Számítógépes nyelvészet bevezető

Nyelv és informatika – Pécs, 2022/23 tavasz
1. óra

Simon Eszter – Vadász Noémi

2023. február 18.

Tartalom

1. Bemutatkozás

2. A félév bemutatása

3. Bevezetés a számítógépes nyelvészetbe

4. Kis történeti áttekintés

    Az MI-kutatás kezdetei

5. Módszerek

    Szabályalapú és statisztikai metodológia

Bemutatkozás

Bemutatkozás

- mi
- ti

A félév bemutatása

Tanárok és alkalmak

Tanárok:

- helyiek: Kleiber Judit, Viszket Anita
- gyüttmentek: Simon Eszter, Vadász Noémi

Alkalmak:

- február 18. 9:15-16:45: tömbösített órák
- március 18.: (regisztrált) részvétel a MANYE kongresszuson, melynek témái: tudásmegosztás, információkezelés, alkalmazhatóság: https://semmelweis.hu/szaknyelv/manye-xxix/
- április 1. 9:30-17:30: tömbösített órák

Teljesítés feltétele

- aktív órai részvétel
- 2500n beszámoló a MANYE-n meghallgatott előadásokról
- felkészülés egyik alkalomról a másikra: egy nyelvtechnológiai alkalmazás rövid bemutatása VAGY teszt az ápr. 3-i héten, egy közösen egyeztetett időpontban

A mai alkalom

- 9:15-10:45 óra
- 10:45-11:00 szünet
- 11:00-12:30 óra
- 12:30-13:30 ebéd
- 13:30-15:00 óra
- 15:00-15:15 szünet
- 15:15-16:45 óra

Tematika

1. Számítógépes nyelvészet bevezető
    - Bemutatkozás
    - A félév áttekintése
    - Számítógépes nyelvészet bevezető: definíció, története stb.
2. Bevezetés a korpuszok csodálatos világába
    - Mik azok a korpuszok?
    - Korpuszépítés
    - Korpuszannotáció
    - Annotációs szintek
3. Számítógépes morfológia
    - Morfológiai elemzés transzducerekkel, kétszintes morfológia
    - Egyéb módszerek: folytatási osztályok, unifikációs modellek
    - Számítógépes morfológiai elemzés
    - Címkekészletek

Tematika – folyt.

4. Korpuszlekérdezés
    - Magyar Nemzeti Szövegtár 2
    - Ómagyar Korpusz
    - Parallel Bible
    - Mazsola
5. Korpuszépítés
    - NerKor
    - KorKor
    - Csángó korpusz
    - Ómagyar Korpusz
    - UraLUID korpusz
6. Gépi tanulás
    - A gépi tanulás forgatókönyve
    - A tulajdonnév-felismerésen keresztül példázva

Tematika – folyt.

7. Magasabb szintű feldolgozás
   - Koreferencia-feloldás
   - KorKor korpusz
   - Winograd-sémák
8. Magyar szövegfeldolgozó eszközök
   - *emtsv*
   - magyarlánc
   - huSpaCy
   - UDPipe

Ajánlott irodalom

- Dan Jurafsky, James H. Martin: Speech and Language Processing. 3rd ed. draft: https://web.stanford.edu/~jurafsky/slp3/
- Alberti Gábor: Matematika a természetes nyelvek leírásában. Tinta Könyvkiadó, Budapest, 2006.
- Prószéky Gábor, Kis Balázs: Számítógéppel emberi nyelven. Szak Kiadó, Budapest, 1999.
- Szirmai Monika: Bevezetés a korpusznyelvészetbe. Tinta Könyvkiadó, Budapest, 2005. http://korpusz.com/Monika/Bevezetes_a_korpusznyelveszetbe.html
- Bálint Sass: Principles of corpus querying: A discussion note. In: Acta Linguistica Academica 69/4. 2022. https://akjournals.com/view/journals/2062/69/4/article-p599.xml?body=pdf-24714

Technikai részletek

Hozzatok gépet!

Minden elérhető lesz a kurzus GitHub repójában:
https://github.com/esztersimon/nlp_at_pecs

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

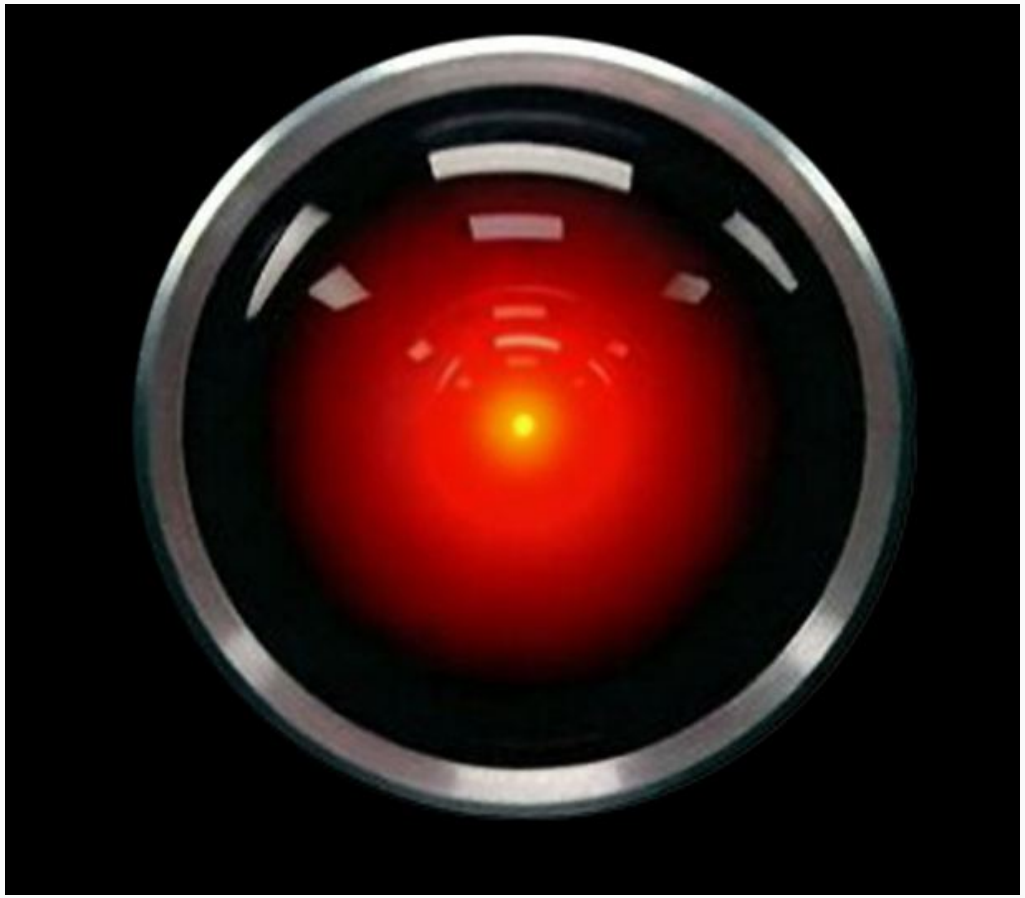
Definíció

- 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

Praktikus gépi alkalmazások

- személyi asszisztensek: Siri, Alexa, Cortana, Google Assistant
- auto-complete
- spell checking: böngészők, editorok, programok (Microsoft Word)
- gépi fordítás: Google Translate, DeepL
- chatbotok, ChatGPT
- szentimentelemzés (pozitív, negatív és semleges értékelések)
- google calendar bejegyzés emailekből

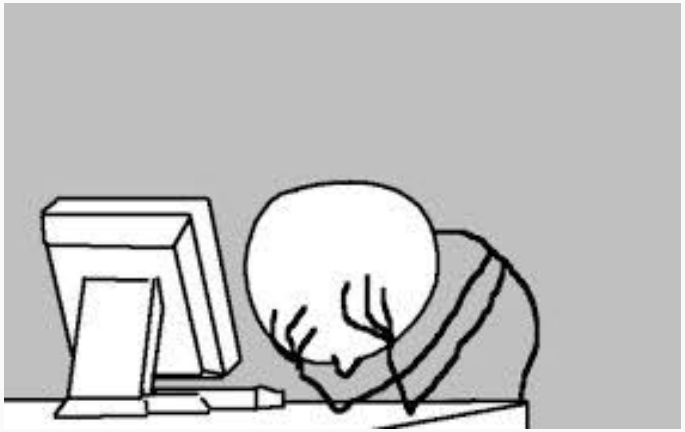
Részterületek

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



Hogy állunk az egyes részterületeken?

nagyon jól:

- spamszűrés
- POS-taggelés
- névelemfelismerés (NER)

egész jól:

- szentimentelemzés
- koreferenciafeloldás
- jelentésegyértelműsítés (WSD)
- mondatelemzés, parsing
- gépi fordítás (MT)
- információkinyerés (IE)

még mindig nem valami jól:

- kérdésmegválaszolás (QA)
- kivonatolás
- dialógus

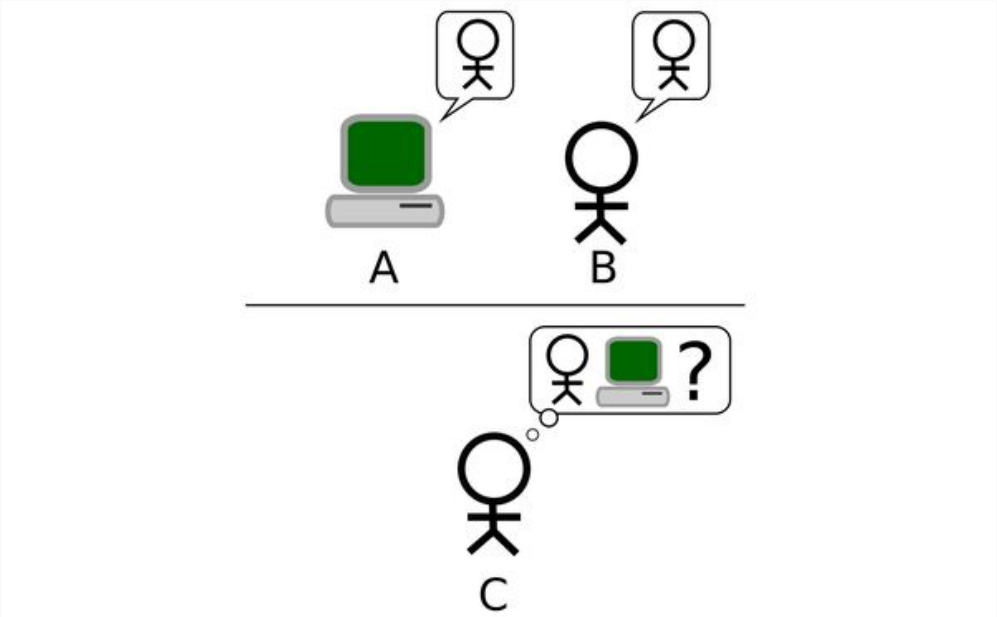
Kis történeti áttekintés

Történelemkönyv

- 1950-60: az első ötletek
- 1960-70: kísérletezés
- 1970-80: használható gépek
- 1980-90: növekvő kapacitás, termékek
- 1990-: új technológiák, kommunikáció
- 2000-: növekvő szövegmennyiség, ipar
- 2010- : internet

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

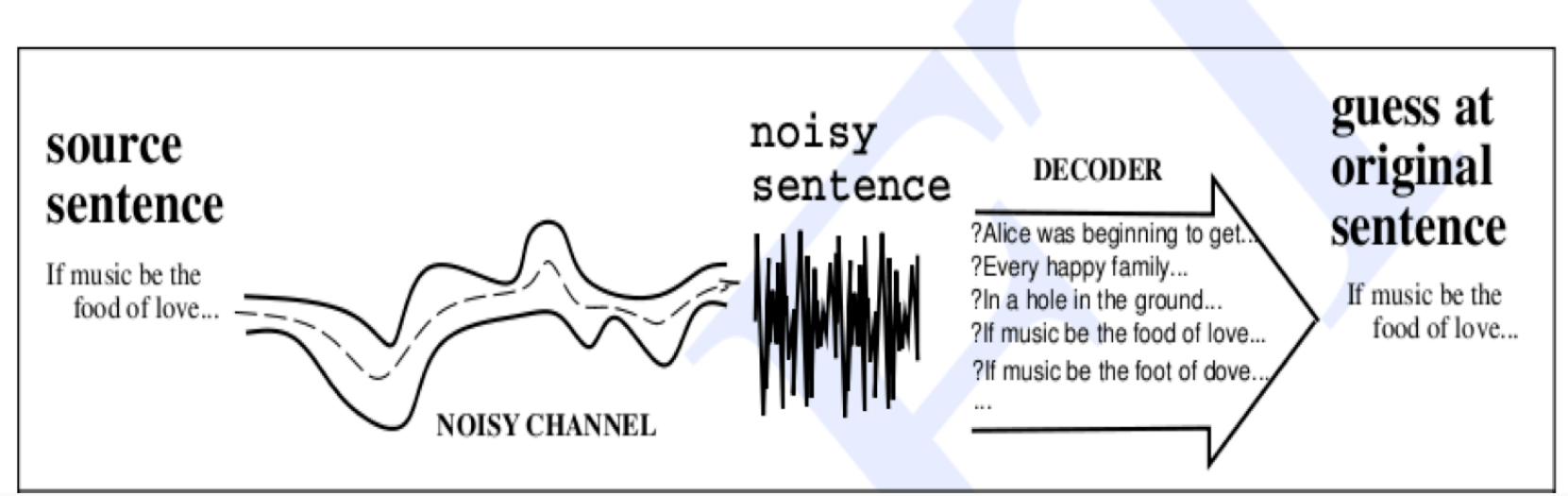


A Turing-teszt kritikája

- 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)

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

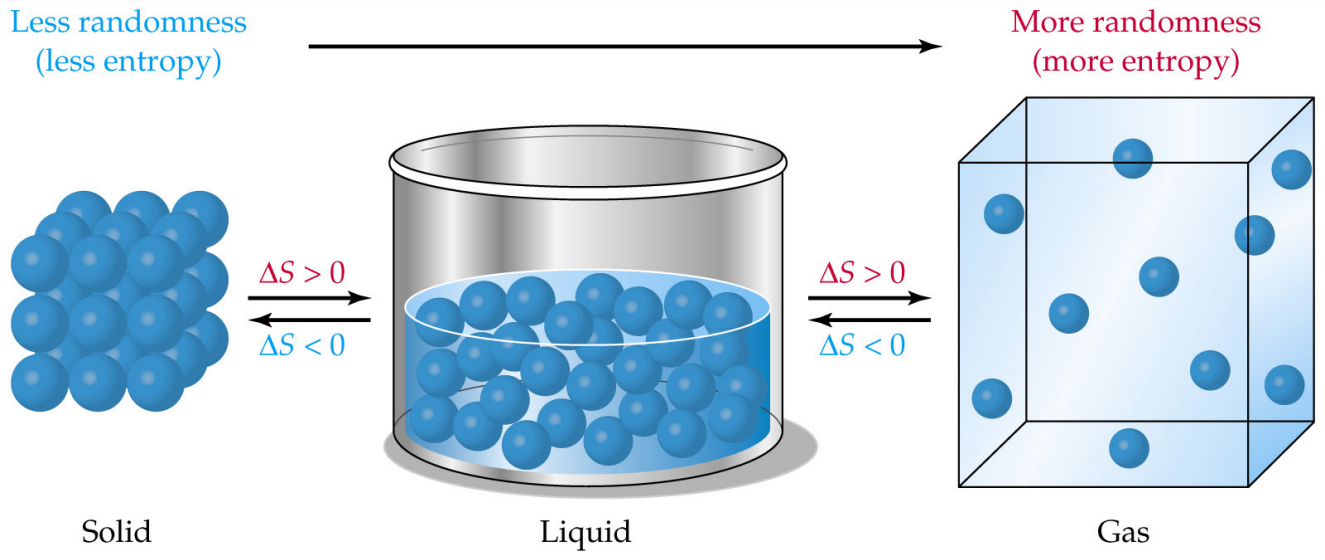
a természetesnyelv-feldolgozási problémák megfeleltethetők dekódolási problémáknak a zajos kommunikációs csatornában



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

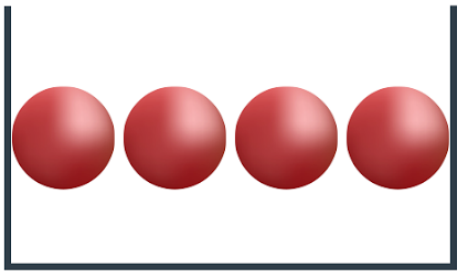
kikölcsönözte az entrópia fogalmát a termodinamikából, és a csatorna információs kapacitásának a mérésére alkalmazta → az információelmélet alapjai

a termodinamikai entrópia egy rendszer rendezetlenségi fokát jellemzi



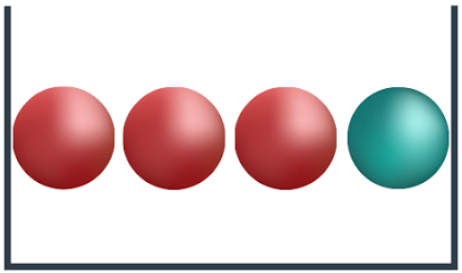
Az információelméleti entrópia

- 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



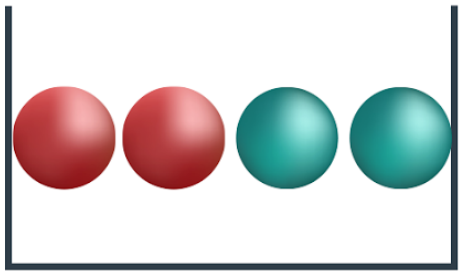
High Knowledge

Low Entropy



Medium Knowledge

Medium Entropy



Low Knowledge

High Entropy

Kornai videó

A Georgetown–IBM kísérlet (1954)

- teljesen automatikus gépi fordítás
- több mint 60 orosz mondatot képes angolra fordítani
- szabályalapú, szótáralapú (a szavakhoz spec. szabályok kapcsolódnak)
  - Operation 0 – An exact equivalent for a translated item exists. Any further steps needed.
  - Operation 1 – Rearrangement of the position of the words. AB > BA
  - Operation 2 – The several choices problem. The result is based on the consecutive words (maximum of three).
  - Operation 3 – Also several problems. But the result depends on the previous words (maximum of three).
  - Operation 4 – Omissions of the lexical (morphological) item. The source item would be redundant.
  - Operation 5 – Insertion of the lexical (morphological) item. The item is not present in the output language.

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 gépi fordítás

- az 50-es évek nagy slágertémája
- *The spirit is willing but the flesh is weak.*

  $\rightarrow$ orosz $\rightarrow$ angol $\rightarrow$

  The vodka is excellent but the meat is rotten.
- ALPAC (Automatic Language Processing Advisory Committee) report (1966): szkeptikus az eddigi eredményekkel kapcsolatban, az USA kormánya drámaian csökkenti a finanszírozást
- statisztikai módszerek
  - kapacitásnövekedés $\rightarrow$ egyre több szöveg válik elérhetővé, kereshetővé
  - sparse data problem: mindig lesznek olyan jelenségek, melyek megfelelő működésének kimutatásához nem elegendő az aktuális méretű adathalmaz
- neurális hálók, vektoros reprezentációk, mélytanulás
  - 2010-es évek: grafikus processzorok (GPU): hatékonyabb számítási teljesítmény + rengeteg adat

STUDENT (1964)

- Daniel G. Bobrow
- középiskolai szöveges matekfeladatokat olvas és old meg
- bemenet: angol szöveg, kimenet: szám
- pl. ilyeneket: *If the number of customers Tom gets is twice the square of 20% of the number of advertisements he runs, and the number of advertisements is 45, then what is the number of customers Tom gets?*
- a mesterséges intelligencia egy korai példája
- szabályalapú, logiaki következtetés

ELIZA (1966)

- Joseph Weizenbaum, MIT Artificial Intelligence Laboratory
- az első chatbot
- mintaillesztés, behelyettesítés

```
  Welcome to
                EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

https://web.njit.edu/~ronkowit/eliza.html

ELIZA

- 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/*

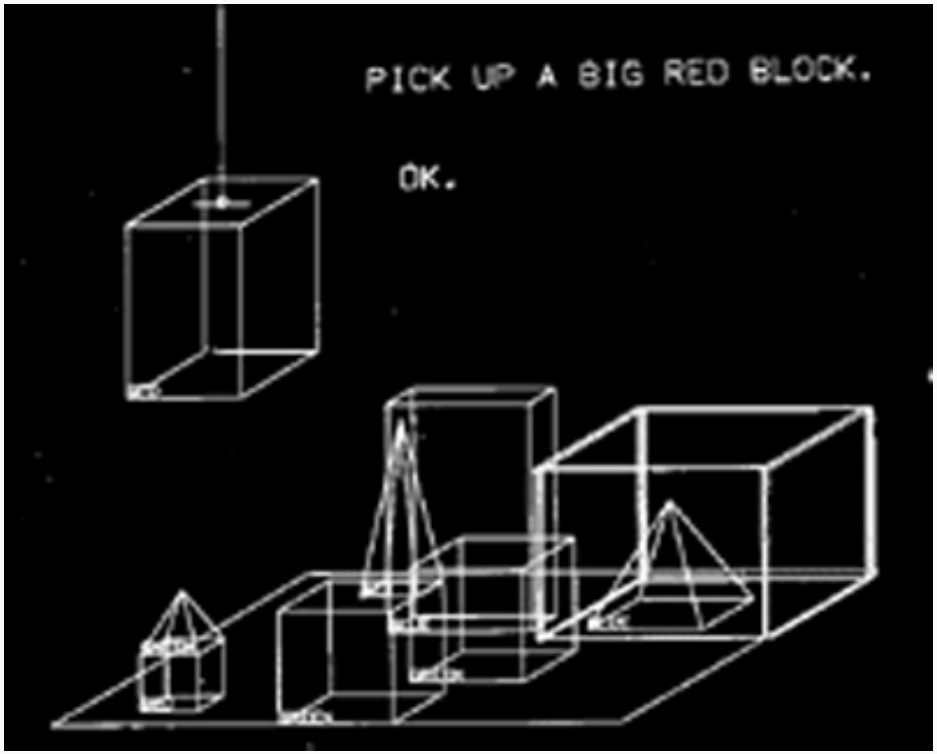
Az első korpuszok & a sztochasztikus paradigma

- 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).

SHRDLU (1970)

- Terry Winograd, MIT
- nyelvfeldolgozó, interakció a userrel angol kifejezéseken keresztül
- memória, statika, névadás
- az első interakciós fikció



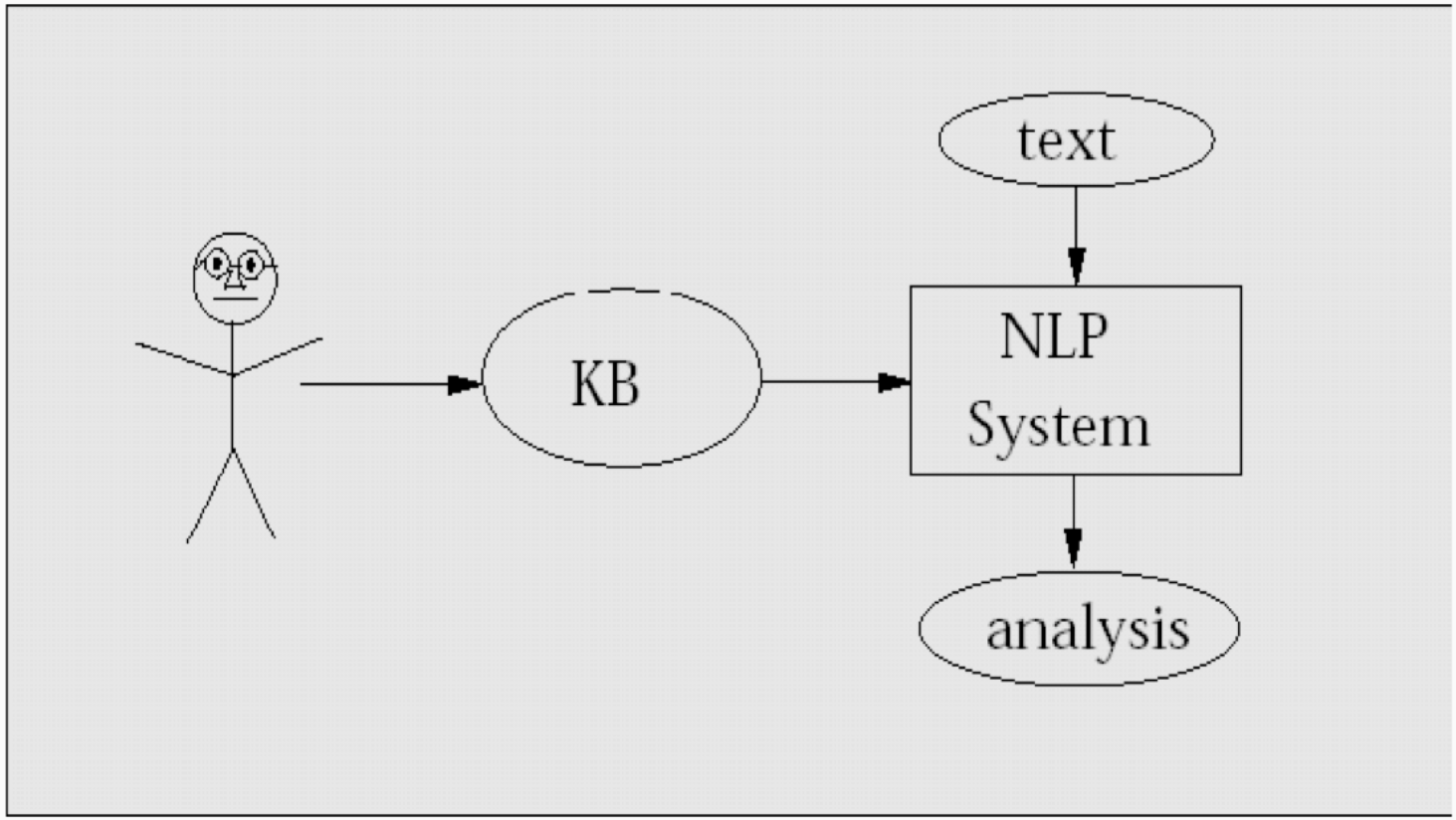
https://www.youtube.com/watch?v=bo4RvYJYOzI

Chatbotok, asszisztensek és az 1 millió dolláros főnyeremény

- 2006: Watson (IBM): 2011-ben megnyeri a Jeopardy!-t
- 2011: Siri (Apple)
- 2014: Cortana (Microsoft), Alexa (Amazon)
- 2016: Google Assistant
- 2022: ChatGPT

Módszerek

Szabályalapú metodológia



Szabályalapú

☺ a fejlesztőnek nagy kontrollja van a rendszer fölött
☺ könnyen értelmezhető visszacsatolás
☺ magas pontosság
☺ nyelvi adatok, amik könnyen megragadhatók szabályokkal (reguláris kifejezésekkel), pl. dátumok szerkezete

☹ sok kézimunka, nagy szakértelem kell hozzá
☹ nem hibatűrő
☹ bonyoult a fejlesztése, törékeny
☹ nehezen átvihető más doménre, nyelvre
☹ lehetetlen olyan szabályrendszert írni, ami mindent lefed, amit kell, de semmit, amit nem
☹ a fedés a listák és a szabályok számának növelésével javítható, de a szabályok száma, a lexikon mérete korlátozott

- pl. morfológiai elemzés, tokenizálás

Szabályalapú metodológia

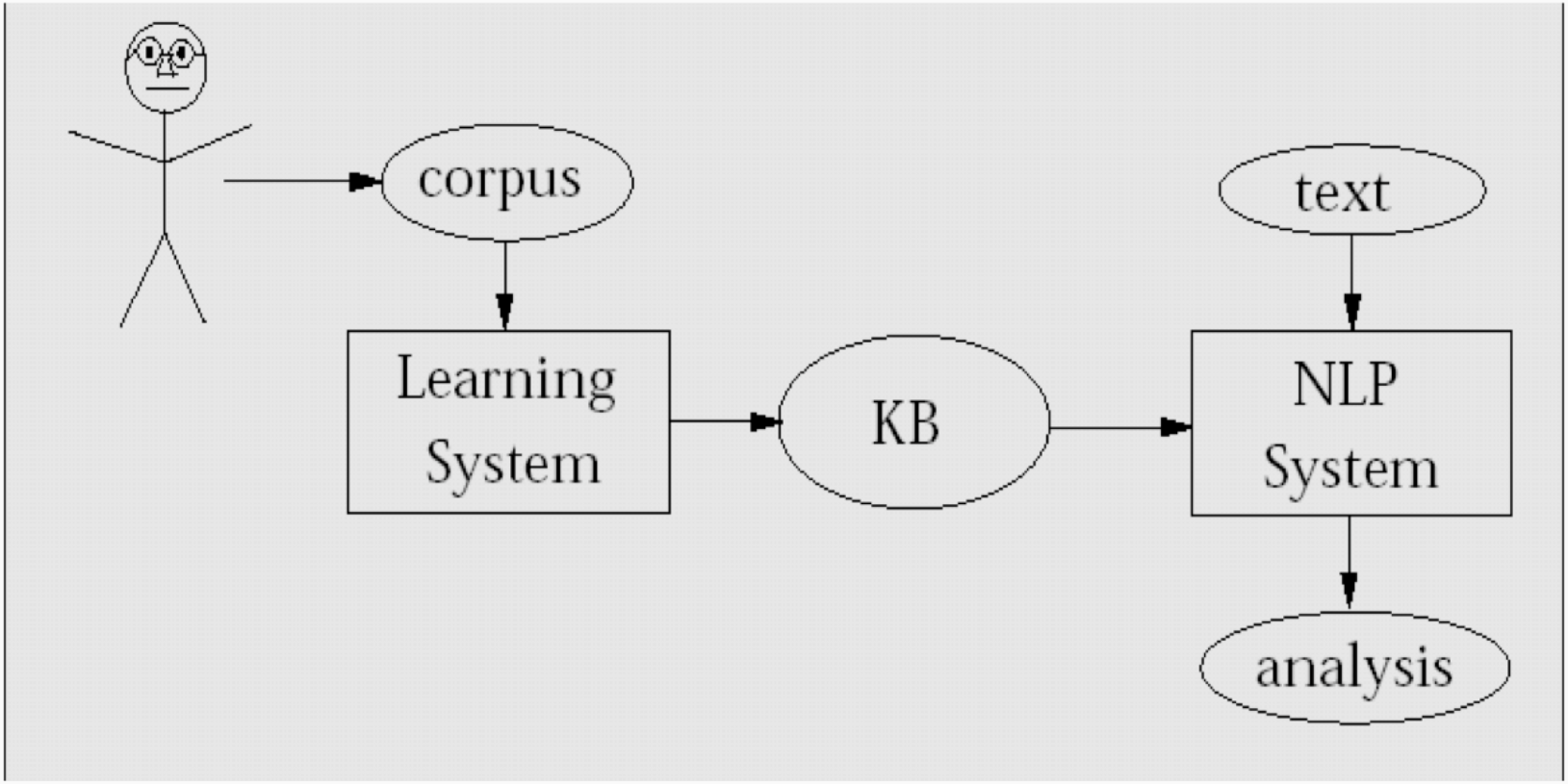
- racionalista filozófiai tradíció (Leibniz, Descartes)
- univerzális nyelvtan
- velünk született nyelvi képesség $\rightarrow$ introspekció
- grammatikalitási ítélet: 0 vagy 1
- kézzel kódolt szabályok
  - reguláris kifejezések

Példák

e-mail cím: [a-z]+@[a-z]+\.[a-z]+

pl.: bubo@doktor.hu

Statisztikai metodológia



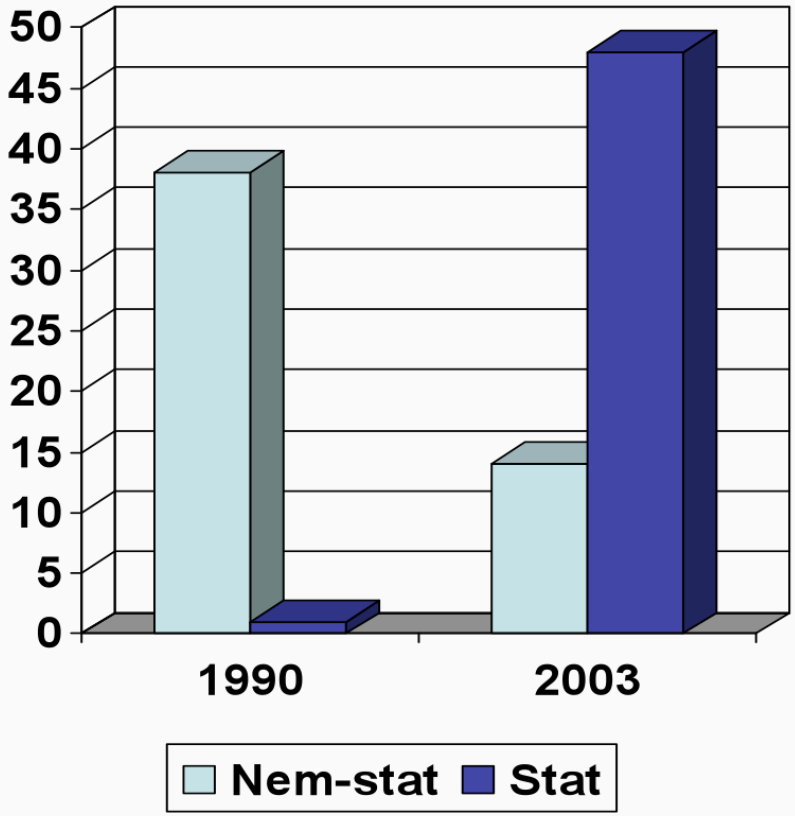
Statisztikai (sztochasztikus), klasszikus gépi tanulás

- adatorientált, gyakorisági adatokból indul ki
- a nyelv általánosabb megértése, modellálása
- kézzel kinyert feature-ökre támaszkodik (pl. mondathossz, POS-tagek, spec. szavak előfordulása)
- gépi tanuló algoritmusok (pl. Naive Bayes, SWM, döntési fa stb.)
- nehézség: az adatok ritkasága (“rare words are very common”)
- pl. szekvenciális címkézési feladatok, szintaktikai elemzés

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



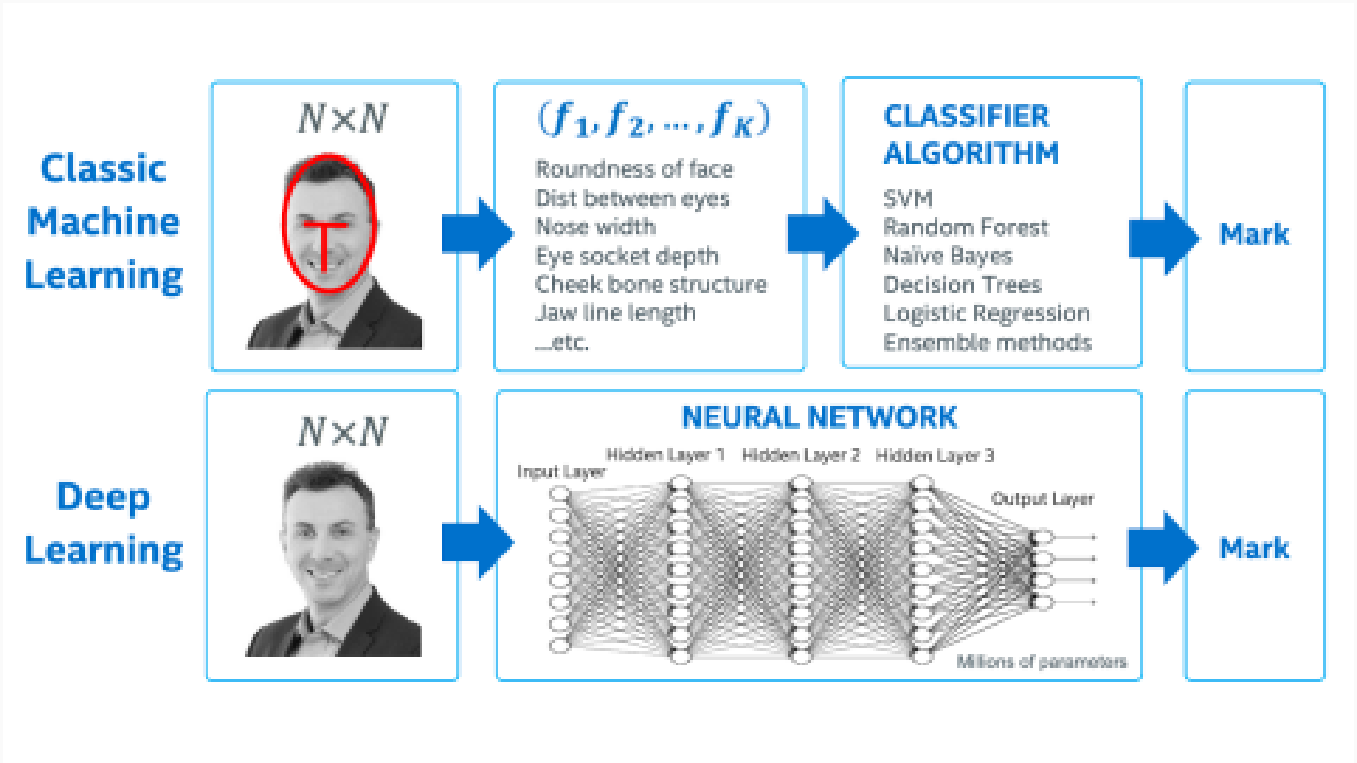
Összehasonlítás :)

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.”

Neurális

- 2010-es évek óta ez a legforróbb terület
- nincsenek kézzel kinyert jegyek (self-supervised learning)
- end-to-end modellek
- GPU-k, párhuzamosítás, nagyobb számítási kapacitás
- deep learning: azért “mély”, mert a neurális hálónak ált. több rétege van




"That's all Folks!"