»Computational linguistics«

Introduktion til datalingvistik

Kursus på RUC-Datalogi/Tværfaglig IT for ITU

Henning Christiansen

ph.d., lektor i datalogi på RUC

henning@ruc.dk, http://www.ruc.dk/~henning

Kursusweb: http://www.ruc.dk/~henning/CompLing2004

Idag:

- Introduktion,
- præsentation
- Lære Prolog og enkle Definite Clause Grammars at kende:
- udvikle vores første lille dialogsystem!

Målet med kurset er ...

- at introducere til at man kan lave spændende, computerbaserede modeller af sprog
- oversigt over området
- at man får kendskab til sproglige fænomener og modellering ved logik- og constraint-baseret tilgang
- kendskab til værktøj og metoder til prototyper/kørende modeller
- Mulige anvendelser: NLP grænseflader, tale, ..., tekstanalyse (f.eks. ontologisk)
- Metoder–begreber–erfaringer kan udnyttes på andre felter, HCI, signalbehandling, ...
- Motivere til projekter/specialer

Målet er ikke ... at uddanne fikse og færdige udviklere til sprogteknologiske systemer

Idé i kurset:

- Vi kan ikke nå alt på 10 gange
- Supplere med grundbog

  Jurafsky, Martin, 2000:

  Speach and Language Processing,

  An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognotion

  Baggrundslæsning, evt. dykke ned i enkeltdele

  ... og at have bogen bagefter!!

Annoncerede deltagerforudsætninger:

- » ... **enten** en baggrund i lingvistik og mod på at omsætte beskrivelser til kørende programmer,
- **eller** et kendskab til indledende programmering, en interesse for sprog og et elementært kendskab til grammatik.«

Lærerens forudsætninger:

- Programmeringssprog, logikprogrammering (LP), constraint-logikprogrammering (CLP), ...
- Anvendelser af LP- og CLP-teknologier
- Udvikling af CHR Grammars, http://www.ruc.dk/~henning/chrg
- (Data)lingvistiske konferencer, Int'l Pragmatics Conf. 1993, RANLP 2003
- »Tværfaglige« specialer på RUC
- OntoQuery-projektet (RUC, DTU, CBS-datalingvistik, CST)
- Forskningssamarbejde med
  Veronica Dahl, Simon-Fraser U. Canada,
  Philippe Blache, U. Aix-en-Provence Frankrig
- CONTROL-projektet: HC, VD, PB, Jørgen Villadsen, John Gallagher, Troels Andreasen

Mit første NLP-kursus! som jeg håber på at lære meget af!

Deltagerrunde: Baggrund, Forventninger??

Kursusform:

- Så vidt muligt udglatte skel forelæsning/øvelser, tænke i hele kursusdage.
- Kan (til vis grad) tilpasses efter deltagernes forudsætninger og ønsker
- Så meget som muligt eksperimentere med små kørende prototyper, hvor forskellige fænomener eller teknikker demonstreres
- Idéer fra jer?
- Nogen som ligger inde særlig sproglig viden?

Der skal foreligge en »lidt større« opgave med præsentation

Analyse af et stykke litteratur? En lille implementation?

Eller kombination?

»**Eksamensform:**
Mundtlig eksamen.
Eksamen vil tage udgangspunkt i dels et spørgsmål som trækkes ved eksamen og dels en samtale om en opgave de studerende forventes at have løst på kurset og præsenteret for de øvrige deltagere.«

Mulige emner på kurset?

- Intro til computerbaseret sprogbehandling, problemstillinger & anvendelser.
- Prolog & Definite Clause Grammars
- Grammatikker og analyse af delmængder af dansk e.lign.
- Constraint Handling Rules & CHR Gram's (kontekst-sens. regler, hypoteser, ...), Assump.Grammars
- Modellering af fænomener så som håndtering af tvetydighed, elipsis og koordination, anaphora (henførende stedord m.v.), ...
- Abduktion som begreb, abduktion til kontektsanalyse
- Evt. fejlrettende grammatikker
- Evt. sætningsskemaer à la Diderichsen/Funktional grammatik à la Heltoft+Hansen overført til CHRG eller DCG
- Evt. traditionelle datalogiske metoder: endelige tilstandsmaskiner, parsing

  (også relevant for prog.sprog., signalbehandling, biologi...)
- ?? Feature structure? HPSG ...
- Jeres forslag?

Nu introduktion til emnet: Datalingvistik med den meget store pensel

- Computere og sprogkompetence?
- Illusioner
- Anvendelser i dag
- Elementer af maskinel sprogbehandling
- Problemer med tvetydigheder på forskellige niveauer
- Metoder og ressourcer

Derefter workshop

- Prolog for minimalister
- Definite Clause Grammars
- udvikle dialogsystem som interagerer med vidensbase

Computere og sprogkompetance

Læs selv J&M kap. 1 ... for detaljer og referencer

Tidligere visioner 1960 ±

»Om få årtier vil computerne kunne forstå og bruge talt sprog lige så godt som mennesker«

»Perfekt maskinoversættelse, computer som man samtaler vil blive dagens orden ... når vi når årtusindskiftet«

(Tilsvarende forståelse af »kunstig intelligens« ...)

Erfaring: Fiaskoer — nu bedre forståelse + mere ydmyge mål

Eksempel på tidlig »succes«: Eliza, Weizenbaum 1966 Psykiaterprogram

Eksempel på vision:

Stanley Kubrick, 1968, 2001 - A Space Odyssey HAL 9000

Hvilke kompetencer forventes, minimum for HAL?

- Talegenkendelse,
- Forståelse af naturligt sprog (sætningsstruktur...)
- Generering af naturligt sprog og tale
- Finde og extrahere information og viden
- Bruge viden og information: Ræsonering/logisk inferens
- Model for »god og balanceret dialog«
- Mundaflæsning...
- en slags følelsesliv... /emotionel model

Anvendelser i dag

- Brugergrænseflade
- Turistinformation »Hvon kom' a te æ tiwoli?«
- Billetbestillingssystemer pr. telefon
- Oplysningssystemer via telefon
- Kommandér rundt med din mobiltelefon eller rullestol
- Alm. teksbehandling: Stavekontrol, grammatikkontrol, stilkontrol, automatisk »summary«
- Maskinoversættelse (rå oversættelse, suppleret af prof. oversætter)
- information extraction, information retrieval
- ....

Elementer af maskinel sprogbehandling: Typiske faser af sproganalyse

Fonetisk analyse: Talestrøm (bølgeformer) -> fonemer

Fonemer -> ord og »punktuationer«

Leksikalsk analyse: (Obs: betegnelse svinger fra fag til fag)

Tegnsekvenser -> ord og punktuationer

Morfologisk analyse:

Ord -> Grundform + Ordklasse + form/bøjning

f.eks. Hundene -> hund1/SB/PLU/DET

Syntaktisk analyse

Sekvenser af »annoterede ord« -> sætningsstrukturer
(træer eller udfyldning af skemapladser)

... »deep« eller »shallow« efter anvendelse

Semantisk analyse

sætningsstrukturer -> »kontekstuafhængige betydninger« (???)

Pragmatisk analyse

Hvordan sætninger bruges i forskellige situationer, hvordan det påvirker tolkning (?)

Diskursanalyse

Henførende stedord: »den« »han« »førbemeldte«,
Temporale aspekter: »ikke nu men NU«

Bearbejdning af »indhold«

På »ontologisk« niveau: »Det handler vist om madlavning«

Dybere forståelse »Hmm, opskrift på Sauce Bernaise«
»... anderledes and den opskrift jeg plejer at bruge...«

Meget-meget mere!

Opdeling i niveauer, hvilke niveauer medtages, hvilken »granularitet«

afhænger af anvendelsen!

Eksempler på interessante problemer: Tvetydighed

Leksikalsk

»bank«

»Der er alarm i sektor B. A. Jensen tog ikke notits af det.«

»Betragt f.eks. Andersens skriverier, eller Kierkegaards, f.eks. Rifbjerg, derimod ...«

Morfologisk

»kvindelig«

Sætningsniveau

»Jeg så manden med kikkerten«

»Jeg så [manden] [med kikkerten]«

»Jeg så [manden med kikkerten]«

<Jens, Peter, Jens Peter og Jens Ejnar spillede kort i går>

»Snød Jens Peter?«

Stedord

»Politiet forbød de studerende at demonstrere fordi de frygtede vold.«

»Politiet forbød de studerende at demonstrere fordi de truede med vold.«

Konklusion: Korrekt analyse kræver indsigt på et meget højt niveau!

Eksempel fra Thomas Nykrog, Politikens fotobog (1. udgave, 3. oplag, 2003)

»De nyeste digitale modeller er de smarteste, men når man først har vænnet sig til dem, ...«

»De nyeste digitale modeller er de smarteste, men når man først har vænnet sig til dem,
er de gamle med drejeskive faktisk hurtigere at arbejde med«

Hvordan fandt I ud af hvad »dem« refererede til?

Danske verber er noget underligt noget

trække X fra Y

X = »momsen«, Y = »regningen«
X = »den moms vi skulle have betalt hvis det skulle gå rigtigt for sig«, Y=underforstået

trække X for Y

X = »gardinet«, Y = »vinduet« eller underforstået, eller intet Y?

trække X til

træk døren til, ellers trækker det fælt

Konklusioner (?)

Det ideelle system må have en indsigt som nærmer sig et menneskes. (?)

Alle faser skal evealueres i parallel (overlejret; som corutiner)

Man kan nå langt med tommelfingerregler og listning af specialtilfælde (?)

Simpel semantisk kategorisering hjælper et langt stykke hen ad vejen:

> »Peter spiste kage og Marie og Børge kringle«

Vel er der et eller andet sted en tvetydighed (i meget generelt system), men [min påstand], hvis sprogbrugeren ville indikere den ekstraordinære tolkning, at Peter spiste kagen og Marie, ville han have formuleret sig så tvetydigheden forsvandt.

Nok ikke sådan her, men det viser idéen:

> »Peter spiste Marie og kage og Børge kringle«

((Marie ≈ privat sprog for mariekiks?))

Metaforer og stående udtryk:

»Nå, så der ligger hunden begravet!«

handler ikke (nødvendigvis) om hunde.

Generelt om håndtering af tvetydighed på forskellige niveauer:

- lægge sig fast på én tolkning ud fra aktuelt niveau (tommelfingerregler)
  misser en sætning i ny og næ
- generere alle tolkninger og sortere fra efterhånden som de fejler
  kan eksplodere i kompleksitet
- vælge mest sandsynlig tolkning, backtracke til næst-mest sandsynlige hvis...
- afvente valg af tolkning, indtil vi måske har information til det
- ???

Metoder og resourcer

Fonetisk analyse

Se J&M kapitler 4–7
Vigtigste viden her: Der findes softwarekomponenter til formålet

Leksikalsk-morfologisk (ikke krystalklar på skelnen)

Endelige tilstandsmaskiner

Ordbøger »lexicon«

Standardalgoritmer til at klippe endelser af bagfra, engelsk »stemming«

Sammensatte ord på dansk: Standardalgoritmer som leder efter binde-e og -s
kombineret med ordbogsopslag

Udfoldet ordbog:
»hundene« er et opslagsord

Men ukendte ord vil der altid være!

Leksikalsk-morfologisk (fortsat)

»Tagging« J&M kap 8

algoritmer til bestemmelse ordklasse, regler kombineret med ordbogsopslag:

Peter/PN og/CON Børge/PN qwakser/?? øl/SB/0

Ordet »qwakser« — ender på »-er« og står mellem PN og SB så der er nok »qwakser/VB«

Typisk baseret på maskinindlæring (Brill):

kompetente lingvister håndtagger korpus

standard AI-algoritmer udtrækker regler

Træksikkerhed: 90–98%

Syntaktisk analyse

Parsealgoritmer, top-down, bottom-up kendt fra datalogisk litteratur siden 1950'erne
I logikprogrammering repræsenteret ved DCG og CHRG som vi vender tilbage til

Egenskaber: Robusthed, effektivitet, ...

Semantisk analyse

?kompositionel semantik, Montague o.lign., semantiske kategorier i ordbog

Pragmatisk/diskurs

Tommelfingerregler »pronomener refererer bagud til nærmeste relevante entitet«
undtagen .... og ..... og .....

Abduktion for kontekst og implikatur
»dommeren ... dømte ... forsvarer ... sparkede ... linjevogter .... hjørnespark ... «

»nu er er han fuld igen« »Mary er køn, men intelligent«

Generelt: Blandede AI-bolcher

NB: Sprog er forskellige og må behandles forskelligt

Kompliceret bøjningssystem eller næsten ikke eksisterende?

Græsk og slaviske sprog.
Latin-fransk-italiensk-...
Dansk-tysk-...
Engelsk.

Eksempel på græsk-dansk-engelsk

Ton protognorisa ston Pireia. (4 ord)

Jeg mødte ham første gang i Pireus. (7 ord)

I met him for the first time in Pireus. (9 ord)

»Free-order versus fixed order«?

Latin siges at være tæt på »free-order«.

Selv på dansk noget der minder om det:

»Peter spiste kagen. Lakridskonfekten spiste Dorte.«

Sammensatte termer

Datalingsvistikuddannelsen

The Computational Linguistics Education

Engelsk har massevis af lokal tvetydighed også fordi de fleste ord opræder i flere ordklasser kombineret med at sammensatte termer skrives med mellemrum!!

Dansk sprogbehandling har problemer med at knække sammensatte ord!

Vær kritisk over litteraturen, specielt den som *implicit* tager engelsk som udgangspunkt.

Afgrænsning i kurset

elementært-leksikalske og syntaktiske niveauer

med afstikkere til det pragmatisk og lidt diskursanalyse

kobling til »semantiske« vidensbaser

mere eksemplariske end realistiske anvendelser

Prolog: Et usædvanligt programmeringssprog

Høj udtrykskraft + interaktivt $\longrightarrow$

nemt at lære (til husbehov!)

\+ egnet til eksperimenter: Idé -> model -> test -> revidere/udvide idé -> ...

Symbolske strukturer $\longrightarrow$ Sprogbehandling

Kerne af matematisk logik $\longrightarrow$ enkelhed i begreber + generalitet + 100årigt teoriapparat

Baggrund

Automatisk bevisførelse: Resolution, bevisregel m. unification; Robinson, **1965**

A.Colmerauer & co. (Marseille), **ca. 1970**: Programmeringssproget »Prolog« ≈ »Programmation en Logic«
udviklet sammen m. Grammaires de Métamorphose

D.H.D. Warren: Effektiv compiler abstrakt maskine "WAM", **1975**, ...

Udbredelse af sproget R.Kowalski »Logic for Problem solving«, **1979**, ...

GdM -> Definite Clause Grammars indbygget i stort set alle Prolog-systemer (ISO?)

Vi bruger SICStus Prolog fra Swedish Institute of Computer Science
som vort værktøj på kurset