Sanaluokkien automaattisen tunnistamisen menetelmät

Aleksi Pekkala

aleksi.v.a.pekkala@student.jyu.fi

Tietotekniikan kandidaattiseminaari

4.12.2013

Sisällys

- Johdanto
- Haasteet
- Käyttötarkoitus
- Menetelmät:
  - Sääntöpohjaiset
  - Tilastolliset

Johdanto

- Sanaluokan automaattinen tunnistaminen tekstiyhteyden perusteella
- […] he/PPS said/VBD it/PPS would/MD force/VB banks/NNS to/TO violate/VB their/PP$ contractual/JJ obligations/NNS with/IN depositors/NNS […]

Haasteet

- Monitulkintaisuus:
  - *Time flies like an arrow*
  - *Fruit flies like a banana*
  - Apuna lokaalit sekä kontekstuaaliset vihjeet
- Tuntemattomat sanat
- Nykyinen suorituskyky noin 97%
  - Lähtötaso on 90%
  - Kokonaisten lauseiden tunnistustarkkuus ~56%

Käyttötarkoitus

- Tärkeä **esikäsittelyvaihe** luonnollisten kielten prosessointiketjussa
- Ei ratkaise tekstin monitulkintaisuutta, mutta rajaa mahdollisten tulkintojen määrää

Google

Sign in

Translate

From: French - detected | To: English | Translate

English Spanish French

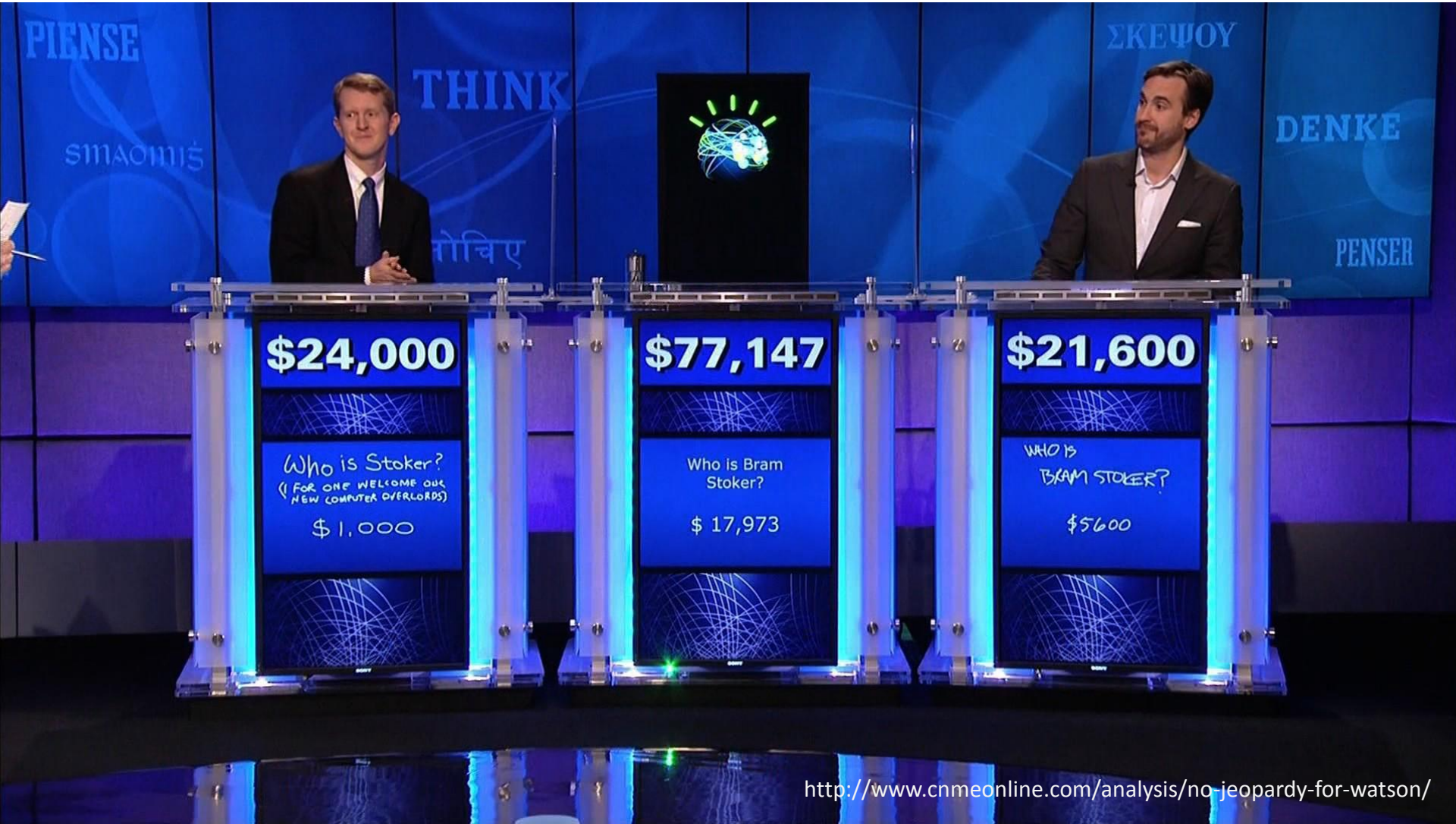
comme un éléphant dans un magasin de porcelaine

English French Spanish

like a bull in a china shop

New! Click the words above to edit and view alternate translations. Dismiss

William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel



http://www.cnmeonline.com/analysis/no-jeopardy-for-watson/

Sääntöpohjaiset menetelmät

- Kielitieteellinen näkökulma
- Sanaluokan yksikäsitteistämiseen käytetään käsin koottuja sääntöjä
  - *korvaa substantiivi erisnimellä, jos sanalla on iso alkukirjain*
  - *korvaa substantiivi verbillä, jos edeltävä sanaluokka on pronomini*
- Sääntöjen kokoaminen työlästä
  - Brillin (1992) tunnistin

Tilastolliset menetelmät

- Sanaluokkien tunnistaminen on **sarjanluokitteluongelma:**

$$\underset{t_1, t_2, \ldots, t_n}{\operatorname{argmax}} P(t_1, t_2, \ldots, t_n \mid w_1, w_2, \ldots, w_n)$$

- Voidaan ratkaista yleisillä koneoppimismenetelmillä, mm.
  - Markovin piilomallit
  - Log-lineaariset mallit

Kiitos!

Kysyttävää?