Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Språkteknologi: siste trender og fallgruver

Pierre Lison
plison@nr.no

Nasjonaltbiblioteket: Digitalt seminar, 3. mars 2020

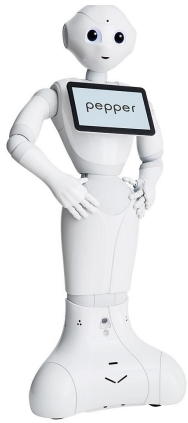
Språkteknologi @ NR

- Forskningsinstitutt (80+ ansatte)
- Utfører oppdragsforskning innen statistisk modellering, maskinlæring, språkteknologi og IKT

CLEANUP
(https://cleanup.nr.no)
automatisert anonymisering av tekstdata

SAFERS
Tale- og språkteknologi for nødmeldetjenester

GraphDial
(https://graphdial.nr.no)
dialogmodellering for komplekse, åpne interaksjoner



DialMT
Dialogmodellering for maskinoversettelse

Oslo Analytics
Tekstmining for trusseletteretning

Plan

- Siste trender innen språkteknologi:
  - Nevrale språkmodeller
  - Etikk og forklarbarhet
  - Små datamengder
- 5 vanlige fallgruver

Plan

- **Siste trender innen språkteknologi:**
  - **Nevrale språkmodeller**
  - **Etikk og forklarbarhet**
  - **Små datamengder**
- 5 vanlige fallgruver

Språkteknologi

Fra forskningsprototyper (modeller, algoritmer, evalueringsmetoder) til konkrete programvare

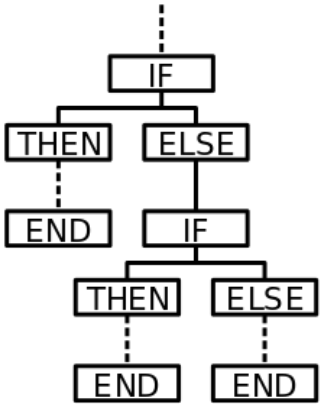
= paraplybegrep for teknologiske løsninger som behandler **språkdata** in en eller annen form

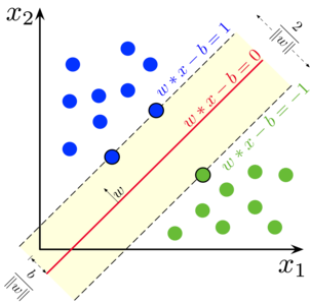
Tale eller tekst: dokumenter, nettsider, samtaler, lydopptak, osv.

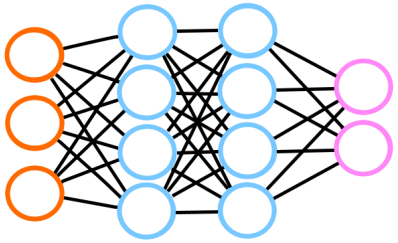
- Transkribere mellom tale og tekst (*talegjenkjenning*, *talesyntese*)
- Utvinne informasjon fra store mengder tekstdata (*søk*, *tekstmining*)
- Oversette mellom språk (*maskinoversettelse*)
- Føre en samtale i naturlig språk (*dialogsystemer*, *pratereroboter*)
- Støtte skriving eller lesing
- ...

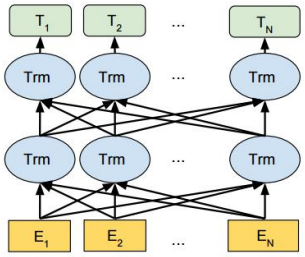
Historisk utvikling

- 1970-nå: regelbaserte systemer
- 1990-2012: datadrevne modeller ("klassisk" maskinlæring, statistikk) blir stadig mer populære
- 2012-nå: Nevrale modeller (ofte "dype", altså med mange prosesseringslag) tar over feltet
- 2018-nå: Utvikling av stadig større nevrale språkmodeller





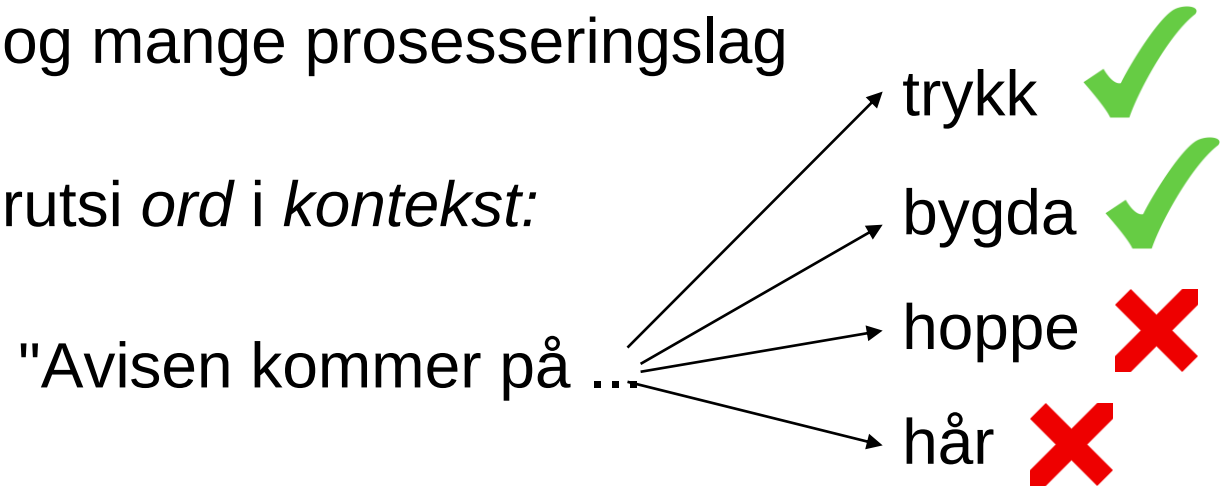




Nevrale språkmodeller

(BERT, GPT-3 osv.)

- Nevrale arkitekturer med opptil flere milliarder parametere, og mange prosesseringslag
- Trent på å forutsi *ord* i *kontekst:*



- "Gjettelek" gjentatt milliarder ganger, på store mengder tekster tilgjengelige på nettet (Wikipedia osv.)
  - Lærer underveis både grammatikk og faktakunnskap

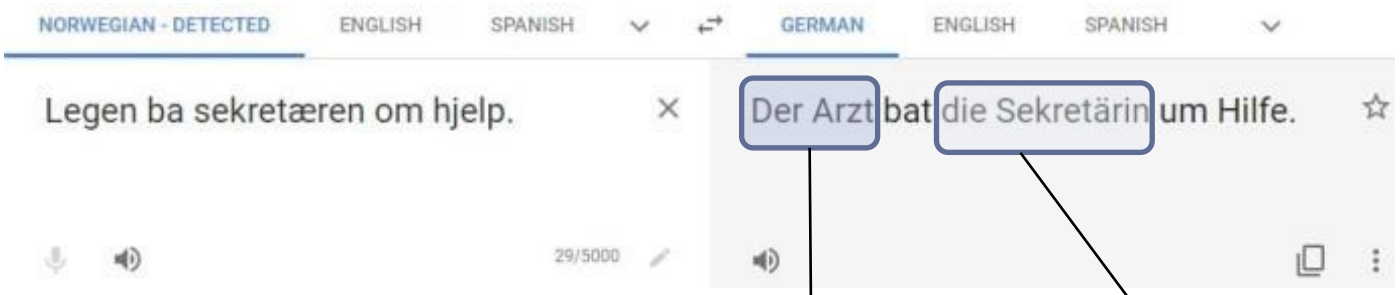
Nevrale språkmodeller

- Etter trening kan modellen tilpasses ("fine-tune") til ulike språkteknologiske oppgaver
  - Basert på (markerte) data spesifikke til oppgaven

- Overraskende gode til å generere troverdige tekster
  - Men ikke fall for "hypen": det betyr ikke at modellen har "forstått" hva som ble generert!
- Språkmodeller for norsk (nb og ny): NorBERT, NoTraM

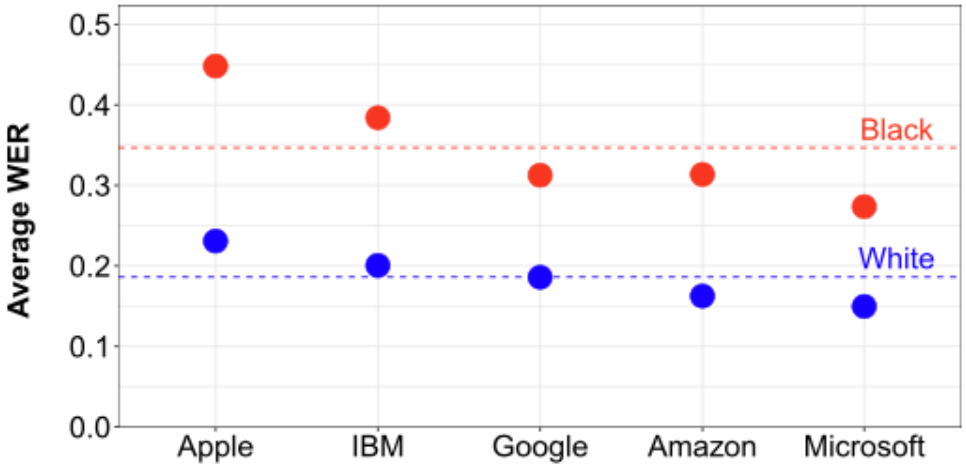
Etikk

Datadrevne modeller kan reprodusere stereotyper og fordommer som uttrykkes i treningsdata



Etikk

- Mange demografiske grupper (språklige og etniske minoriteter, lavt utdannede, osv.) er *underrepresentert* i språkdata vi samler
- Det kan føre til *reelle konsekvenser* i bruken av teknologi for disse gruppene:



[A. Koenecke et al (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences]*

Forklarbarhet

- Prediksjoner fra nevrale modeller er vanskelig å forklare
  - "Sorte bokser" som produserer tilsynelatende gode resultater
  - ... Men man forstår ikke hvorfor!
- Dette er selvsagt svært problematisk
- Mye nye forskning om utvikling av mer "forklarbare" modeller innen maskinlæring og språkteknologi
  - Hva har modellen egentlig lært?
  - Hvilke språklige mønstre "fanges opp" av modellen?

Små datamengder

- Moderne språkteknologi sterkt avhengig av (gode) treningsdata
  - Men i praksis har man ofte ikke tilgang til store mengder (markerte) data!
- Hva kan man gjøre? Ulike strategier:
  1. **Overføringslæring**: "gjenbruke" kunnskap fra beslektede tekstdomener / oppgaver / språk
  2. **Dataøkning**: lage nye treningseksempler ved å lage modifiserte versjoner av eksisterende eksempler
  3. "**Weak supervision**": bruke ekspertkunnskap (heuristikker, databaser, osv.) til å automatisk markere data

Plan

- Siste trender innen språkteknologi:
  - Nevrale språkmodeller
  - Etikk og forklarbarhet
  - Små datamengder
- **5 vanlige fallgruver**

Noen vanlige fallgruver

1. For høye forventninger

- Språkteknologi har riktignok blitt mye bedre de siste årene
- Men modellens "forståelse" er fortsatt veldig overfladisk (gjenkjenning av mønstre i tekst)
- *Ikke fall for hypen*: hvis en leverandør forsøker å selge dere et "revolusjonerende" system som hevdes å "forstå" språk like bra som mennesker, løp unna

Noen vanlige fallgruver

2. Oppgaven er for vagt definert

- Utvikling av språkteknologiske modeller krever å definere på en veldig presise måte:
  - Hva slags "inputs" modellen skal motta
  - Hva slags "outputs" modellen skal produsere
  - Hva som regnes som en "god" eller "dårlig" output
- Ikke undervurder viktigheten av preprosessering (formattering, normalisering, osv.)!

Noen vanlige fallgruver

3. Dårlig eller utilstrekkelig data

- Data er helt avgjørende for utviklingen av gode språkteknologiske løsninger
  - (også for regelbaserte systemer!)
- Datainnsamling og markering er tid- og ressurskrevende
- Datakvalitet er også en viktig moment
  - Samt hvorvidt treningsdata gjenspeiler språkdata som hentes inn når systemet er i bruk

Noen vanlige fallgruver

4. Starte for tidlig med for kompliserte modeller

- Virksomheter ønsker ofte å ta i bruk de seneste, mest avanserte språkteknologiske modeller
- ... Ofte en dårlig idé: vanskelig å få til å fungere (og forstå feilene som produseres), fare for overtrening, osv.
  - Enklere løsninger basert på bl.a. regler eller statistiske modeller kan fungere like bra!
- Hvis begrensede ressurser: investeringer i datainnsamling er ofte mye mer lønnsomme!

Noen vanlige fallgruver

5. Ikke nok fokus på evaluering

- Mange virksomheter kjøper "ferdige" språkteknologiske løsninger fra en leverandør
  - Ofte kun anekdotiske evalueringer av systemet
- Det er synd - systematiske evalueringer lønner seg!
  - Både kvantitative og kvalitative (feilanalyse)
  - Må gjentas over tid: nye temaer / forespørsler som dukker opp, språkbruk i stadig endring, osv.

Spørsmål / kommentarer?