow \mathbb{R}$ istnieje bezprzedrostkowy kod $\mathbf{C} : \mathbb{X}^{\mathbf{n}} \longrightarrow \mathbb{X}^{*}$ taki, że

$$|\mathbf{C(u)}| = \left\lceil -\log_{|\mathbb{X}|} \mathbb{P}(\mathbf{u}) \right\rceil + \mathbf{1}.$$

Dla tegoż kodu mamy

$$\mathbf{0} \leq \sum_{\mathbf{u} \in \mathbb{X}^{\mathbf{n}}} \mathbb{P}(\mathbf{u}) \, |\mathbf{C(u)}| - \mathbf{H(n)} \leq \mathbf{2}.$$

Kod j.d. nazywa się uniwersalnym, gdy asymptotycznie optymalnie kompresuje on dowolny proces stacjonarny.

Kody uniwersalne

Ponieważ $\mathbf{H(n)/n}$ jest nieujemne i maleje z $\mathbf{n}$, istnieje granica

$$\mathbf{h = \lim_{n\to\infty} H(n)/n} \quad \text{(intensywność entropii)}.$$

Mówimy, że kod j.d. $\mathbf{C}$ jest uniwersalny, gdy spełnia

$$\mathbf{\lim_{n\to\infty} \sum_{u\in\mathbb{X}^n} \mathbb{P}(u)\,|C(u)|\,/n = h}$$

dla dowolnego rozkładu stacjonarnego $\mathbb{P}$.

Przykładem kod uniwersalnego jest kod Lempela-Ziva.

Informacja wzajemna (Dębowski 2006)

$$\mathbf{E(n) = 2H(n) - H(2n)} \quad \text{(entropia nadwyżkowa)},$$

$$\mathbf{E^C(n) = \sum_{u,v \in \mathbb{X}^n} \mathbb{P}(uv) \underbrace{[|C(u)| + |C(v)| - |C(uv)|]}_{\text{nadwyżka długości kodu}}}.$$

Jeżeli $\mathbf{C}$ jest uniwersalny, to

$$\mathbf{\limsup_{n\to\infty} \left[E^C(n) - E(n) \right] \geq 0.}$$

Jeśli $\mathbf{C}$ i $\mathbf{C'}$ są uniwersalne oraz $|\mathbf{C}(\cdot)| \geq |\mathbf{C'}(\cdot)|$, to

$$\mathbf{\limsup_{n\to\infty} \left[E^C(n) - E^{C'}(n) \right] \geq 0.}$$

Transformacja gramatykowa a kod uniwersalny

Nie istnieje kod uniwersalny o długości równej długości najmniejszej gramatyki dla napisu $\mathbf{u}$, gdyż ta ma długość

$$\mathbf{O(|u| \,/\, \log |u|).}$$

Intuicja

Nie można zakodować nieskończonej liczby symboli nieterminalnych za pomocą pojedynczych symboli (lub słów ustalonej długości) nad skończonym alfabetem.

Mimo to istnieje kod uniwersalny,

- który koduje gramatykę minimalną w uogólnionym sensie,
- i którego nadwyżka długości jest mniejsza od iloczynu rozmiaru słownika i długości najdłuższego powtórzenia.

Konstrukcja kodu

$\mathbf{B} : \mathcal{G} \longrightarrow \mathbb{X}^+$ nazywam lokalnym koderem gramatykowym, gdy

$$\mathbf{B(G) = B_S(B_N(G))},$$

gdzie $\mathbf{B_N} : \mathcal{G} \longrightarrow (\{\mathbf{0}\} \cup \mathbb{N})^*$ koduje gramatykę

$$\mathbf{G = \{A_1 \longrightarrow \alpha_1, A_2 \longrightarrow \alpha_2, ..., A_n \longrightarrow \alpha_n\}}$$

jako ciąg liczb $\mathbf{B_N(G) = F_1(\alpha_1)DF_2(\alpha_2)D...DF_n(\alpha_n)(D+1)}$,

$$\mathbf{F_i(x) := x, \quad x \in \mathbb{X} = \{0, 1, ..., D-1\}},$$
$$\mathbf{F_i(A_j) := D + 1 + j - i, \qquad F_i(\beta\gamma) = F_i(\beta)F_i(\gamma),}$$

zaś $\mathbf{B_S} : (\{\mathbf{0}\} \cup \mathbb{N})^* \longrightarrow \mathbb{X}^*$ spełnia $\mathbf{B_S(\beta\gamma) = B_S(\beta)B_S(\gamma)}$.

Uogólnienie minimalnej transformacji gramatykowej

Dla kodera gramatykowego $\mathbf{B}$, transformację gramatykową $\mathbf{\Gamma}$ oraz kod $\mathbf{B(\Gamma(\cdot))}$ nazywam **B**-minimalnymi, jeżeli

$$|\mathbf{B(\Gamma(w))}| \leq |\mathbf{B(G)}|$$

dla każdej gramatyki $\mathbf{G}$ generującej słowo $\mathbf{w}$.

Twierdzenie (Dębowski 2007)

Niech $\mathbf{B_S}$ będzie koderem bezprzedrostkowym zaś $|\mathbf{B_S(\cdot)}|$ będzie funkcją rosnącą i spełnia

$$\limsup_{\mathbf{n\to\infty}} |\mathbf{B_S(n)}|/\log_{\mathbf{D}} \mathbf{n} = \mathbf{1}.$$

Jeżeli kod $\mathbf{C}$ jest $\mathbf{B_S(B_N(\cdot))}$-minimalny, to jest uniwersalny.

Główne twierdzenie

Twierdzenie (Dębowski 2007)

Dla lokalnego kodera gramatykowego $\mathbf{B} = \mathbf{B_S(B_N(\cdot))}$, oznaczmy

$$\mathbf{W_m} := \max_{\mathbf{0 \le n \le D+2+m}} |\mathbf{B_S(n)}|.$$

Jeżeli kod $\mathbf{C = B(\Gamma(\cdot))}$ jest $\mathbf{B}$-minimalny, to

$$|\mathbf{C(u)}| + |\mathbf{C(v)}| - |\mathbf{C(w)}| \le \mathbf{W_0} \mathbb{V}\mathbf{[\Gamma(w)]}(1 + \mathbb{L}(\mathbf{w})),$$

gdzie $\mathbf{w = uv}$ dla $\mathbf{u, v} \in \mathbb{X}^+$.

(— Tudzież kod ten ma kilka innych przyjemnych własności.)

Entropia języka naturalnego

Spróbujmy wyobrazić sobie proces tworzenia tekstów w języku naturalnym jako proces stacjonarny $(\mathbf{X_k})_{\mathbf{k}\in\mathbb{Z}}$:

- $\mathbf{X_i}$ — pojedyncze litery tekstu ($\mathbb{X}$ skończone),
- $(\mathbf{X_i})_{\mathbf{m}\leq\mathbf{i}\leq\mathbf{n}}$ — konkretne teksty tworzone przez ludzi.

W oparciu o pomiary entropii warunkowej (Shannon 1950) sformułowano hipotezę (Hilberg 1990), że proces ten ma entropię nadwyżkową $\mathbf{E(n)} \asymp \sqrt{\mathbf{n}}$.

- W. Hilberg, (1990). Frequenz, 44:243–248.
- Ł. Dębowski, (2006). *On Hilberg's law and its links with Guiraud's law.* Journal of Quantitative Linguistics, 13:81–109.

Hipoteza

Jeżeli tekst długości **N** liter opisuje w sposób niesprzeczny
$\mathbf{N}^{\beta}$ losowych niezależnych faktów o świecie,
a proces generowania tekstu jest dostatecznie losowy,
to tekst ten zawiera co najmniej $\mathbf{N}^{\beta}/\log \mathbf{N}$ różnych słów
(rozumianych jako symbole nieterminalne najkrótszej gramatyki).

Pewien wyidealizowany proces stochastyczny

Rozważmy proces $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ o zmiennych losowych

$$\mathbf{X_i} := (\mathbf{K_i}, \mathbf{Z_{K_i}}),$$

gdzie niezależne procesy IID $(\mathbf{K_i})_{\mathbf{i}\in\mathbb{Z}}$ oraz $(\mathbf{Z_k})_{\mathbf{k}\in\mathbb{N}}$ spełniają $\mathbf{P(K_i = k) > 0}$ dla $\mathbf{k} \in \mathbb{N}$ oraz $\mathbf{P(Z_k = z)} = \frac{1}{2}$ dla $\mathbf{z} \in \{\mathbf{0}, \mathbf{1}\}$.

Interpretacja quasi-lingwistyczna

Proces $(\mathbf{X_i})_{\mathbf{i}\in\mathbb{Z}}$ jest ciągiem losowych stwierdzeń $\mathbf{X_i}$ niesprzecznie opisujących stan „wcześniej” wylosowanego obiektu $(\mathbf{Z_k})_{\mathbf{k}\in\mathbb{N}}$ przyjmującego nieprzeliczalnie wiele wartości.

Stwierdzenie $\mathbf{X_i = (k, z)}$ orzeka, że $\mathbf{k}$-ty bit obiektu $(\mathbf{Z_k})_{\mathbf{k}\in\mathbb{N}}$ ma wartość $\mathbf{z}$, w taki sposób, że można ustalić zarówno $\mathbf{k}$ jak $\mathbf{z}$.

Dla stwierdzeń $\mathbf{X_i = (k, z)}$ i $\mathbf{X_j = (k', z')}$ nie wiadomo, które bity opiszą i jakie wartości im przypiszą, ale jeżeli $\mathbf{k = k'}$, to $\mathbf{z = z'}$.

Dolne ograniczenie entropii nadwyżkowej

$\mathbf{E(n)}$ można ograniczyć przez liczbę bitów procesu $(\mathbf{Z_k})_{\mathbf{k}\in\mathbb{N}}$, które można przewidzieć dostatecznie dobrze na podstawie $\mathbf{X_{1:n}}$:

$$\mathbf{H_\delta^U(n)} := \left|\left\{\mathbf{k}\in\mathbb{N} : \mathbf{P}\left(\mathbf{f_{nk}}\left((\mathbf{X_i})_{1\le \mathbf{i}\le \mathbf{n}}\right) = \mathbf{Z_k}\right) \ge \delta\right\}\right|.$$

Twierdzenie (Dębowski 2007)

Dla $\mathbf{H^I(n)} := \mathbf{I}\left((\mathbf{X_i})_{1\le \mathbf{i}\le \mathbf{n}}; (\mathbf{Z_k})_{\mathbf{k}\in\mathbb{N}}\right)$ mamy $\lim_{\mathbf{n}} \mathbf{H^I(n)/n} = \mathbf{0}$ oraz

$$\mathbf{E(n)} \ge \mathbf{2H^I(n)} - \mathbf{H^I(2n)},$$

$$\mathbf{H^I(n)} \ge \mathbf{H_\delta^U(n)} \cdot \frac{\log 2 + \delta\log\delta + (1-\delta)\log(1-\delta)}{\log|\mathbb{X}|}.$$

Zatem jeżeli $\mathbf{H_\delta^U(n)} \ge \mathbf{B}\cdot\mathbf{n}^{\beta}$ (np. dla $\mathbf{P(K_i = k)} \sim \mathbf{k}^{-1/\beta}$), to

$$\limsup_{\mathbf{n}\to\infty}\left[\mathbf{E(n)} - \mathbf{B'}\cdot\mathbf{n}^{\beta}\right] \ge \mathbf{0}.$$

Procesy nieprzeliczalnego opisu

Twierdzenie z poprzedniego slajdu obowiązuje też dla p.n.o.

Proces stacjonarny $(X_i)_{i\in\mathbb{Z}}$ nazywam p.n.o., jeżeli istnieją

- zmienne $(Z_k)_{k\in\mathbb{N}} \sim IID$, $P(Z_k = z) = \frac{1}{2}$ dla $z \in \{0,1\}$,
- i funkcje $f_{nk} : \mathbb{X}^n \to \{0,1\}$

takie, że

- $P\left(f_{nk}\left((X_i)_{j+1\leq i\leq j+n}\right) = Z_k\right)$ nie zależy od $j \in \mathbb{Z}$,
- $\lim_n P\left(f_{nk}\left((X_i)_{j+1\leq i\leq j+n}\right) = Z_k\right) = 1$.

Są to procesy nieergodyczne.

Procesy o skończonej energii

Proces stacjonarny $(\mathbf{X_i})_{i\in\mathbb{Z}}$ nazywa się p. o s.e., jeżeli

$$\mathbb{P}(\mathbf{uv})/\mathbb{P}(\mathbf{u}) \leq \mathbf{K}\exp\left[-\mathbf{c}\,|\mathbf{v}|\right], \quad \mathbf{u},\mathbf{v}\in\mathbb{X}^{+}.$$

Przykładem p. o s.e. jest ciąg zmiennych

$$\mathbf{X_i} \equiv \mathbf{Y_i} + \mathbf{U_i} \pmod{\mathbf{D}},$$

gdzie proces $(\mathbf{Y_i})_{i\in\mathbb{Z}}$ jest ergodyczny zaś $(\mathbf{U_i})_{i\in\mathbb{Z}} \sim \mathbf{IID}$.

Twierdzenie (Shields 1997)

Jeżeli $(\mathbf{X_i})_{i\in\mathbb{Z}}$ jest p. o s.e., to długość najdłuższego powtórzenia jest ograniczona nierównością

$$\liminf_{n\to\infty}\left[\mathbf{K'}\log\mathbf{n} - \mathbb{L}\left((\mathbf{X_i})_{1\leq i\leq n}\right)\right] \geq 0 \quad \text{prawie na pewno.}$$

Składając wszystko razem

Twierdzenie

Jeżeli kod $\mathbf{C = B(\Gamma(\cdot))}$ jest $\mathbf{B}$-minimalny dla lokalnego kodera $\mathbf{B}$, zaś $\mathbf{(X_i)_{i\in\mathbb{Z}}}$ jest p. o s.e., to

$$\mathbf{\limsup_{n\to\infty}\left[\sum_{w\in\mathbb{X}^n}\mathbb{P}(w)\mathbb{V}[\Gamma(w)] - \frac{E(n)}{K'\log n}\right] \geq 0.}$$

Jeżeli $\mathbf{(X_i)_{i\in\mathbb{Z}}}$ jest p.n.o. takim, że

$$\mathbf{\left|\left\{k\in\mathbb{N} : P\left(f_{nk}\left((X_i)_{1\leq i\leq n}\right) = Z_k\right)\geq\delta\right\}\right| \geq B\cdot n^{\beta},}$$

to

$$\mathbf{\limsup_{n\to\infty}\left[E(n) - B'\cdot n^{\beta}\right] \geq 0.}$$

Niestety $\mathbf{\limsup_n(a_n+b_n) \geq \limsup_n a_n + \liminf_n b_n}$.

Moje prace

- Ł. Dębowski, (2007). *On vocabulary size of grammar-based codes.* `http://xxx.lanl.gov/abs/cs.IT/0701047`. (ISIT 2007, Nicea)
- Ł. Dębowski, (2006). *Ergodic decomposition of excess entropy and conditional mutual information.* Prace IPI PAN nr 993.
- Ł. Dębowski, (2006). *On Hilberg's law and its links with Guiraud's law.* Journal of Quantitative Linguistics, 13:81–109.

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