Warszawa, 2 grudnia 2020

dr hab. Łukasz Dębowski
Instytut Podstaw Informatyki PAN
ldebowsk@ipipan.waw.pl

Zainteresowania naukowe

Od ukończenia studiów magisterskich z zakresu fizyki teoretycznej, moje zainteresowania naukowe koncentrują się na matematycznych podstawach statystycznego modelowania języka naturalnego. Motywuje mnie ciekawość, jakiego typu proces stochastyczny może modelować tworzenie tekstów przez ludzi. W ramach tych zainteresowań prowadzę badania matematyczne w zakresie rachunku prawdopodobieństwa i teorii informacji, jak też okazjonalnie badam empirycznie statystyczne własności tekstów w języku naturalnym (tzw. lingwistyka kwantytatywna). Badania moje mają charakter podstawowy w stosunku do prac inżyniersko-informatycznych w sztucznej inteligencji i inżynierii lingwistycznej (tzw. lingwistyce komputerowej), ale mogą one zaowocować zastosowaniami. Podsumowaniem moich dotychczasowych prac jest monografia *Information Theory Meets Power Laws: Stochastic Processes and Language Models* wydana jako ebook w listopadzie 2020 przez wydawnictwo Wiley (wydanie drukowane ukaże się w lutym 2021).

W poniższych akapitach przybliżę swoje kluczowe wyniki:

1. Od prac Chomskyego z lat 1950-tych uważa się, że proces tworzenia tekstu przez ludzi nie może być ukrytym procesem Markowa o skończonej liczbie stanów. W latach 1990-tych inżynierowie lingwistyczni próbowali modelować ów proces przez ukryty proces Markowa z bardzo dużą, acz skończoną liczbą stanów. Takie procesy skończenie stanowe są do pewnego stopnia użyteczne, ale nie uwzględniają one ważnych zależności w tekście, co współcześnie potwierdza większa skuteczność modeli opartych na głębokich sieciach neuronowych. W próbach wyjścia poza klasę procesów skończenie stanowych pierwszą inspiracją była dla mnie hipoteza Hilberga zakładająca, że entropia Shannona ciągu n kolejnych liter tekstu w języku naturalnym jest proporcjonalna do $\sqrt{n}$. Hipoteza ta nie była traktowana serio, gdyż implikuje zerową intensywność entropii Shannona i determinizm ludzkich wypowiedzi (tzn. każda litera tekstu jest prawie na pewno mierzalną funkcją nieskończonego ciągu liter poprzedzających). Crutchfield i Feldman zaproponowali jednak interesujące osłabienie hipotezy Hilberga zakładające, że informacja wzajemna pomiędzy dwoma następującymi po sobie ciągami n kolejnych liter tekstu jest proporcjonalna do $\sqrt{n}$.

Z osłabionej hipotezy Hilberga wynika, że odpowiedni proces stochastyczny nie może być procesem skończenie stanowym. W pierwszej kolejności skonstruowałem przykłady procesów stochastycznych spełniające ten warunek. Pomocne okazało się pojęcie entropii nadwyżkowej, czyli informacji wzajemnej między półnieskończoną przeszłością i przyszłością procesu. Entropia nadwyżkowa jest niekończona, jeżeli (ale nie tylko wtedy, gdy) proces jest silnie nieergodyczny (tzn. σ-ciało niezmiennicze jest bezatomowe). Silna nieergodyczność ma naturalną interpretację lingwistyczną. Zmienne mierzalne względem σ-ciała niezmienniczego można sobie wyobrażać jako tematy nieskończonych losowych tekstów, a silna nieergodyczność oznacza, że pula takich tematów jest nieprzeliczalna. Dysponując tą interpretacją odkryłem procesy Santa Fe, o elementarnej konstrukcji, które są silnie nieergodyczne i spełniają osłabioną hipotezę Hilberga. Prosta modyfikacja procesów Santa Fe pozwala uzyskać procesy mieszające także spełniające osłabioną hipotezę Hilberga.

2. Najsłynniejszym prawem ilościowym spełnionym przez teksty w języku naturalnym jest prawo Zipfa. Głosi ono, że rangi słów w tekście są w przybliżeniu odwrotnie proporcjonalne do ich częstości. (Z definicji ranga słowa n-tego co do częstości wynosi n.) Prawo to ma swoją wersję całkową, prawo Herdana, które głosi, że liczba różnych słów w tekście długości n jest proporcjonalna do n^β. W swoich pracach powiązałem prawo Herdana z osłabioną hipotezą Hilberga. Powiązanie ma następującą postać:

   Twierdzenie 1 *Jeżeli informacja wzajemna pomiędzy dwoma następującymi po sobie ciągami n kolejnych liter jest większa niż n^β, gdzie $\beta \in (0,1)$, to liczba różnych słów w tekście długości n jest większa niż $n^\beta/L(n)$, gdzie $L(n)$ jest długością maksymalnego powtórzenia w tekście długości n.*

   Twierdzenie 1 wymaga doprecyzowania, jak definiowane są słowa w dowolnym losowym ciągu liter. Odpowiedniej definicji dostarcza procedura kompresji danych oparta na gramatykach bezkontekstowych — słowa są określone jako symbole nieterminalne w najkrótszej gramatyce generującej dany tekst jako jedyną produkcję. Eksperymenty w lingwistyce komputerowej potwierdzają, że taka procedura identyfikacji słów zwraca w przybliżeniu to samo, co definicja ortograficzna — w myśl tej ostatniej, słowem jest dowolny ciąg znaków między dwiema spacjami.

Twierdzenie 1 dopełnione jest przez kolejny mój wynik. Dla procesu silnie nieergodycznego, faktami będę nazywać niezależne zmienne binarne mierzalne względem σ-ciała niezmienniczego. — Faktów w tym sensie jest nieskończenie wiele. Prawdziwe jest następujące zdanie:

Twierdzenie 2 *Jeżeli liczba faktów, które można przewidzieć wystarczająco dobrze z tekstu długości n, jest większa niż n^β, gdzie $\beta \in (0,1)$, to informacja wzajemna Shannona pomiędzy dwoma następującymi po sobie ciągami n kolejnych liter jest większa niż n^β.*

Z Twierdzeń 1 i 2 wynika, że liczba różnych słów w tekście musi być większa niż liczba różnych faktów opisanych w tekście podzielona przez długość maksymalnego powtórzenia. Wynik ten wiąże ilościowo znaczenie losowego tekstu z jego strukturą.

3. Długość maksymalnego powtórzenia $L(n)$ w tekście długości n jest ważną statystyką charakteryzującą siłę zależności w procesie. Dla szerokiej klasy procesów stochastycznych, zwanych procesami o skończonej energii (są nimi procesy skończenie stanowe i procesy ψ-mieszające), zachodzi proporcjonalność $L(n) \propto \log n$. W przypadku tekstów w języku naturalnym badania empiryczne przeprowadzone przeze mnie wykazały, że mamy do czynienia z zależnością $L(n) \propto (\log n)^\alpha$, gdzie $\alpha \approx 3$. Otwartym pytaniem jest to, do jakiego stopnia tempo wzrostu długości maksymalnego powtórzenia można powiązać z entropią procesu. Pokazałem, że jeżeli $L(n) \propto (\log n)^\alpha$, gdzie $\alpha > 1$, to intensywność warunkowej entropii Renyiego wynosi zero. Skonstruuowałem także procesy RHA (Random Hierarchical Association), które cechuje własność $L(n) \propto (\log n)^\alpha$ oraz zerowa intensywność entropii Shannona. Powszechnie uważa się jednak, że intensywność entropii Shannona dla języka naturalnego jest większa od zera. Interesującym problemem otwartym jest zatem skonstuowanie motywowanych lingwistycznie procesów, dla których intensywność warunkowej entropii Renyiego wynosi zero a intensywność entropii Shannona jest dodatnia.

W powyższych punktach omówiłem wyniki bezpośrednio związane z tematyką statystycznego modelowania języka naturalnego. Oprócz tego mam na swoim koncie prace o charakterze bardziej podstawowym, motywowane potrzebą stworzenia podglebia dla prac bardziej stosowanych (z punktu widzenia czystej matematyki). Dotyczą one przykładowo:

1. Algebraicznych własności warunkowej informacji wzajemnej dla σ-ciał. (Potrzeba takiego pojęcia pojawia się przy dyskusji rozkładu ergodycznego entropii nadwyżkowej.)

2. Własności wykładników Hilberga, mierzących potęgowe tempo wzrostu informacji wzajemnej dla procesu stacjonarnego. (Te wyniki wykorzystują algebraiczne własności złożoności Kołmogorowa.)

3. Kodowania miar stacjonarnych przy pomocy kodu zmiennej długości. (Badałem, jak powiązane są własności procesu, którego wartościami są litery, z własnościami procesu, którego wartościami są słowa.)

4. Teorii uniwersalnej kompresji i predykcji dla ciągów algorytmicznie losowych w sensie Martina-Löfa względem nieobliczalnych miar stacjonarnych ergodycznych. (Te wyniki powstały we współpracy z moim doktorantem Tomaszem Steiferem.)

5. Zgodnego estymatora liczby stanów ukrytych dla procesów stacjonarnych ergodycznych i jego związków z wykładnikiem Hilberga dla algorytmicznej informacji wzajemnej. (Tutaj także pojawiają się kody uniwersalne.)

Sądzę, że interesująca mnie tematyka może być płodna w przyszłości w teorię i zastosowania. Zależy mi na przyciągnięciu do niej innych probabilistów oraz informatyków teoretycznych.