INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Andre forelesning

Lilja Øvrelid

23 januar, 2017

Språkteknologi

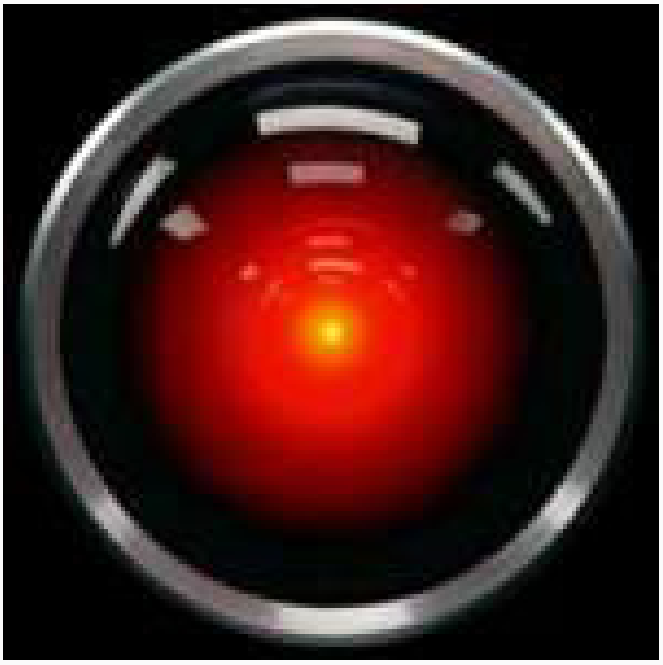
Språkteknologi

- Kjært barn:
  - Språkteknologi ("Language Technology")
  - Datalingvistikk ("Computational Linguistics")
  - Menneskespråkprosessering ("Natural Language Processing")

Hva er språkteknologi?

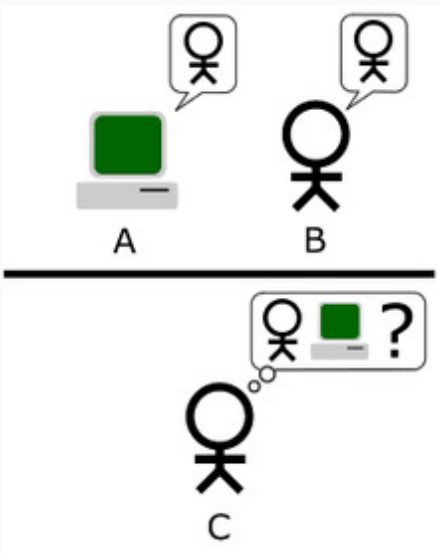
- "Lære datamaskiner å forstå menneskelige språk"
- Teknikker for å automatisk behandle språklige data
- Tverrfaglighet:
  - informatikk
  - lingvistikk
  - logikk, statistikk, matematikk, filosofi, ...

The holy grail

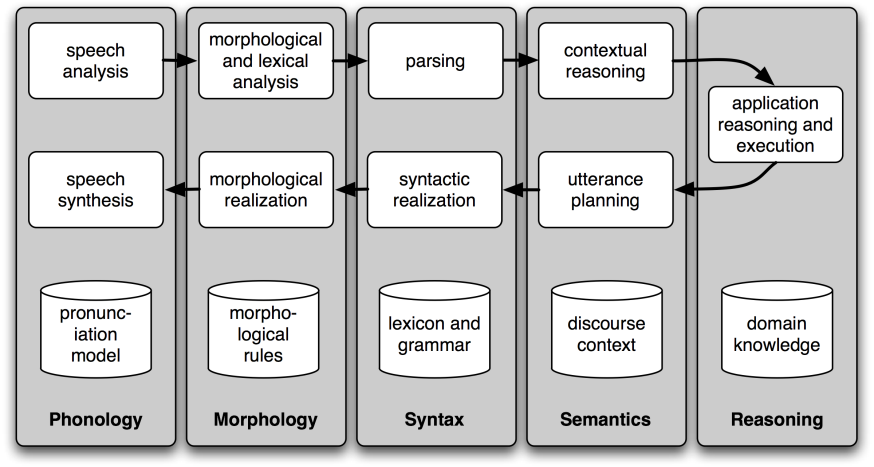


Turingtesten

- AI: test på en maskins evne til å vise intelligens
- Alan Turing: *Computing Machinery and Intelligence* (1950)
- "Kan en maskin tenke"
- "Finnes det en tenkelig datamaskin som kan klare Turingtesten?"



Et dialogsystem



Språkteknologiske komponenter

- *Fonetikk/fonologi:* kunnskap om lingvistiske lyder
- **Talegjenkjenning/talesyntese:**
  - tale ⇒ tekst
  - tekst ⇒ tale

Eksempel problem

Homofoner (homonymer) – ord som uttales likt men har forskjellig betydelse

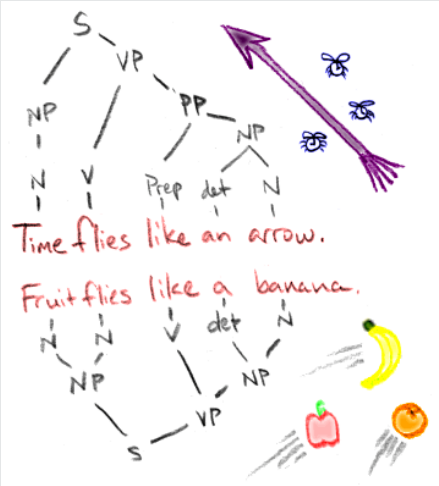
- weak — week
- to — too — two

`http://www.economist.com/technology-quarterly/2017-05-01/language` `https://deepmind.com/blog/wavenet-generative-model-raw-audio/`

Språkteknologiske komponenter

- *Morfologi:* kunnskap om ordstruktur
  Morfologisk analyse, ordklassetagging

Eksempel problem



Flertydighet av *flies* og *like* gir opphav til ulike tolkninger

Språkteknologiske komponenter

- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

Eksempel problem

Noen syntaktiske konstruksjoner gir opphav til flere tolkninger

liten, pen skole
liten pike, pen skole
liten, pen pike
ganske liten skole
ganske liten pike

Språkteknologiske komponenter

- *Semantikk:* kunnskap om mening — ord, setninger
  "Word Sense Disambiguation" (WSD)

Eksempel problem

En form – flere meninger

- Mine **mål** er egentlig ganske forskjellige
  - uttalt av en fotballspiller
  - uttalt av en modell som sammenligner seg med Kate Moss
  - uttalt av en ISK masterstudent

Språkteknologiske komponenter

- *Semantikk:* kunnskap om mening — ord, setninger

Eksempel problem

Flertydighet i rekkevidde ('scope')

- Alle studenter hater et kurs:
  $\forall x.student(x) \rightarrow (\exists y.kurs(y) \wedge hater(x, y))$
  $\exists y.kurs(y) \wedge (\forall x.student(x) \rightarrow hater(x, y))$

Språkteknologiske komponenter

- *Diskurs:* kunnskap om enheter ut over enkelte ytringer
  Anaforresolusjon, dialogsystemer
  - . . . det er diskutabelt hvor mye Watson egentlig "forstår". **Den** driver snarere en form for etterlikning av noen av måtene menneskehjernen prosesserer språk på. (Dagbladet, 14/1/11)

Språkteknologiske komponenter

- NLP-systemer: moduler som representerer forskjellige lingvistiske nivåer
- "Pipeline"-arkitektur
- "Høyere" nivåer avhenger typisk av "lavere"

- Hvorfor blir resultatene (noen ganger) dårlige?
  - Språkforståelse er komplisert
  - Den nødvendige kunnskapen er enorm
  - De fleste stadier viser flertydighet

Flertydighet

- De fleste språkteknologiske applikasjoner må håndtere *flertydighet* ("ambiguity")
- Kjennetegner naturlige språk, på alle nivåer
  - I saw her duck
  - Krasjet med rådyr på moped (Agderposten)

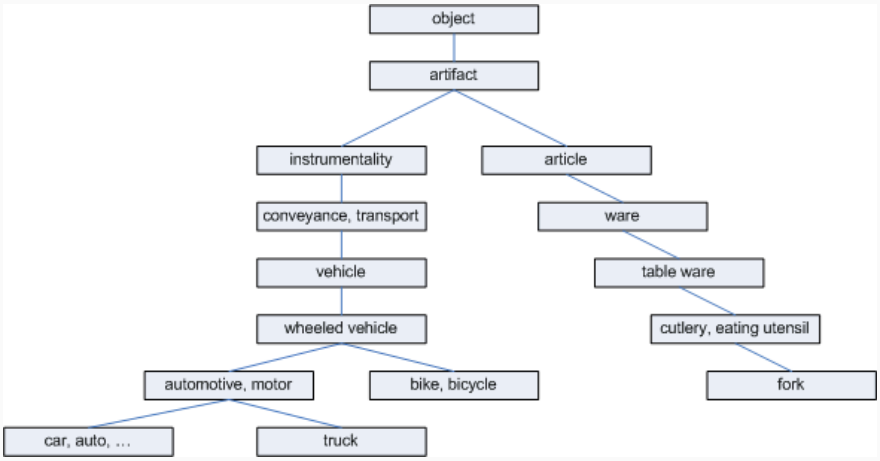
The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted — recognizing a face, lifting a pencil, walking across a room, answering a question — in fact solve some of the hardest engineering problems ever conceived. . . As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Steven Pinker, *The language instinct*

- En robot som bretter et håndkle (videoen er 50 ganger normal hastighet):
  `http://www.youtube.com/watch?v=gy5g33S0Gzo`

Metoder

→ 2000-tallet: manuelt utformede regeler og leksikon



Metoder

- 2000-tallet $\rightarrow$: empirisk revolusjon
- **Maskinlæring**
  - Datamaskiner kan lære fra data: fange opp mønstre og generalisere til nye eksempler

Metoder

- Formelle modeller hentet fra matematikk, logikk, statistikk
- Maskinlæring brukes for å håndtere flertydighet

Maskinlæring

Kunstig intelligens: delområde innen informatikk (fra 60-tallet)

- fokus på oppgaver som er lette for mennesker, men vanskelige for maskiner
- Språkforståelse er en slik oppgave
- Andre oppgaver: planlegging, bevegelse i verden, objektgjenkjenning, osv.

Maskinlæring:

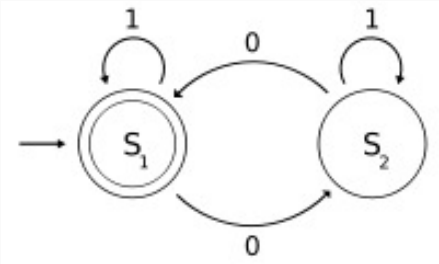
- gitt et AI problem og en masse data om verden (eksempler): la datamaskinen finne riktig svar
- unngår hardkoding av svarene

Formelle modeller

- Kan fange inn den språklige kunnskapen v.hj.a et lite antall formelle modeller og teorier
- Hentet fra informatikk, matematikk og lingvistikk
- Disse modellene kan prosesseres ved et lite antall algoritmer — søk (feks dynamisk programmering), maskinlæring (feks klassifikasjon)

Formelle modeller

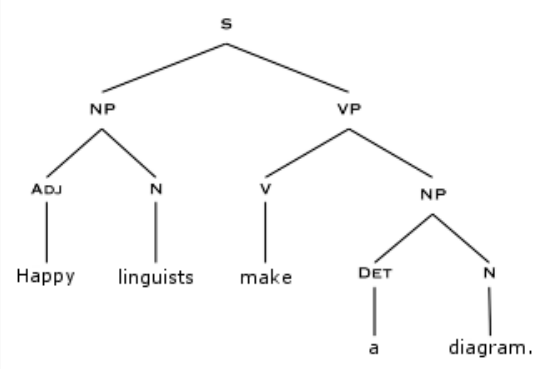
- Endelige tilstandsmaskiner ("finite state automata"):
- Består av tilstander, overganger ("transitions") og en input-representasjon
- Variasjoner: deterministiske og ikke-deterministiske, endelige tilstandsmaskiner og endelige tilstandstransdusere



Fonologi, morfologi

Formelle modeller

- Formelle regelsystemer
- feks kontekstfrie grammatikker

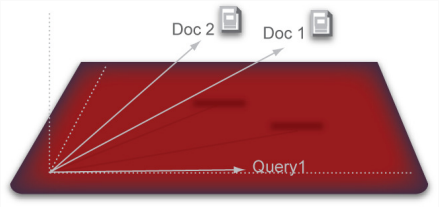


- S $\rightarrow$ NP VP
- NP $\rightarrow$ ADJ N
- NP $\rightarrow$ Det N
- VP $\rightarrow$ V NP

Syntaks

Formelle modeller

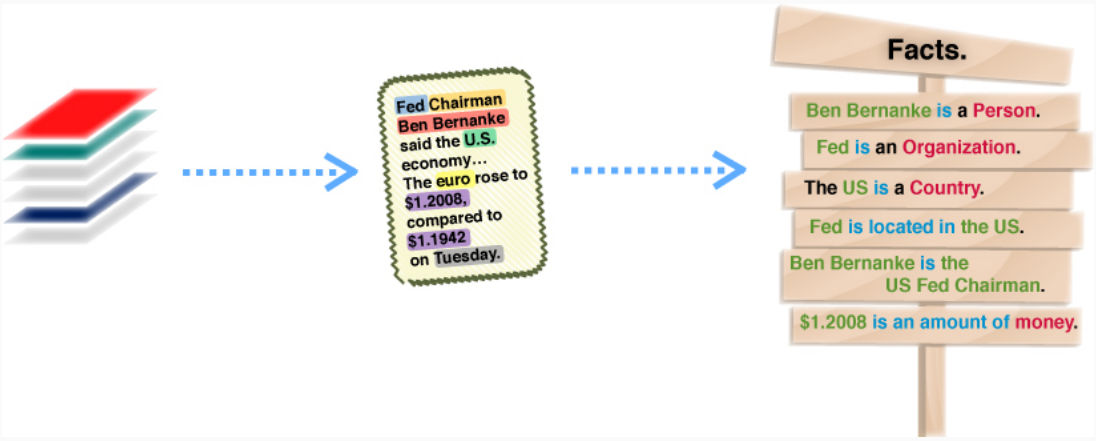
- Første ordens logikk **Semantikk, pragmatikk**
- Probabilistiske modeller – utvidelser til probabilistiske versjoner, disambiguering
- Vektormodeller



Leksikal semantikk, søk

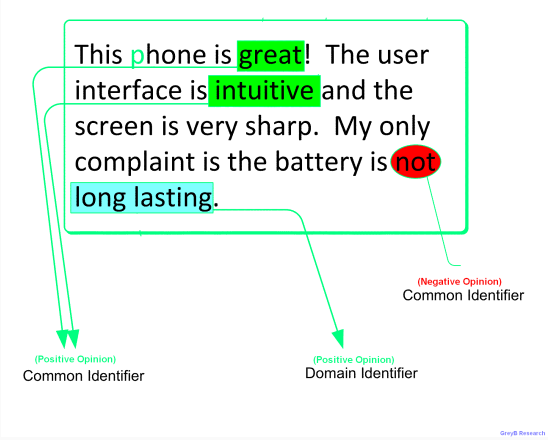
Hva kan vi bruke språkteknologi til?

Informasjonsekstraksjon

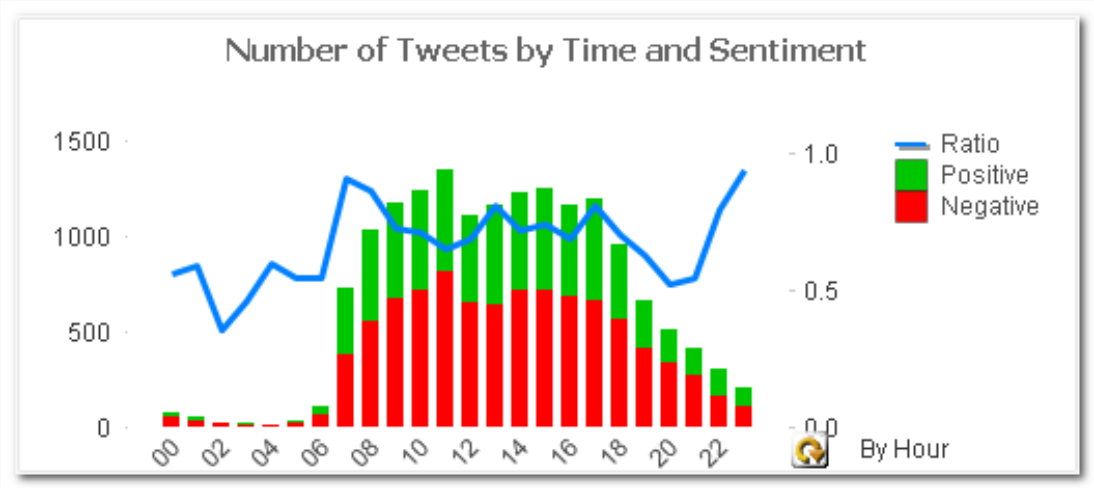


Sentiment Analyse

Automatisk analyse av subjektivt språk



Medieovervåkning

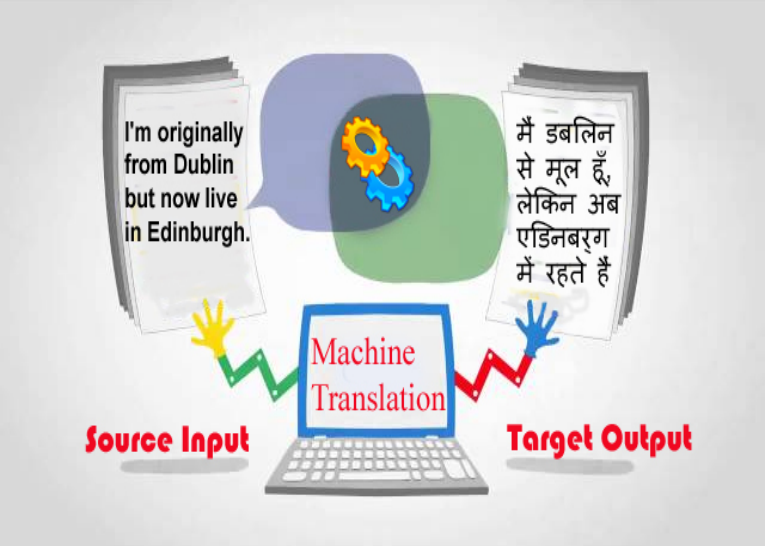


Demo

- System for informasjonsekstraksjon og sentiment analyse
- `http://www.alchemyapi.com/`

Maskinoversettelse

- Automatisk oversette fra et menneskelig språk til et annet



Dialogsystemer

- IT-systemer som kan kommunisere med brukere ved hjelp av sprk

Python-programmering og NLTK

Hva er et program?

- På et abstrakt nivå er et program en sekvens med kommandoer, som gir output for en viss input.
- **Eksempel 1:**
  1. Input: en tekstfil med hele *Kristin Lavransdatter*
  2. Program: noe skjer (Input $\mapsto$ Output)
  3. Output: alle bigram (sekvenser av to ord) med frekvenser
- **Eksempel 2:**
  1. Input: Lønn-og trekkoppgave
  2. Program: noe skjer (Input $\mapsto$ Output)
  3. Output: Hvor mye skatt du må betale

Algoritmer

- Et program er en algoritme, dvs en sekvens av kommandoer
- Eksempel på algoritme som skriver ut alle ord som ender med -er:
  1. Les inn alle ord fra teksten
     - 1.1 Sjekk om hvert ord ender i -er
     - 1.2 Lagre (unike forekomster) av ord som ender i -er
  2. Skriv ut hvert ord
- Men hvordan leser vi inn noe eller lagrer noe?

Programmeringsspråk

- Programmeringsspråk har mye til felles
  - Lignende datastrukturer (lister, funksjoner, moduler, ...)
  - Krever at du bruker en eksplisitt syntaks
    - Kun veldefinerte funksjoner kan brukes
    - Må følge visst format
- Ofte forskjeller i syntaks, men god programmeringsskikk overføres ofte, samme prinsipper

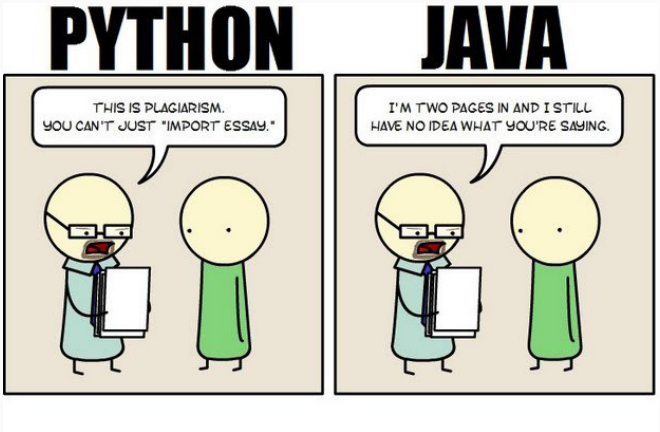
Python

```
for line in open("file.txt"):
    for word in line.split():
        if word.endswith('er'):
            print word
```

1. Whitespace
2. Objektorientert: hver variabel er en entitet som har visse definerte attributter og metoder
   Feks `line` – strengobjekt med bl.a. metoden `split`
3. Metoder har argumenter (i parentes)

Hvorfor Python?

- Enkelt, kraftig
- Meget god funksjonalitet for prosessering av lingvistiske data (strengbehandling, tekstprosessering)
- Lesbart
- Objektorientert: lett å kapsle inn kode og bruke om igjen
- Brukes mye! Organisasjoner som Google, Pixar og NSA bruker Python



NLTK

- Natural Language Toolkit (NLTK):

  Open source Python modules, linguistic data and documentation for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux.

  (http://www.nltk.org)
- Open Source
- Kommmer med klasser for datarepresentasjon, grensesnitt for oppgaver som ordklassetagging, parsing, tekstklassifisering
- Veldokumentert