Bevezetés

Nyelvi adatok feldolgozása – 2019/20 tavasz
1. óra

Simon Eszter – Ferenczi Zsanett

MTA Nyelvtudományi Intézet

Tartalom

Bemutatkozás

Bemutatkozás

- mi
- ti

A félév bemutatása

Tanrend

Szorgalmi időszak

Első nap: 2020. február 10. (hétfő)
Tavaszi szünet: 2020. április 10. – április 17. (péntek–péntek)
Dékáni szünet: 2020. április 22. (szerda)
Ünnepnap: 2020. május 1. (péntek)
Utolsó tanítási nap: 2020. május 22. (péntek)
Pótlásokra szolgáló időszak: 2020. május 25-29. (hétfő–péntek)

Vizsgaidőszak

Első nap: 2020. június 2. (hétfő)
Utolsó nap: 2020. június 29. (hétfő)

Az órák

- összesen 14 óra
- egy órán belül:
  - elméleti bevezetés slide-okkal
  - gyakorlatok gépen
  - házi feladat

Eredeti tematika

Adatgyűjtés. Végesállapotú technológiák. Környezetfüggetlen nyelvtanok. A szavak megszámolása. Zipf törvényei, hatványtörvények. Indexépítés. A keresőmotorok alapjai. Amit a nyelvészetből tudni kell. A szavak osztályozása. Szótárépítés. Kollokációk, idiómák, többértelműség. Nyelvmodellezés. Súlyozott automaták, Markov modellek, rejtett Markov, n-gram. Helyesírás-ellenőrzés, nyelvtan-ellenőrzés. Beszédfelismerés, írásfelismerés, beszédkeltés. Névelemfelismerés. Funkcionális mondatelemzés. Mondat feletti egységek. Érzület-elemzés. Jelentésreprezentáció. Szójelentés, mondatjelentés, diskurzus-jelentés. Logikai modellek, vektormodellek. Gépi fordítás.

Témák

- nyelvészeti bevezető
- metodológia
- formális nyelvek, nyelvtanok, automaták
- korpuszok
- a szövegfeldolgozás szintjei
- gépi tanulás
- nyelvmodellezés n-gramokkal
- szintaktikai elemzés
- szemantika
- vektorok, szóbeágyazások, neurális hálók
- gépi fordítás
- ...
- meghívott előadók?

Adminisztráció

Teljesítés

Jelenléti követelmények: Legfeljebb 3 óráról lehet hiányozni – ez és az órai aktivitás az aláírás feltételei.

Félévközi számonkérések: A félév során házi feladatok kerülnek kiadásra. Ezekből minimum hármat kell beadni a kiadástól számított 2 héten belül.

A félév végi osztályzat: A beadott házi feladatokból és a félév végi szóbeli vizsgából áll össze az osztályzat.

Konzultáció: E-mailben egyeztetett időben (simon.eszter@nytud.mta.hu,ferenczizsani@gmail.com).

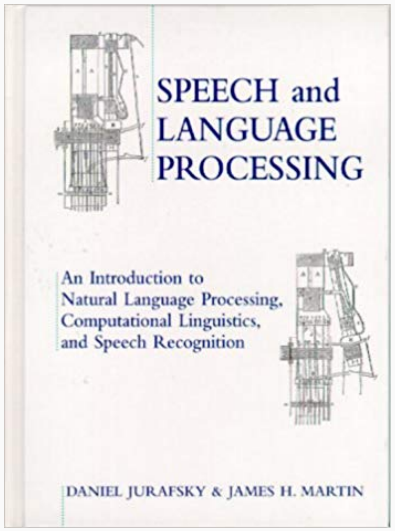
Házi feladatok

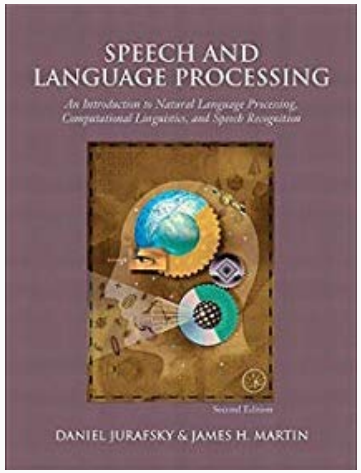
- összesen 5-8 házi feladat kerül kiadásra
- ebből legalább 3-at kell beadni a teljesítéshez
- a feladat kiadásától számítva 2 héten belül
- a feladatokból néhány megoldható programozás nélkül is
- a beadott házikat leosztályozzuk, a 3 legjobb számít
- jegy = 50% házik + 50% szóbeli vizsga

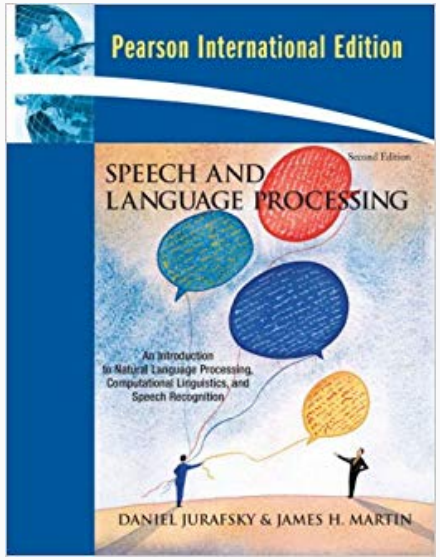
Olvasnivaló

Dan Jurafsky – James H. Martin: Speech and Language Processing

3rd edition draft: https://web.stanford.edu/~jurafsky/slp3/
2nd edition
1st edition







Technikai részletek

Technikai részletek

Python 3:

- Linux, OS X: ✓
- Windows: python, Anaconda
- online lehetőségek: PythonAnywhere, repl.it
- tananyagok:
  - hivatalos tutorial
  - egyéb anyagok gyűjteménye

Repó

URL: https://github.com/esztersimon/nlp_at_bme.git

Git:

- TryGit, The Simple Guide
- `git clone https://github.com/esztersimon/nlp_at_bme.git`

Jupyter Notebook:

- tutorial
- `pip install jupyter` vagy `pip3 install jupyter`
- Anacondában benne van – miniconda esetén: `conda install jupyter`

kérés, kérdés, óhaj, sóhaj?

Bevezetés a számítógépes nyelvészetbe

Szinonimák?

- számítógépes nyelvészet
- természetesnyelv-feldolgozás (natural language processing, NLP)
- nyelvtechnológia (human language technology, HLT)
- korpusznyelvészet

Besorolás


nyelvészet
számítógépes nyelvészet
számítástechnika
villamosmérnöki tudomány
beszédfelismerés
mesterséges intelligencia
természetes nyelv feldolgozás
számítógépes nyelvészet
számítógépes pszicholingvisztika
pszichológia
számítógépes nyelvészet
nyelvészet

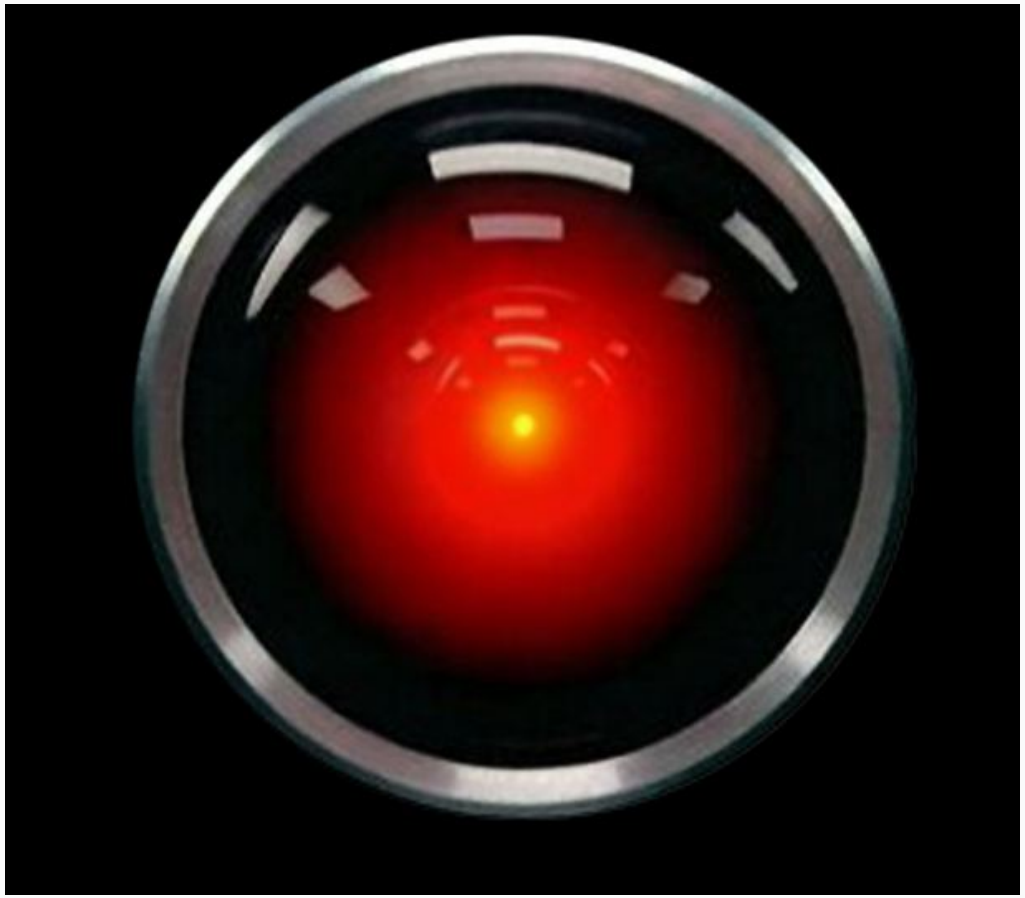
Definíció

- átfedésben van a mesterségesintelligencia-kutatással
- a természetes nyelvek számítógépes feldolgozásával foglalkozik
- a kutatások a nyelv szerkezetének gépi modellezésére irányulnak

Wikipédia:

A számítógépes nyelvészet olyan műszaki tudomány, amely a természetes nyelvű szövegek számítógépes feldolgozásával foglalkozik, de minden olyan elméleti és gyakorlati tevékenység ide tartozik, amely kapcsolatban van a természetes nyelvekkel. Egy interdiszciplína, vagyis olyan szakterület, amely több terület eredményeire és tudására épül, mint pl. az informatika, a matematika és a nyelvészet.

A Szent Grál



A nyelvtechnológia célja

olyan rendszer építése, amely fel tudja dolgozni és elő tudja állítani az emberi nyelvet – úgy, ahogy az ember teszi

elméleti motiváció: az emberi nyelvhasználatot leíró formalizált és konzisztens nyelvi modellek létrehozása

gyakorlati motiváció: a modellek gyakorlati, számítógépes megvalósítása → praktikus gépi alkalmazások

Részterületek

a nyelvtechnológia egyes részfeladatai tükrözik az emberi nyelvértés pszicholingvisztikai részfeladatait

- beszédfelismerés -és szintézis
- morfológiai és szintaktikai elemzés
- szemantikai elemzés
- generálás
- következtetés

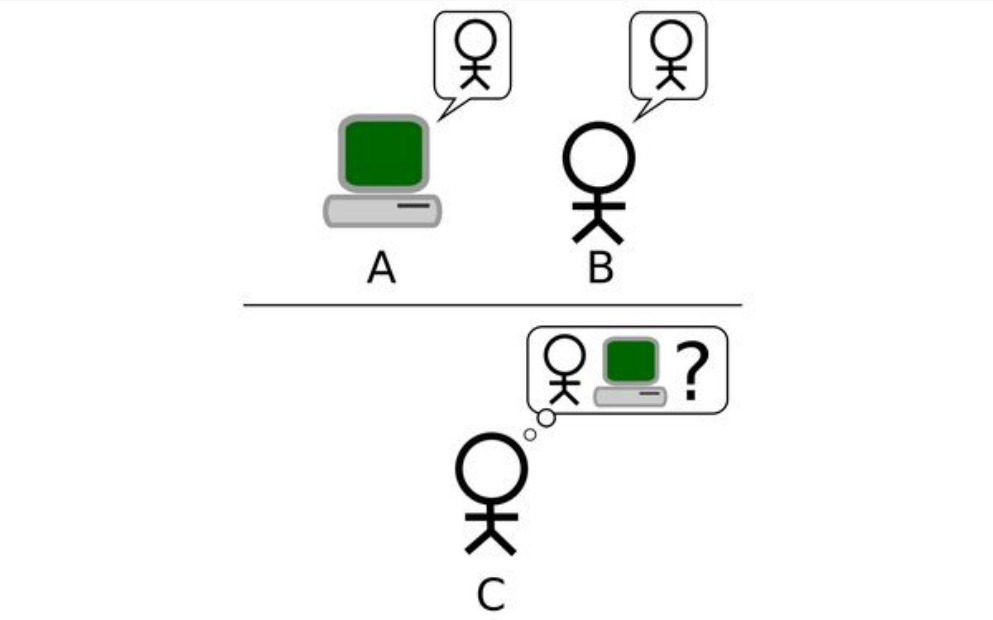
A problémák

- a nyelvfeldolgozás rendkívül bonyolult
- a szükséges tudás hatalmas
- szabályalapú: a szabályok száma, a lexikon mérete
- statisztikai: az adatok ritkasága (“rare words are very common”)
  $\rightarrow$ a 15 leggyakoribb szó adja a szöveg 25%-át, a 100 leggyakoribb a 60%-át, 1000 a 85%-át, 4000 pedig a 97,5%-át
- többértelműség
- magasabb szintű feldolgozási problémák (előfeltevések, mondatok közötti anaforafeloldás stb.)
- robusztusság

Kis történeti áttekintés

Turing-teszt

- három résztvevő: két tesztalany – egy ember és egy gép – és egy kérdező
- a kérdező billentyűzet és monitor közvetítésével kérdéseket tesz fel a két tesztalanynak
- mindkét tesztalany megpróbálja meggyőzni a kérdezőt arról, hogy ő gondolkodó ember
- ha a kérdező öt perces faggatás után sem tudja megállapítani, hogy melyik a gép, akkor a gép átment a teszten



A Turing-teszt kritikája

- a párbeszéd szimulálása csak kevéssé tekinthető az intelligencia jelének → a hagyományos értelemben vett intelligenciának csak egy szegletét tudja mérni;
- attól még lehet intelligens egy gép, hogy nem képes emberi módon kommunikálni;
- az emberek közül se teljesítené mindenki sikerrel a Turing-tesztet (kisgyerekek, fogyatékosok), holott ők is lehetnek más tekintetben intelligensek;
- a teszten olyan ember is megbukhat, aki nem hajlandó a feltételek szerint együttműködni → az együttműködés megtagadása nem egyenlő az értelem hiányával (lásd HAL);
- a kísérleti szituáció jellegénél fogva a lehetséges beszélgetésfolyamat-variációk száma korlátozott → egy kellően kiterjedt adatbázissal ellátott számítógép előre eltárolt kérdés- és válaszminták felhasználásával tényleges intelligencia hiányában is sikerrel teljesítheti a tesztet (lásd Jeopardy)

ELIZA (Weizenbaum, 1966)

- A férfiak mind egyformák.
- Ezen mit értesz?
- Mindig cseszegetnek minket valamiért.
- Mondanál egy konkrét példát?
- Hát, a fiúm küldött ide.
- A fiúd küldött ide.
- Azt mondja, mindig depressziós vagyok.
- Sajnálom, hogy depressziós vagy.
- Talán ki kellene békülnöm az anyámmal.
- Mesélj még a családodról.

...ÉS AMI MÖGÖTTE VAN

rogersiánus pszichológia

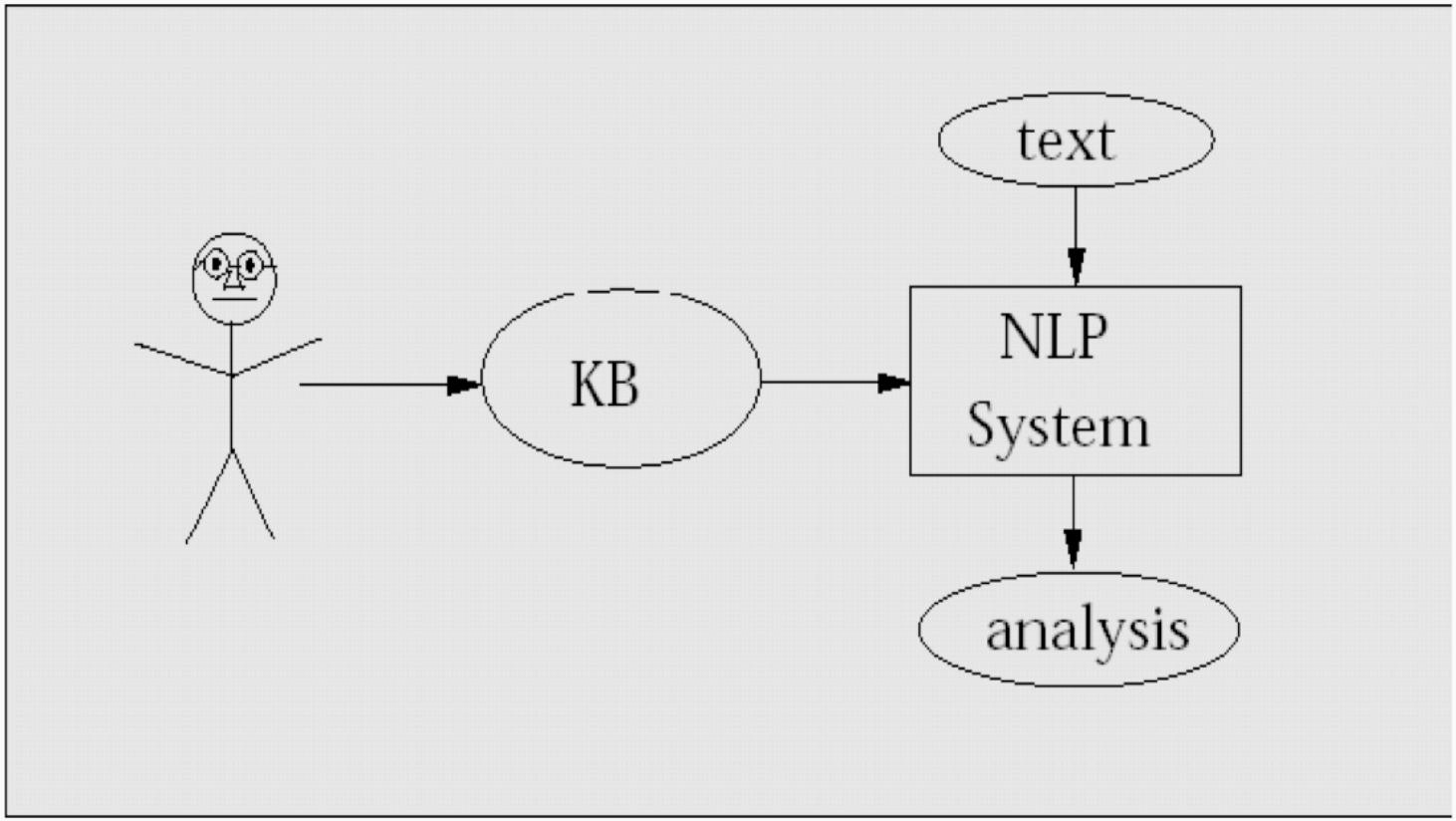
Carl Rogers (1902-1987):

- amerikai pszichológus
- a pszichoterápiás kutatás egyik alapító atyjának tartják
- kliensközpontú terápia:
  - a terapeuta párbeszédbe lép a klienssel
  - bólint, összegzi a hallottakat, ha a másik elakad
  - a feltárás után továbblép
  - nem kérdez, figyel

Példák

- egyszerű kulcsszavak által aktivált utasítások: *my boyfriend* $\rightarrow$ *your boyfriend*
- reguláris kifejezések: *s/.* (depressziós|szomorú) vagyok .*/Sajnálom, hogy \1 vagy/*

Szabályalapú metodológia



Szabályalapú metodológia

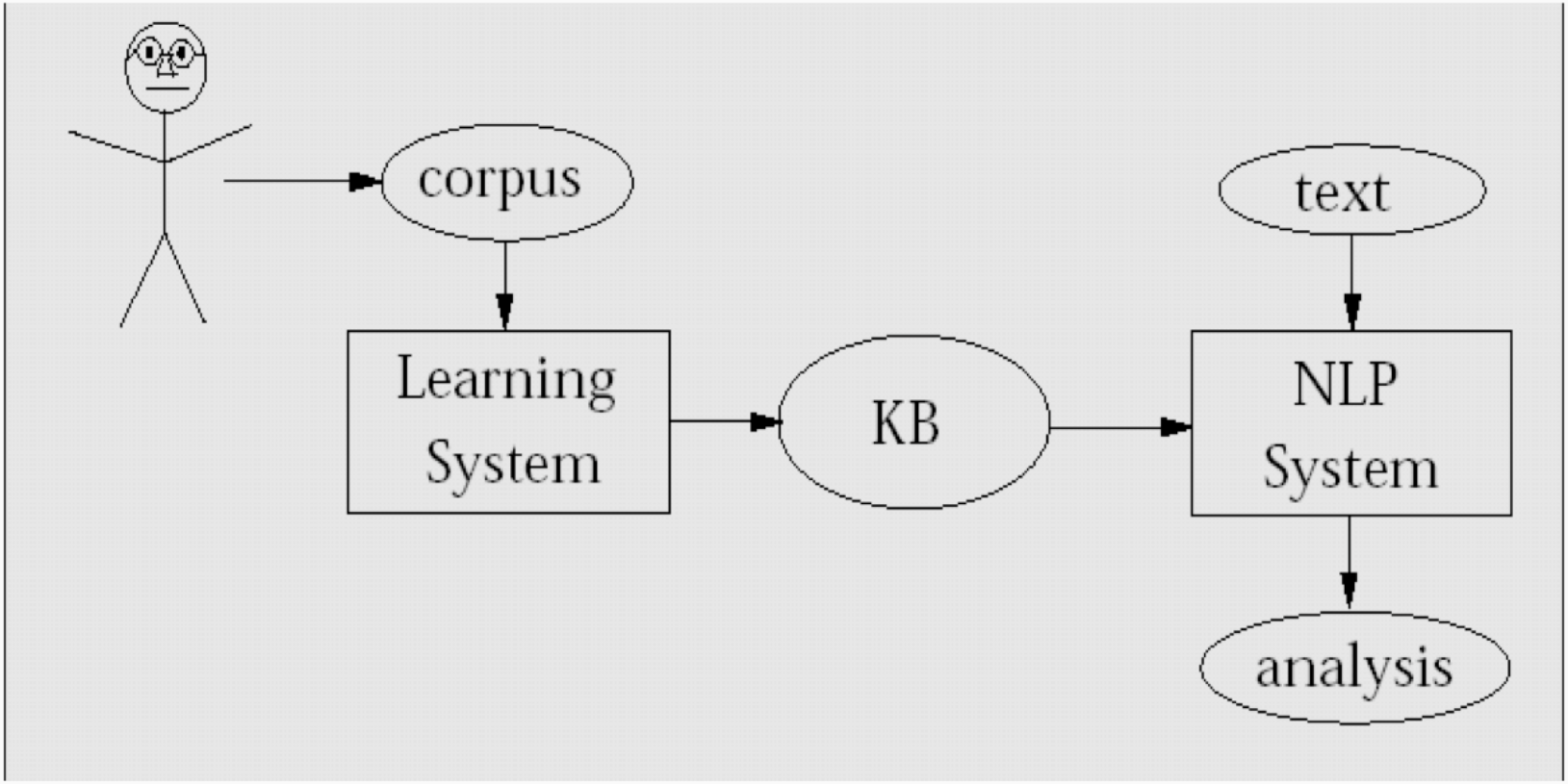
- racionalista filozófiai tradíció (Leibniz, Descartes)
- univerzális nyelvtan
- velünk született nyelvi képesség → introspekció
- grammatikalitási ítélet: 0 vagy 1
- kézzel kódolt szabályok
  - reguláris kifejezések

Példák

e-mail cím: [a-z]+@[a-z]+\.[a-z]+

pl.: bubo@doktor.hu

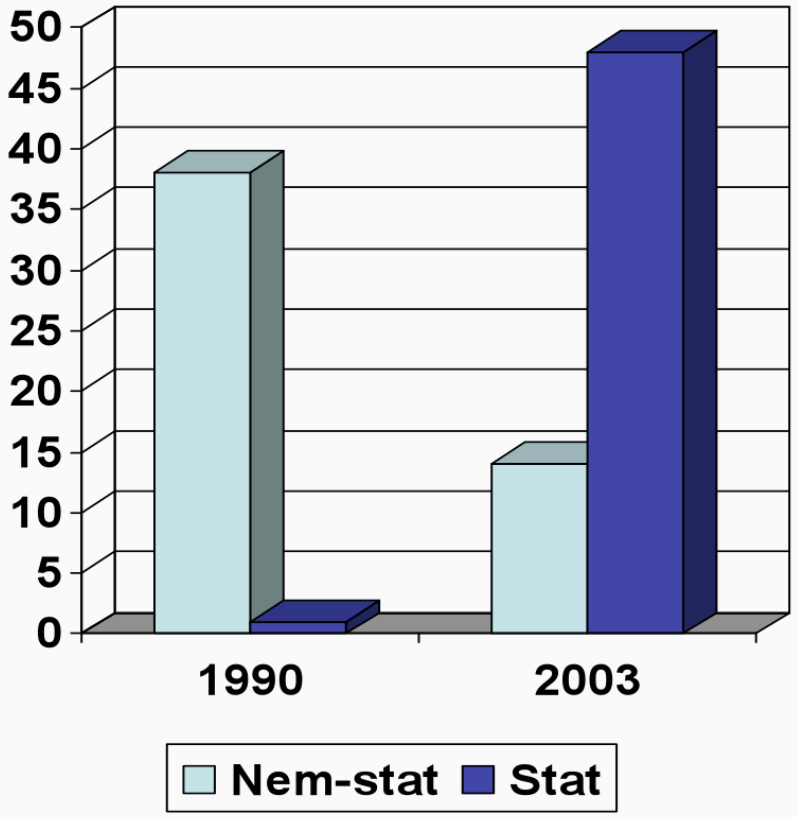
Statisztikai metodológia



Statisztikai metodológia

- empirista filozófiai tradíció (Locke)
- az érzékszervi tapasztalat prioritása → tudásunk elsődleges forrása a tapasztalat
- gyakoriségi adatokból indul ki, adatorientált
- a szövegből gépi tanuló algoritmus tanulja ki a szabályszerűségeket
- a grammatikalitási ítélet nem kétértékű, hanem fokozatai vannak

Összehasonlítás



Összehasonlítás :)

Noam Chomsky 1969

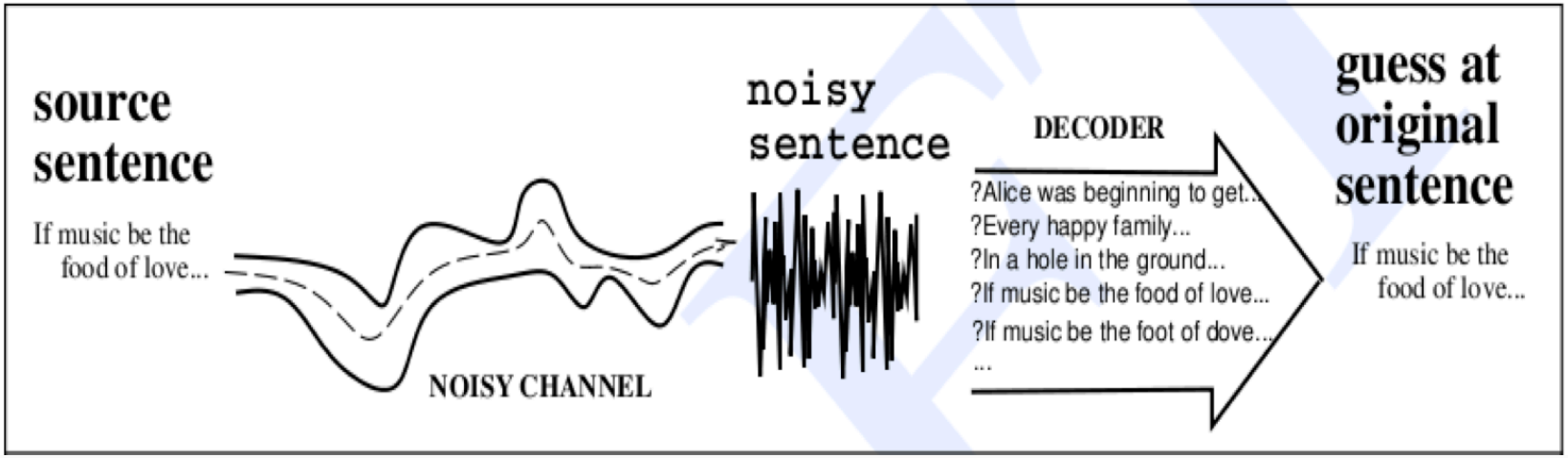
„Meg kell értsük, hogy egy mondat valószínűségéről beszélni teljesen értelmetlen.”

Fred Jelinek 1988

„Ahányszor távozik egy nyelvész a csoportból, felszökik a beszédfelismerési rátánk.”

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3):379–423.

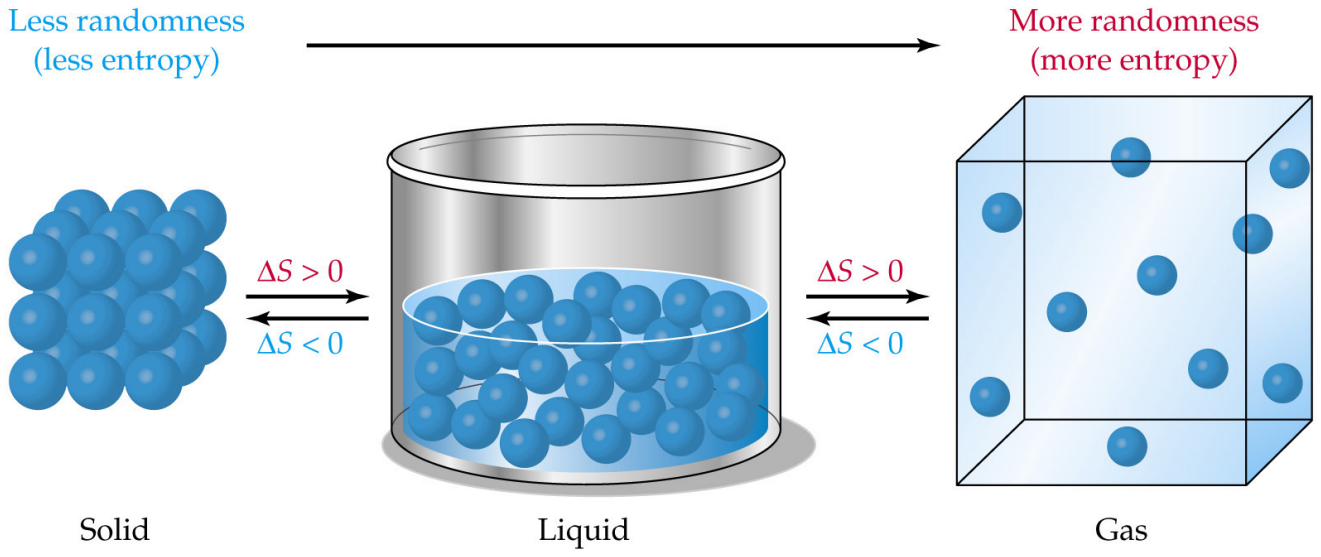
a természetesnyelv-feldolgozási problémák megfeleltethetők dekódolási problémáknak a zajos kommunikációs csatornában



Shannon, C. E. (1951). Prediction and Entropy of Printed English. Bell Systems Technical Journal, 30:50–64.

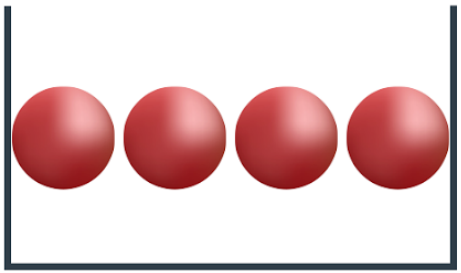
kikölcsönözte az entrópia fogalmát a termodinamikából, és a csatorna információs kapacitásának a mérésére alkalmazta → az információelmélet alapjai

a termodinamikai entrópia egy rendszer rendezetlenségi fokát jellemzi



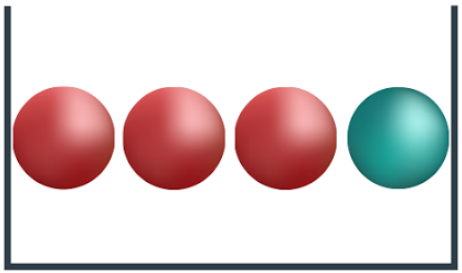
Az információelméleti entrópia

- az entrópia akkor a legkisebb (0), ha a hírforrás biztosan mindig ugyanazt a hírt sugározza → a bizonytalanságunk nulla, vagyis teljesen biztosak lehetünk benne, hogy az adott hír fog érkezni
- az entrópia akkor a legnagyobb, ha az összes hír valószínűsége egyenlő → ekkor a bizonytalanságunk a legnagyobb, hiszen bármelyik hír ugyanakkora valószínűséggel érkezhet



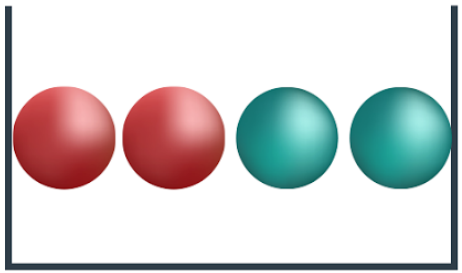
High Knowledge

Low Entropy



Medium Knowledge

Medium Entropy



Low Knowledge

High Entropy

Kornai videó

Chomsky

Chomsky, N. (1957). Syntactic Structures. Mouton, The Hague.
Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. Language, 35(1):26–58.

Újradefiniálta a nyelvészet feladatát: a nyelvésznek nem a nyelvi jelenségek leírása a feladata, hanem annak a vizsgálata, hogy hogyan tanulja meg a gyerek a nyelvet, és mik azok a jegyek, amelyek minden nyelvben közösek. Márpedig ezek a jelenségek a nyelv felszíni megjelenésétől igen távol esnek, így a „sekély" korpuszalapú módszerekkel nem elérhetőek.

A szabályalapú rendszerek hátrányai

- egy mondat lehetséges elemzéseinek a száma hatalmas → ahogy nő a mondat szavainak a száma, úgy exponenciálisan nő a lehetséges elemzések száma → számítástechnikailag nem volt kivitelezhető
- nem hibatűrő: 'Thanks for all you help.' (Abney, 1996)
- bonyolult a fejlesztése, törékeny
- nehezen átvihető más doménre vagy nyelvre

Abney, S. (1996). Statistical Methods and Linguistics. In Klavans, J. and Resnik, P., editors, The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pages 1–26. MIT Press.

Az első korpuszok & a sztochasztikus paradigma

- Brown Corpus (Kucera and Francis, 1967): was created in the US, which then inspired a whole family of corpora:
  - Lancester-Oslo-Bergen Corpus (Leech et al., 1983) (Brown's British English counterpart)
  - London–Lund Corpus (Svartvik, 1990)

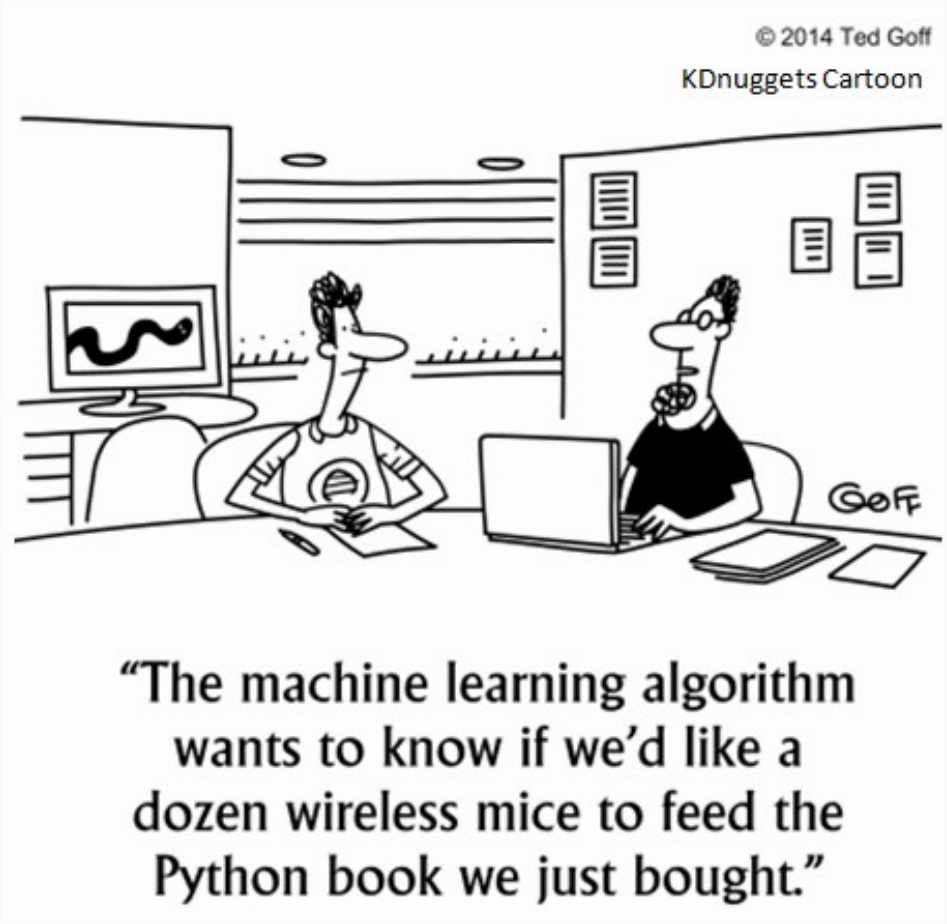
A sztochasztikus módszerek

a beszédfelismerés területén érték el az első sikereket, aztán onnan terjedtek tovább más NLP területekre, pl. POS taggelés (Bahl and Mercer, 1976).

Kilencvenes és kétezres évek

az empirizmus visszavág...

...oda, ahonnan jöttél



Kétezertizes évek

a neurális fordulat

- szóbeágyazások & neurális hálók → representation learning: az input szövegben fellelhető hasznos szóreprezentációkat automatikusan kitanulják
- self-supervised learning → automatikus reprezentációtanulás a kézi fícsörizálás helyett
- „deep learning”: azért mély, mert a neurális hálónak jellemzően több rétege van
- GPU & mátrixszorzás → neurális hálók
- big data
- párhuzamosítás


"That's all Folks!"