Bevezetés

Számítógépes nyelvészet – 2018 tavasz
1. óra

Simon Eszter – Mittelholcz Iván

MTA Nyelvtudományi Intézet

Tartalom

Bemutatkozás

Bemutatkozás

- mi
- ti

A félév bemutatása

Az órák

- összesen 14 hét
  - ebből 11-re van terv
  - két lauf
  - egy elmarad (tavaszi szünet)
- egy órán belül:
  - elméleti bevezetés slide-okkal
  - gyakorlatok gépen
  - házi feladat

Bevezetés a karakterkódolások rejtelmes világába

- Elmélet:
  - szöveges fájlok
  - karakterkódolás általában, karakterkódolás és fontkészlet
  - ASCII és kiegészítései
  - Unicode, Unicode kódolások (UTF és UCS)
  - karakterkódolás detektálása
  - konvertálás kódolások között
- Gyakorlat:
  - file és iconv parancsok
  - karakterkódolás python-ban
  - python2 és python3 közti különbségek

Bevezetés a héjak és szabályos kifejezések csodálatos világába

- Elmélet:
  - shell bevezetés
  - nyelvosztályok
  - regex elméleti alapok (reguláris nyelvek, automaták)
  - regex motorok működése, hatékonyság
- Gyakorlat:
  - sed, grep
  - python regex-ek
  - regexek és karakterkódolás

Automaták, FST, kétszintű morfológia

- Elmélet:
  - mi az automata, hogyan kell csinálni
  - mire lehet használni: különböző morfofonológiai feladatokra
  - Kimmo és a kétszintű morfológia
  - automaták implementálásának alapjai táblázattal
- Gyakorlat:
  - hfst-nek van olyan parancsa, amivel szabályokból FST-t lehet építeni
- Házi feladat:
  - automata építése, ami egy nyelv minden elemét legenerálja, és csak azt
  - szorgalmi: játékautomata leprogramozása pythonban

Korpuszépítés

- Elmélet:
  - a forrás módja: hang, írott, multimodális → innentől csak írott
  - forrás: papír Vs elektronikus → kép Vs szöveg → txt
  - az annotáció formátuma: inline (XML) vagy standoff (tsv & BIE1)
  - annotációs séma → annotációs útmutató
  - kézi annotálás, annotációs eszközök, inter-annotator agreement
- Gyakorlat:
  - crawling: wget, scrapy
  - boilerplate removal: beautifulsoup4
  - odt –> xml-ből kinyerés
  - docx, pdf: tika
  - kézi annotáció segítése: excel, ana2html, GATE
- Házi feladat:
  - játékkorpusz annotálása (NER v. NP-chunk v. dependencia) ketten vagy hárman, inter-annotator agreement számolása
  - szorgalmi: NLTK-ban van rá eszköz, azzal kiszámolgatni

Korpuszannotáció 1.

- Elmélet:
  - kézi vs. automatikus annotáció, gold vs. silver standard
  - az automatikus korpuszannotáló eszközök kiértékelése (P, R, F)
  - mondatra bontás, tokenizálás
  - morfológiai elemzés
  - egyértelműsítés
- Gyakorlat:
  - GATE vagy NLTK, polyglot?
- Házi feladat:
  - egy szöveg végigtolása egy elemzőláncon

Korpuszannotáció 2.

- Elmélet:
  - NER
  - sekély szintaktikai elemzés
  - szintaktikai elemzés (konstituencia és dependencia)
- Gyakorlat:
  - ?
- Házi feladat:
  - az ötödik hét kézzel annotált játékkorpusza legyen a gold standard $\rightarrow$ az e-magyar teljesítményének a kiértékelése ezeken a korpuszokon (precision, recall, f-measure)

Korpuszlekérdezések, -statisztika

- Elmélet:
  - alapfogalmak: korpusz, korpuszlekérdező motor, nyelvek és felület
  - lekérdező nyelvek: CQL (MNSZ), MQL (Emdros)
  - token–type, gyakoriság, relatív gyakoriság, MLE…
- Gyakorlat:
  - MNSZ-en vagy ómagyar korpuszon parancssorból lekérdezgetni dolgokat, valami egyszerűbb statisztikát számolni
- Házi feladat:
  - ómagyar korpuszon egy nyelvi jelenség diakrón vizsgálatát elvégezni: pl. a főnevek száma az egyes kódexekben, relatív gyakoriság, diagram

Gépi tanulás 1.

- Elmélet:
  - történeti kitekintés: szabályalapú vs. statisztikai módszerek
  - supervised és unsupervised tanulás
  - gold standard adat
  - train-devel-test halmazok, keresztvalidáció
  - feature extraction, n-gramok
  - modellépítés
  - taggelés
  - kiértékelés
- Gyakorlat:
  - huntagen végigpróbáljuk az egyes lépéseket
- Házi feladat:
  - NLTK-ban egy korpuszon egy tanuló algoritmussal valamit kipróbálni

Gépi tanulás 2.

- Elmélet:
  - Bayes-tétel, noisy channel, HMM
  - supervised algoritmusok: döntési fa, maxent, CRF, neurális háló stb.
  - unsupervised: klaszterezés
- Gyakorlat:
  - Scikit-learn-ben megnézni egy-két dolgot
- Házi feladat:
  - ?

Kitekintés

- Elmélet:
  - ontológia, linked open data, RDF
  - információkinyerés, NER
  - információ-visszakeresés
  - kulcsszókinyerés
  - metaforák és metonímiák felismerése
  - automatikus szótárgenerálás
  - sentiment analysis
  - gépi fordítás
- Gyakorlat:
  - -
- Házi feladat:
  - játék kulcsszókinyerés

Adminisztráció

Házi feladatok és teljesítés

- összesen 8 házi feladat kerül kiadásra
- ebből legalább 3-at kell beadni a teljesítéshez
- a feladatokból 4 megoldható programozási tudás nélkül is
- a beadott házikra megajánlott jegyet lehet kapni
- akinek ez nem jó, írhat javító ZH-t

Tanrend & elmaradó órák

ELTE BTK tanrend:

- szorgalmi időszak első napja: 2018. február 12. (hétfő)
- tavaszi szünet: 2018. március 28. – április 3. (szerda–kedd)
- utolsó tanítási nap: 2018. május 18. (péntek)
- vizsgaidőszak első napja: 2018. május 22. (kedd)
- vizsgaidőszak utolsó napja: 2018. július 6. (péntek)

Elmaradó órák:

- március 14.: elutazunk $\rightarrow$ pótló alkalom $\rightarrow$ doodle
- március 28.: tavaszi szünet

Technikai részletek

Technikai részletek

OS

- unix-like oprendszerek preferáltak (Linux, OS X)
- windows:
  - cygwin
  - Windows Subsystem for Linux
  - VirtualBox + Linux (Debian, Ubuntu, CentOS)

Shell

- Linux, OS X: ✓
- online: Unix Terminal Online

Python 3

- Linux, OS X: ✓
- Windows: python
- Anaconda
- online lehetőségek: PythonAnywhere, repl.it

Repó

https://github.com/m-ivan/compling

Git

- TryGit, The Simple Guide
- `git clone https://github.com/m-ivan/compling.git`

Jupyter Notebook

- tutorial
- `pip install jupyter` vagy `pip3 install jupyter`
- Anacondában elvileg benne van – ha mégsem: `conda install jupyter`

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

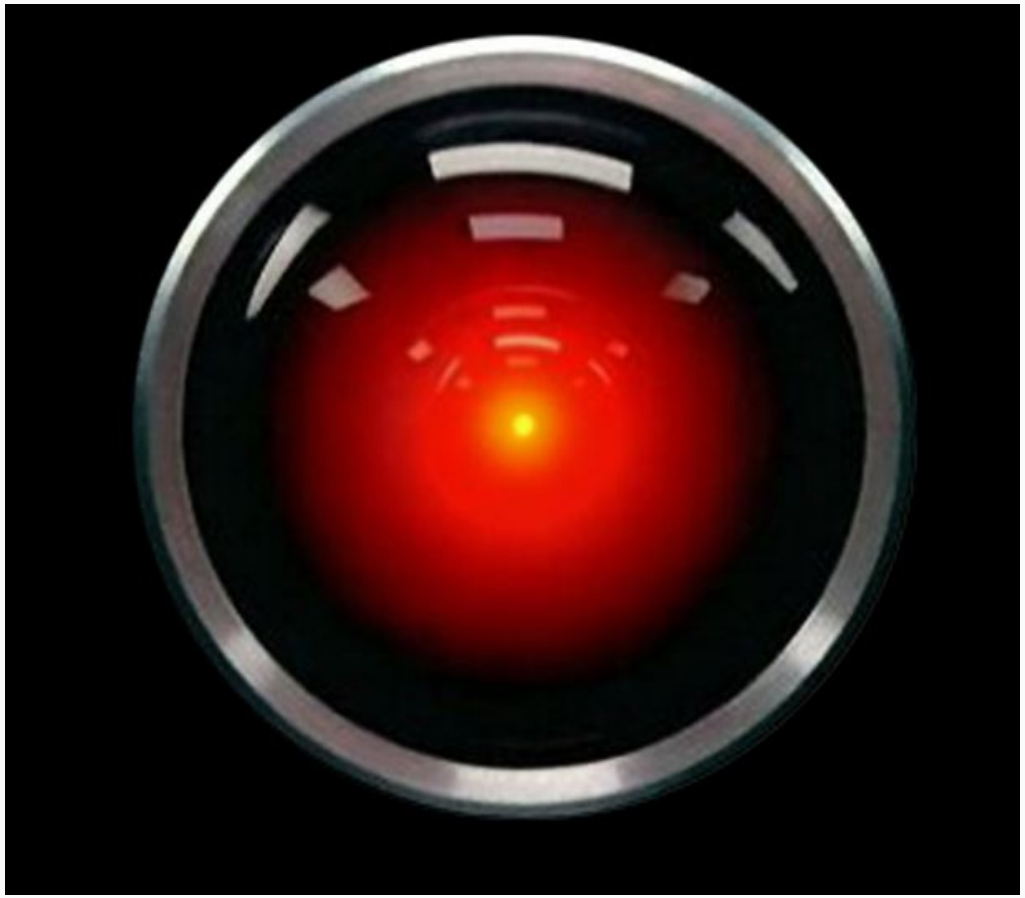
Definíció

- 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

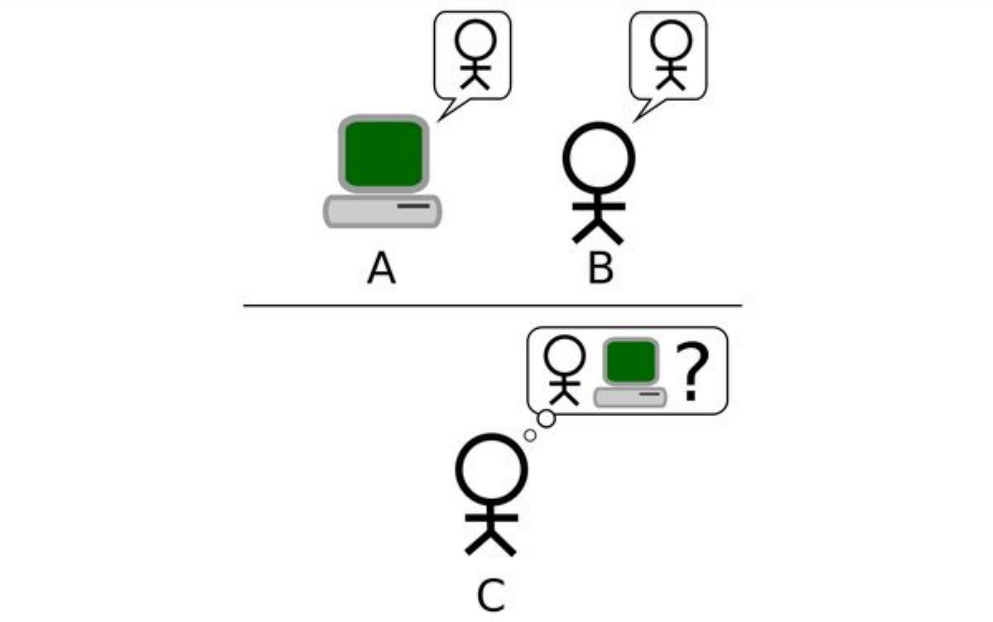
A problémák

- 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

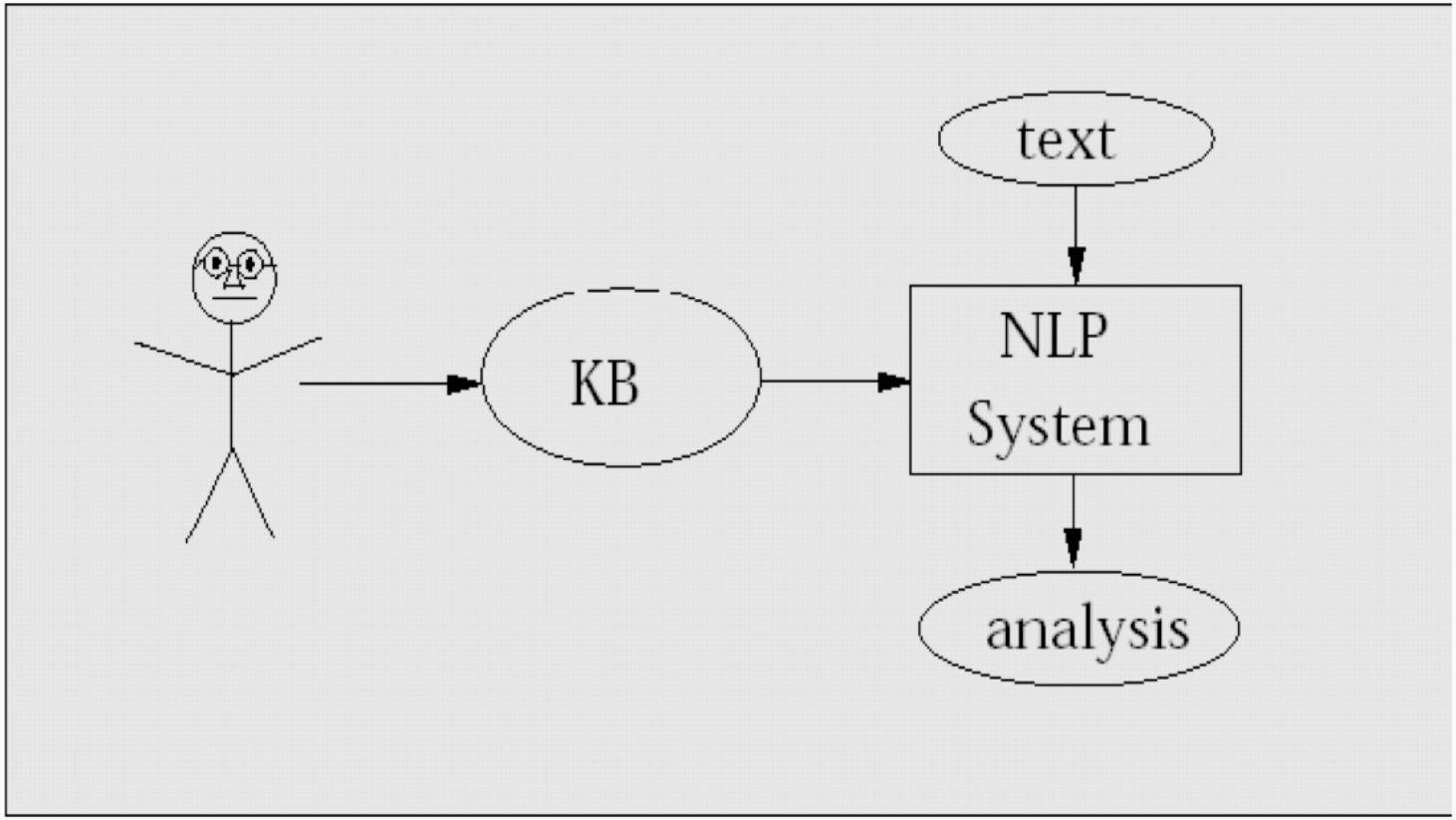
Carl Rogers (1902-1987):

- 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



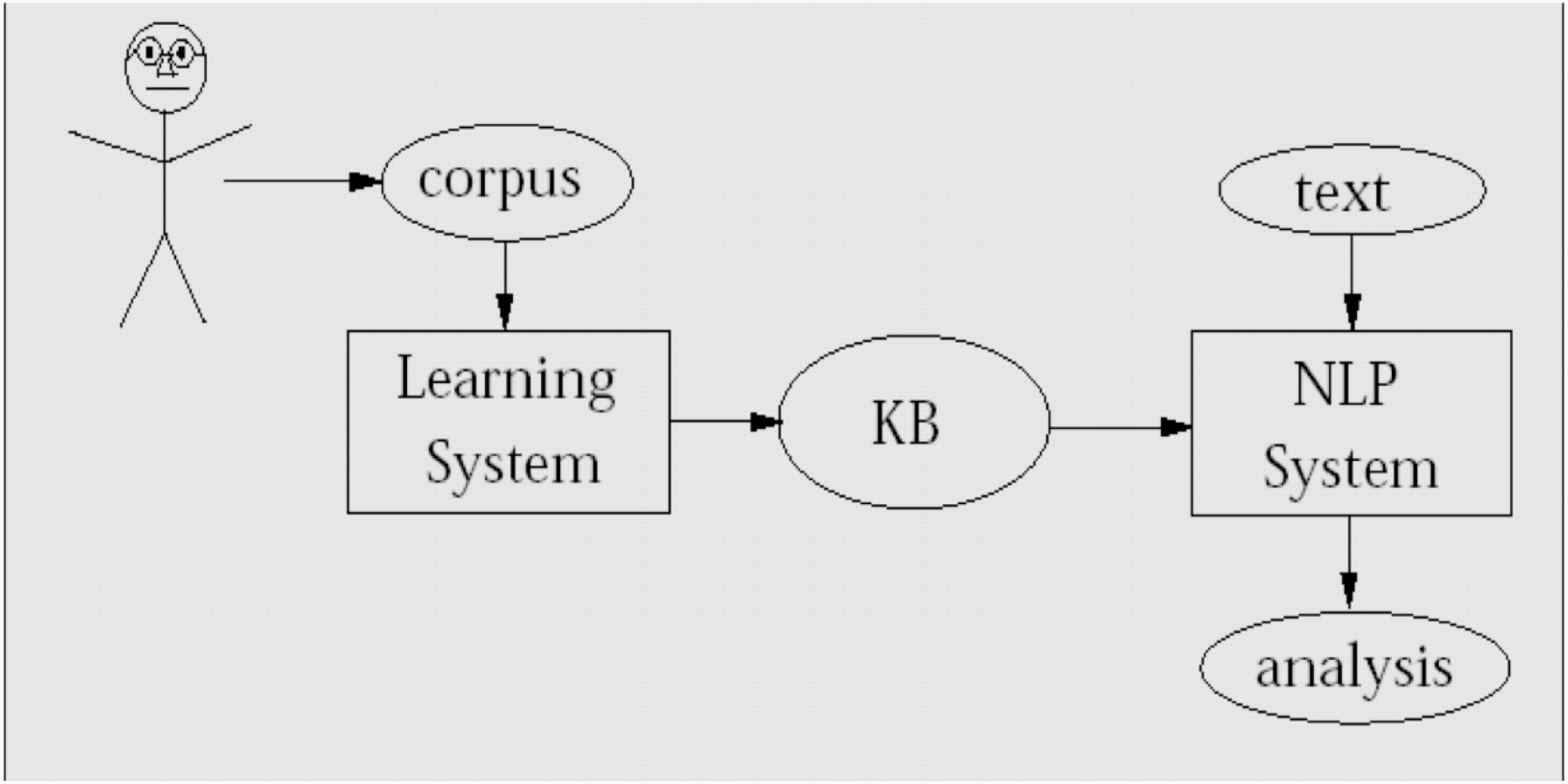
Szabályalapú metodológia

- racionalista filozófiai tradíció (Leibniz, Descartes)
- univerzális nyelvtan
- velünk született nyelvi képesség $\rightarrow$ 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

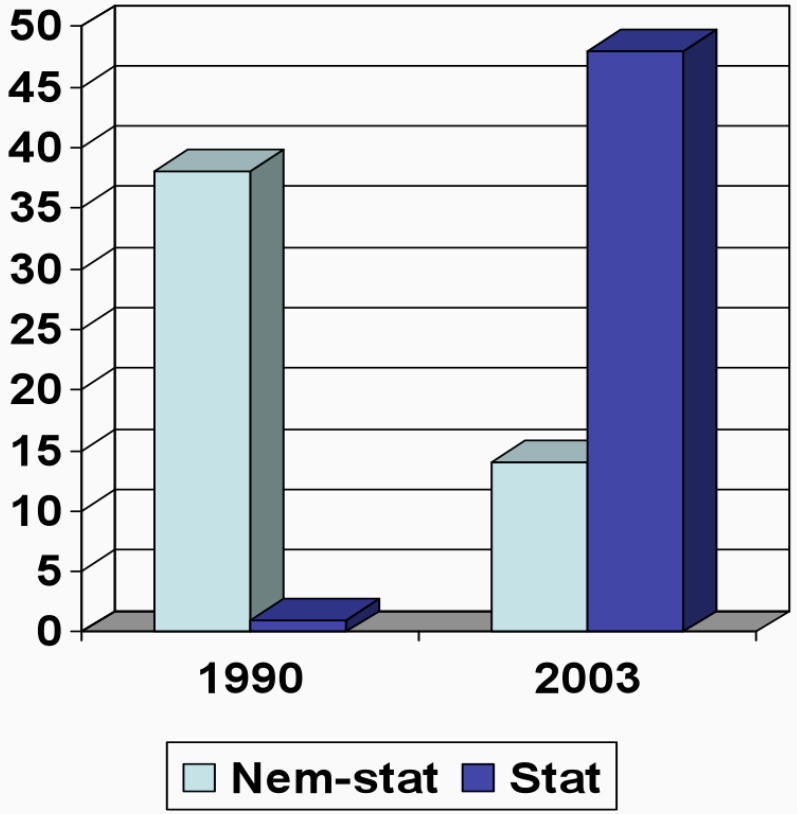
Statisztikai metodológia



Statisztikai metodológia

- empirista filozófiai tradíció (Locke)
- az érzékszervi tapasztalat prioritása → tudásunk elsődleges forrása a tapasztalat
- gyakorisági 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

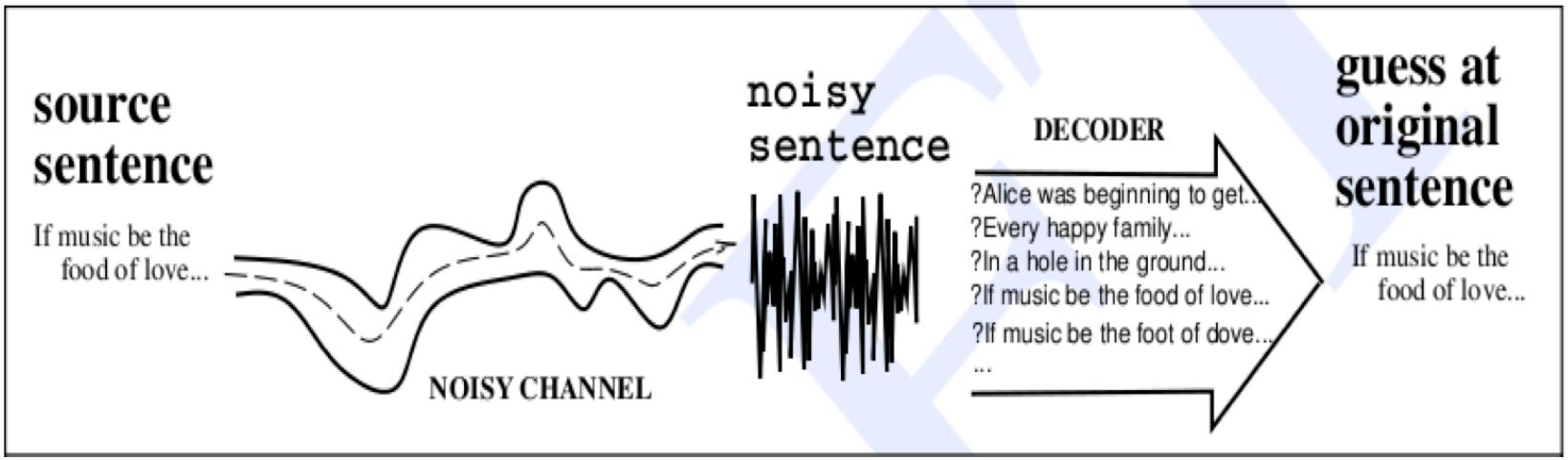
“Meg kell értsük, hogy egy mondat valószínűségéről beszélni teljesen értelmetlen.”

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.

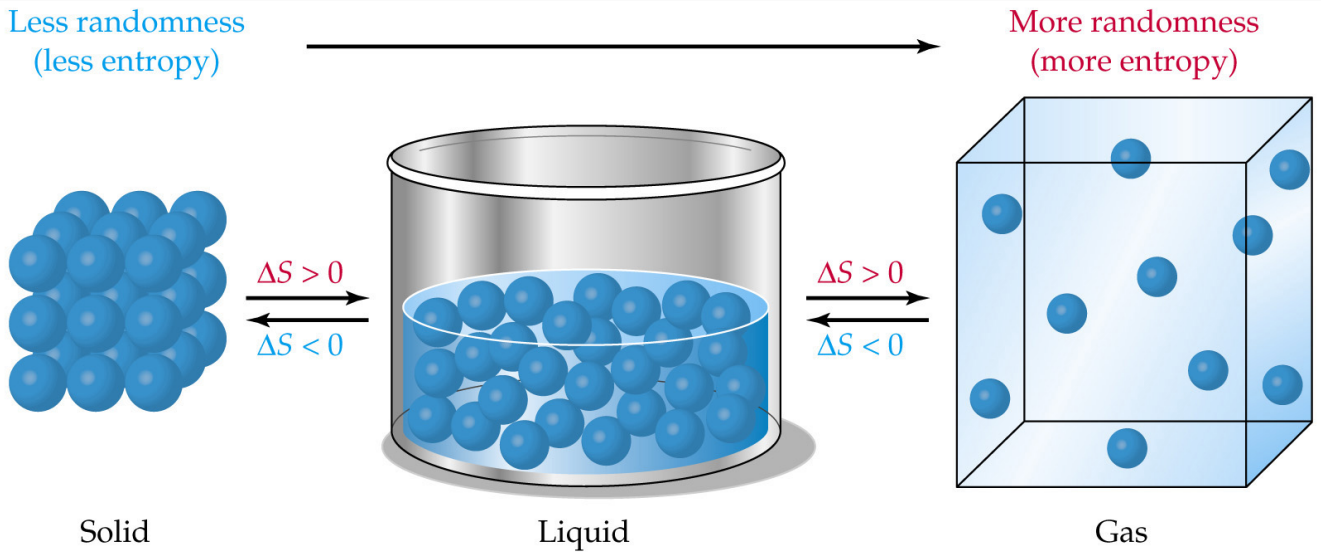
a természetesnyelv-feldolgozási problémák megfeleltethetők dekódolási problémáknak a zajos kommunikációs csatornában



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

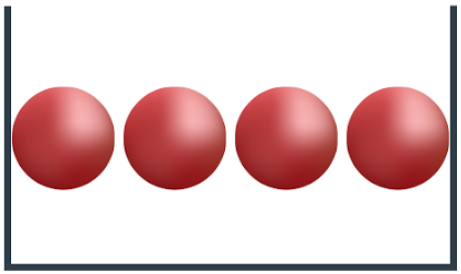
kikölcsönözte az entrópia fogalmát a termodinamikából, és a csatorna információs kapacitásának a mérésére alkalmazta → az információelmélet alapjai

a termodinamikai entrópia egy rendszer rendezetlenségi fokát jellemzi



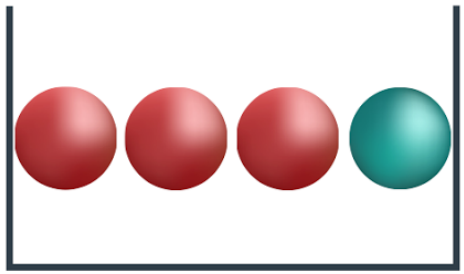
Az információelméleti entrópia

- 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



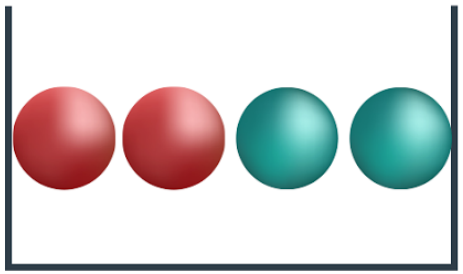
High Knowledge

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)

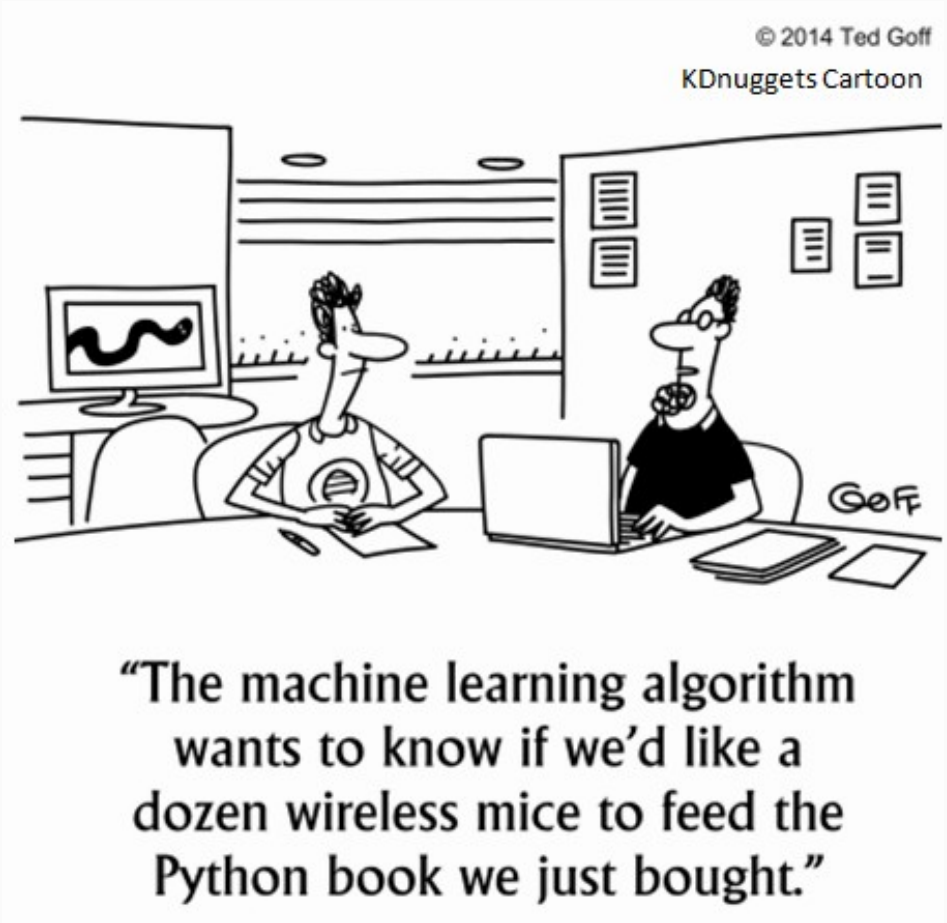
A sztochasztikus módszerek

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

Kilencvenes évek

az empirizmus visszavág...

...oda, ahonnan jöttél




"That's all Folks!"