

Entropia nadwyżkowa a teksty w języku naturalnym

Jak logiczna niesprzeczność losowego opisywania nieskończonego losowego świata może implikować efektywną wykrywalność podziału tekstów na słowa?

Streszczenie

Referat dotyczyć będzie związków pomiędzy językoznawstwem a teorią informacji, konkretnie entropii nadwyżkowej. Entropia nadwyżkowa jest pewną miarą pamięci stacjonarnego procesu stochastycznego. Istnieje hipoteza (Hilberg, 1990), że jeżeli generację tekstu można opisać jako stacjonarny proces stochastyczny, to ów proces ma nieskończoną entropię nadwyżkową. Hipoteza ta została sformułowana w oparciu o pomiary entropii dla języka angielskiego przeprowadzone przez Shannona (1950) metodą zgadywania. Istnieją jednak poważniejsze przesłanki, aby uznać tę hipotezę za wartą dalszych badań. Opowiem mianowicie o zaskakującej lingwistycznej interpretacji ściśle matematycznych związków entropii nadwyżkowej z teorią ergodyczną i teorią kodowania. Związki te są następujące:

1. Rozważmy sytuację, w której każdy tekst opisuje w sposób częściowo losowy, logicznie niesprzeczny i asymptotycznie zupełny pewien losowy stan nieskończonego świata. (Losowość w układzie (tekst, świat) pojawia się na dwóch poziomach: nieograniczonej losowości stanu świata i losowości tekstu, która jest ograniczona przez wylosowany, acz niezmany stan świata.) Można pokazać (Dębowski, 2006c), że dowolny stacjonarny proces stochastyczny modelujący generację tekstów o ww. własnościach ma nieskończoną entropię nadwyżkową.
2. W oparciu o pewne ściśle określone procedury kompresji danych (tzw. gramatyczne kody uniwersalne), dowolny "tekst" (tzn. dowolny ciąg liter) można podzielić na "słowa", które są bliskie maksymalnym powtarzającym się w "tekście" ciągom liter. Można pokazać (Dębowski, 2006a), (Dębowski, 2006c), że dla tekstu będącego realizacją procesu stacjonarnego liczba różnych "słów" w "tekście" jest nie mniejsza niż pewne przybliżenie entropii nadwyżkowej, które zależy od długości "tekstu".
3. Doświadczenia wskazują, że dla "tekstów" będących tekstami w języku naturalnym ww. "słowa" są zbliżone do tego, co zwykle nazywa się słowami w tekście, a z konkretnej postaci hipotezy Hilberga wynika dobrze znany fakt, że liczba różnych słów jest większa od pierwiastka kwadra-

towego długości tekstu. Dla porównania, dla "tekstu" będącego ciągiem rzutów uczciwą monetą, entropia nadwyżkowa równa jest zeru, a liczba różnych "słów" rośnie z długością "tekstu" bardzo wolno (Dębowski, 2006b).

Literatura

- DĘBOWSKI, Ł. (2006A): *On Hilberg's law and its links with Guiraud's law*. Journal of Quantitative Linguistics 13, 81–109.
- (2006B): *Menzerath's law for the smallest grammars*, [w:] The Exact Science of Language and Text, (R. Koehler, P. Grzybek, red.), De Gruyter – w druku
- (2006C): *Ergodic decomposition of excess entropy and conditional mutual information*. Prace IPI PAN nr 993.
- HILBERG, W. (1990): *Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?*. Frequenz 44, 243–248.