Nyelvi modellek
“Kutya nehéz úgy hazudni, ha az ember nem ösmeri az igazságot”

Varjú Zoltán

Weblib Kft.

2012-03-12

Bevezetés

> "Essentially, all models are wrong, but some are useful."
>
> — George Edward Pelham Box

- Chomsky
- Norvig
- Turing
- Shannon
- van Benthem

A vita 1

"Chomsky derided researchers in machine learning who use purely statistical methods to produce behavior that mimics something in the world, but who don't try to understand the meaning of that behavior. Chomsky compared such researchers to scientists who might study the dance made by a bee returning to the hive, and who could produce a statistically based simulation of such a dance without attempting to understand why the bee behaved that way. "That's a notion of [scientific] success that's very novel. I don't know of anything like it in the history of science," said Chomsky."

— Stephen Cass: Unthinking Machines, Technology Review,
`http://www.technologyreview.com/computing/37525/?a=f`

A vita 2

> "Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list."
>
> — Chomsky

Mennyire lehet reprezentatív egy korpusz?

- "I live in New York" sokkal gyakoribb mint az "I live in Dayton Ohio"
- The Unreasonable Effectiveness of Data: "simple models and a lot of data trump more elaborate models based on less data"

Két kultúra

- Hogyan írhatunk le véges eszközökkel egy végtelen jelenséget?
- Modell- és rekurzióelmélet
- Leíró statisztika és korpusznyelvészet
- Algoritmikus tanuláselmélet

Gyökerek 1 - Frege

"For my money, Gentzen's natural deduction and Church's lambda calculus are on a par with Einstein's relativity and Dirac's quantum physics for elegance and insight."

— Philip Wadler, Proofs are Programs

Gyökerek 2 - Turing

- Colossus: a Turing gépek első fizikai implementációja
- Engima: bayesiánus statisztikai módszerek futnak a Colossus-on

Információelmélet 1. - Áttekintés

- Nem térünk ki minden kérdésre
- Miképp lehetséges statisztikailag leírni a nyelvi jelenségeket
- Mintavételezés vs. stacionárius ergodikus forrás
- Az indukció problémája

Információelmélet 2. - stacionárius ergodikus forrás 1.

- Nulladrendű közelítés

"XFOML RXKHRJFFJUJ ZPLWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD"

- Elsőrendű közelítés

"OCRO HLI NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL"

- Másodrendű közelítés

"ON IE ANTSOUTINYS ARE T INCTORE BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE"

Információelmélet 3. - stacionárius ergodikus forrás 2.

- Harmadrendű közelítés

"IN NO IST LAT WHEY CRATIC FROURE BIRS GRODIC PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE"

- Elsőrendű szószintű közelítés

"REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE"

- Másodrendű szószintű közelítés

"THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED"

Információelmélet 5. - stacionárius ergodikus forrás 4.

stacionárius forrás időben nem változik, pl. elsőrendű közelítések

idősor átlag tkp. a relatív gyakoriság

összesített átlag egy infinit forrás végtelen sorozatot hozhat létre

ergodikus forrás minden olyan stacionárius forrás mely idősor átlaga és összesített átlaga megegyezik

Információelmélet 6. - az indukció problémája

- Tök mindegy melyik forrást vizsgáljuk, hiszen annak idősor átlaga megegyezik az ergodikus forrás összesített átlagával
- Ahogy növekszik a vizsgált szekvencia hossza, úgy kerülünk egyre közelebb a forrás átlagához
- Nem statisztikai leírást kapunk, hanem egy algoritmikus módszert arra hogy generáljunk egy közelítő szekvenciát

Modellek

- X nyelvészet, ahol X = kognitív, matematikai, bio, ...
- Társadalomtudományok: a modellek nem leíró jellegűek, "csak" segítik a megértést
- Robert Aumann: `Interactive Epistemology I. & II.`

Zajos csatorna 1

Miért redundáns a nyelv?

- Effektív kódolás problémája, az információnak “át kell jutnia” a zajos csatornán
- Hogyan generáljuk és dekódoljuk az üzenetet?
- Milyen episztemológiai következményei vannak ennek?

Zajos csatorna 2

begin
  $i := 0$
  while *true* do
    begin *read* x_i;
      send x_i *until* $K_S K_R(x_i)$;
      send $K_S K_R(x_i)$ *until* $K_S K_R K_S K_R(x_i)$
      $i := i + 1$
    end
  od
end

Zajos csatorna 3

begin
  when $K_R(x_0)$ *set* $i := 0$
  while *true* do
    begin *write* x_i;
      send $K_R(x_i)$ *until* $K_R K_S K_R(x_i)$;
      send $K_R K_S K_R(x_i)$ *until* $K_R(x_{i+1})$
      $i := i + 1$
    end
  od
end

Dinamikus episztemikus logika

- van Bentem: ‘‘One is a lonely number’’.
- tanulás vs érvelés [learning vs. reasoning about knowledge]

Köszönöm a figyelmet

- Kereső Világ `http://kereses.blog.hu/`
- Számítógépes nyelvészet `http://szamitogepesnyelveszet.blogspot.com/`
- Twitter: @zoltanvarju
- Email: zoltan.varju@weblib.com