INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Trettende forelesning – REPETISJON

Lilja Øvrelid

15 mai, 2017

Fra emnebeskrivelsen

Kurset gir en innføring i lingvistisk teori og relaterer denne til språkteknologiske problemområder, metoder og applikasjoner. Fokus er på å koble teori til praksis. Vi vil ta for oss morfologisk, syntaktisk, samt noe semantisk analyse av naturlige språk, formell språkteori og korpusbaserte metoder. Studentene vil få et første møte med noen datalingvistiske applikasjonsområder.

Lingvistikk

Lingvistikk

- Vitenskapelige studiet av språk
- Vitenskapelig? Systematisk studie av regler, systemer og prinsipper i menneskelige språk
- Kunnskapen om enheter og regler i et språk:
  - Fonologi: lyder $\Rightarrow$ ord
  - Morfologi: morfemer $\Rightarrow$ ord
  - Syntaks: ord $\Rightarrow$ fraser, fraser $\Rightarrow$ setninger
  - Semantikk: ord $\Rightarrow$ mening, setninger $\Rightarrow$ mening

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)
  - *I will meet you by the bank*

Språklige data

- Menneskelig språkprossessering: hvordan modelleres språk i hjernen?
  - afasistudier, hjernescanning
- Språkteknologi: programmer som generaliserer over språklige mønstre
  - **korpusdata** helt sentralt (representativitet)

Morfologi

Handler om **ord**

- hvordan ord er bygd opp (morfemer)
- hvordan nye ord dannes (avledning, sammensetning)
- hvordan ord bøyes

Morfologi

- Morfemet er den elementære (minste) lingvistiske enheten
- To hovedtyper:
  - **Frie** morfemer: ord. *boy*, *desire*, *gentle*, *man*
  - **Bundne** morfemer: affikser.
    - prefikser: *un-*, *pre-*, *bi-*
    - suffikser: *-ing*, *-ish*, *-ness*
- Morfologisk komplekse ord består av :
  - **Rot** + en eller flere affikser (*hus+lig*)
  - En rot er et ordelement som ikke kan deles opp i mindre (meningsbærende) deler

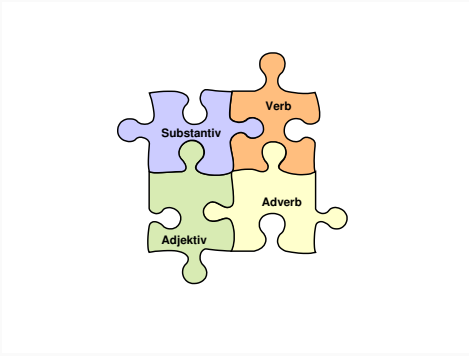
Avledning

- En avledning er et ord som er dannet fra et annet ord ved hjelp av et avledningsaffiks (prefiks eller suffiks),
- Avledningsbasen kan være et rotord (*barn*) eller en avledning (*barnslig*)
- Avledningsaffiksene er bundne morfemer med klart semantisk innhold (som innholdsord, men er ikke ord)
  - *u–* negasjon: *umulig, uvel, urolig*
  - *for-* - foran: *forelese, forbokstav, formann*
  - *-er* - den som utfører handlingen: *fisker, baker*

Bøyning

- **Bøyningsmorfemer** markerer kategorier som tempus, numerus (tall), kasus, etc.
- Noen eksempler:
  - **Tall**: entall og flertall *bil-biler, car-cars*
  - **Bestemthet**: uttrykkes i hovedsak ved suffiks (*bilen, huset*) eller (jf. engelsk bestemt artikkel *the*)
  - **Tempus**: angir tidspunktet for handlingen eller tilstanden som setningen beskriver. Presens (nåtid) og preteritum (fortid) *liker-likte, likes-liked*

Ordklasser



- **Taksonomi** - uttømmende, gjensidig utelukkende, styrt av et prinsipp
- Alle ord havner i en klasse og ingen ord havner i mer enn én klasse
- Vi trenger **kriterier** for ordklasseinndeling

Ordklassekriterier

3 slags kriterier:

1. formelle eller morfologiske kriterier
    - hvilke bøyningsformer har ordet?
2. funksjonelle eller syntaktiske kriterier
    - hvordan kan ordet kombineres med andre ord?
3. betydningsmessige eller semantiske kriterier
    - hva er typiske betydninger hos ord i ordklassen?

Inndeling av ordklasser

- **åpne** vs. **lukkede** ordklasser
  - nye medlemmer?
- **innholdsord** vs. **funksjonsord**
  - semantisk innhold?

Syntaks

"Studiet av hvordan setninger bygges opp av ord og ordkombinasjoner"

- **Syntaktisk form** - konstituenter beskrives i form av ordklasser, fraser
- **Syntaktisk funksjon** - konstituenter beskrives i form av sin funksjon i setningen som helhet

Konstituenter

- Konstituenter – grupperinger av ord i en setning, fungerer som en enhet
  - The dog ate my homework
  - The dog ate my homework
- Lingvistiske tester:
  - "stå alene"-testen
  - "erstattes med pronomen"
  - "Flyttes som enhet"

Syntaktiske regler

- Beskriver hierarkisk gruppering av ord
  - Alle eldre menn og kvinner kan forlate skipet
  - [gamle menn] og [kvinner]
  - [gamle [menn og kvinner]]
- Strukturell flertydighet
  - flertydighet grunnet flere mulige strukturer for en setning
  - forklarer hvordan gruppering av ord relaterer til betydning

Funksjonell analyse

- Hvilken funksjon et språklig uttrykk har i frasen eller setningen den forekommer i (ikke hva slags frase)
- **primære setningsledd**: fraser

My old friend		has bought		a car		in Dallas
SUBJ		PRED		D.OBJ		RAL

Semantikk

- Studiet av betydning slik det uttrykkes gjennom språk
- Betydning til morfemer, ord, fraser og setninger
  - Leksikal semantikk
  - Setningssemantikk
  - (Pragmatikk: hvordan konteksten påvirker betydning)

Leksikal semantikk

- Leksikal semantikk (ordsemantikk):
  - representere betydningen til ord
  - vise hvordan betydningene er relatert
- Semantiske trekk, f.eks. FEMALE, HUMAN (substantiver), CAUSE, GO, BECOME (verb)
- Leksikale relasjoner:
  - homonymi og polysemi
  - synonymi
  - antonymi (forskjellige typer)
  - hyponymi

Komposisjonell (Formell) Semantikk

- Beregner **sannhetsverdien** for setninger basert på betydningen til mindre deler (ord, fraser)
- Semantisk analyse ved oversettelse fra engelsk (eller norsk eller . . .) til et universelt metaspråk (førsteordenslogikk)
- Semantiske regler
  - John synger = Sj
  - **[**Sj**]** = 1 hvis og bare hvis **[**j**]** $\in$ **[**S**]**

Semantiske roller

- Annet aspekt ved setningsbetydning: hvilke roller de forskjellige deltagerene inntar i handlingen beskrevet av verbet
- Eksempel:

Gina	*hevet*	*bilen*	*med*	*jekken*
AGENT		THEME		INSTRUMENT

- PropBank og FrameNet inneholder informasjon om semantiske roller (korpus og leksikon)

Formelle modeller

Formelle modeller

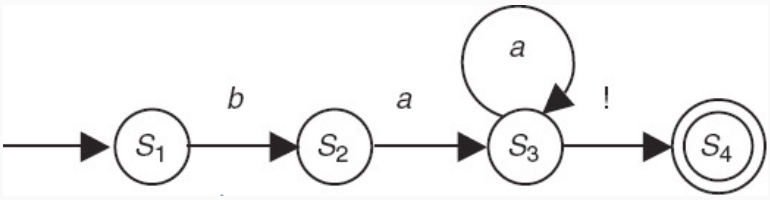
- Hentet fra matematikk, statistikk og (generell) informatikk
- Representere lingvistisk kunnskap
  - tilstandsmaskiner
  - formelle regelsystemer
  - logikk
  - probabilistiske modeller
  - vektorrommodeller (lineær algebra)

Regulære uttrykk

- Et regulært uttrykk er en beskrivelse av en mengde strenger
- Brukes til tekstsøk
- Regulære uttrykk består av:
  - strenger bestående av tegn: `b`, `INF1820`, `informatikk`
  - disjunksjon: `penge(r|ne)`, `[Dd]en`, `[A-Z]`, `[0-9]`
  - negasjon: `[^b]` `[^A-Z0-9]`
  - tellere: Kleenes `*`, `+`, `?`
  - "wildcard" for et hvilket som helst tegn: `.`
    `beg.n` `beltedyr.*beltedyr`

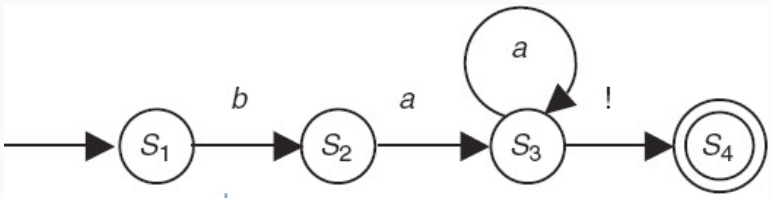
Endelige tilstandsmaskiner (FSM)

- Ethvert regulært uttrykk kan implementeres som en endelig tilstandsmaskin (og vice-versa)
- Måte å beskrive et spesielt formelt språk, nemlig regulære språk
- Hva er en endelig tilstandsmaskin (FSM)?
  - Abstrakt beregningsmaskin
  - Består av en mengde tilstander (noder i en graf), og en mengde transisjoner (kanter i en graf)
  - Tre typer tilstander: vanlig, start og slutt



Endelige tilstandsmaskiner (FSM)

- Kan også spesifiseres ved en transisjonstabell:



	a	b	!
S_1	θ	S_2	θ
S_2	S_3	θ	θ
S_3	S_3	θ	S_4
S_4	θ	θ	θ
θ	θ	θ	θ

Endelige tilstandsmaskiner (FSM)

- Ikke-deterministiske FSMer
  - En FSM er ikke-deterministisk dersom den
    - for hvertfall en tilstand og ett symbol fins det mer enn en transisjon som passer

	b	a	!	ϵ
S_0	S_1	Ø	Ø	Ø
S_1	Ø	S_2	Ø	Ø
S_2	Ø	S_2, S_3	Ø	Ø
S_3	Ø	Ø	S_4	Ø
S_4:	Ø	Ø	Ø	Ø

Kontekstfrie grammatikker (CFGer)

- Formelt: en CFG er en 4-tuppel $< N, \Sigma, P, S >$, der
  - N er en mengde **ikke-terminale** symboler (syntaktiske kategorier)
  - Σ er en mengde **terminale** symboler (ord)
  - R er en mengde **regler** (produksjoner) på formen $A \rightarrow \alpha$, der
    - A er en ikke-terminal
    - α er en streng av symboler hentet fra mengden $(\Sigma \cup N)*$, dvs både terminaler og ikke-terminaler
    - Et særskilt startsymbol S

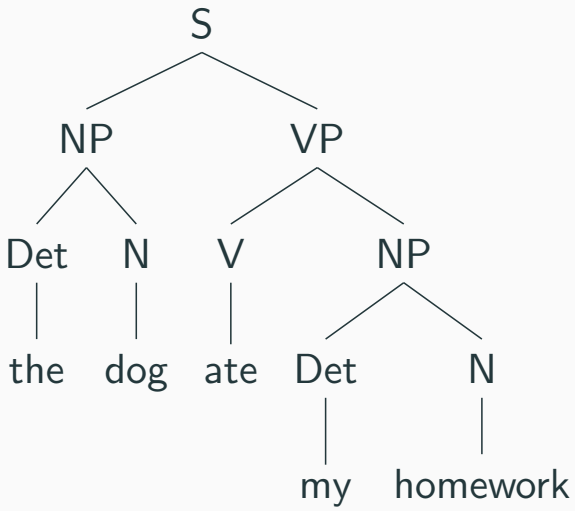
Kontekstfrie grammatikker (CFGer)

Eksempel CFG

- La $G =< N, \Sigma, R, S >$ der
  - $N = \{$S, NP, VP , DT, N', V, N$\}$
  - $\Sigma = \{$*et, fly, ankom*$\}$
  - $R = \{$S $\rightarrow$ NP VP,
    NP $\rightarrow$ Det N',
    N' $\rightarrow$ N,
    VP $\rightarrow$ V,
    Det $\rightarrow$ *et*,
    N $\rightarrow$ *fly*,
    V $\rightarrow$ *ankom*,
    $\}$
  - $S =$S

Trær

- En **derivasjon** av en streng fra en ikke-terminal A er resultatet av en rekke applikasjoner av reglene (fra G) til A
- Derivasjoner kan også visualiseres som **trær**



Rekursive stukturer

Direkte rekursjon:	
Nom $\rightarrow$ Nom PP	*flight to Boston*
VP $\rightarrow$ VP PP	*departed Miami at noon*
Indirekte rekursjon:	
S $\rightarrow$ NP VP	
VP $\rightarrow$ V CP	
CP $\rightarrow$ C S	*said that the flight was late*

Sannsynlighet

- **Felles** sannsynlighet ("joint probability") for to hendelser A og B er sannsynligheten for at begge hendelser finner sted, $P(A \cap B)$
- **Uavhengighet**: sannsynligheten for en påvirker ikke sannsynligheten for den andre. To hendelser er uavhengige hvis

$$P(A \text{ og } B) = P(A)P(B)$$

(Multiplikasjonsregelen)

Sannsynlighet

- **Betinget** sannsynlighet ("conditional probability")
- Lar oss håndtere **avhengige** hendelser
- Multiplikasjonsregelen:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

- **Kjederegelen** generaliserer multiplikasjonsregelen til flere hendelser:

$$P(A \cap B \cap C \cap D \cap E \cap \ldots) = P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C)\ldots$$

Språkmodeller

- Språkmodeller (n-grammodeller): sannsynlighetsmodeller som angir sannsynligheten for neste ord gitt de $n-1$ foregående ordene
- Kan også brukes til regne ut sannsynligheten for en hel setning
- Et n-gram er en sekvens av n ord (tokens)
  - 2-gram (bigram; $n = 2$) er en sekvens av to ord, feks *Johaug blir, blir utestengt, utestengt i, i ett, ett år*
  - 3-gram (trigram; $n = 3$) er en sekvens av tre ord, feks *Johaug blir utestengt, blir utestengt i, utestengt i ett, i ett år*

Språkmodeller

- Med en bigrammodell blir sannsynligheten for en streng $w_1, \ldots, w_k$ produktet av de individuelle ordenes sannsynlighet, slik:

$$P(w_1 \ldots w_k) = \prod_{i=1}^{k} P(w_i|w_{i-1})$$

- Vi bruker **kjederegelen** for sekvenser av hendelser samt **Markovantagelsen**
- Vi kan beregne disse sannsynlighetene ved tellinger fra et korpus

$$\frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

- Dette kalles **Maximum Likelihood Estimation**

Bayes regel

- Betinget sannsynlighet

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplikasjonsregelen:

$$P(A \cap B) = P(A|B)P(B)$$

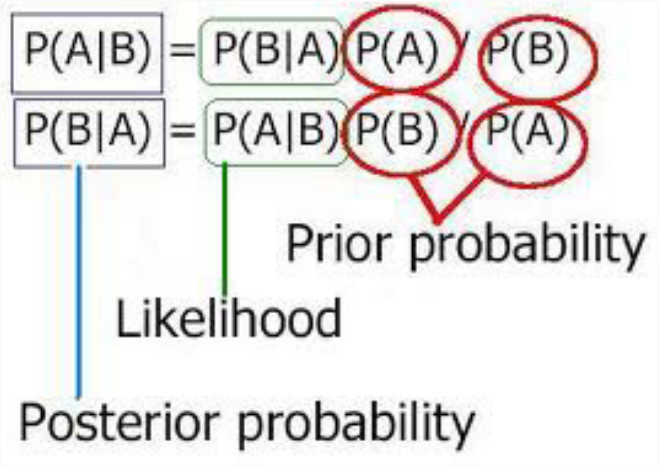
$$P(A|B)P(B) = P(B|A)P(A)$$

- Bayes regel

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes regel

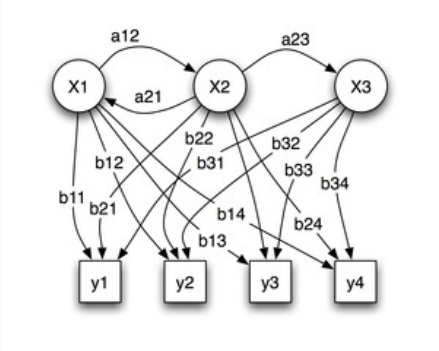
- Bayes regel/teorem viser hvordan vi kan beregne inverse sannsynligheter, dvs dersom vi vet $P(A|B)$, hvordan kan vi beregne $P(B|A)$?
- Noen begreper



Hidden Markov Model (HMM)

- HMM'er er en utvidelse av endelig tilstandsmaskiner (FSM'er)
- FSM'er defineres ved en mengde tilstander og en mengde transisjoner mellom disse i henhold til input-observasjoner
- En **vektet** FSM er en utvidelse av en FSM der hver transisjon er forbundet med en sannsynlighet: uttrykker hvor sannsynlig den overgangen er

HMM-tagging



- x = tilstander, y = observasjoner, a = transisjonssannsynligheter, b = observasjonssannsynligheter
- To typer sannsynligheter
  1. transisjonssannsynligheter
  2. observasjonssannsynligheter

HMM-tagging

- Bruker Bayes Teorem samt to forenklende antagelser

$$\hat{t}_1^n = \underset{t_1^n}{\text{argmax}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

- Med denne formelen kan vi beregne en taggsekvens for en gitt ordsekvens
- Sannsynlighetene beregnes fra korpus (MLE)

Naive Bayes

- Naive Bayes klassifiserer

$$\hat{s} = \underset{s \in S}{\text{argmax}} P(s) \prod_{j=1}^{n} P(f_j|s)$$

- Vi **trener** klassifisereren ved å beregne sannsynligheter fra et korpus (MLE)
- 2 sannsynligheter:
  1. prior-sannsynligheten for betydningen $P(s)$

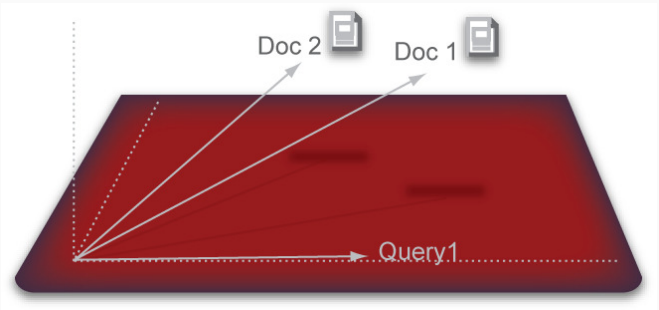
$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)}$$

  2. sannsynligheten for individuelle trekk $P(f_j|s)$

$$P(f_j|s) = \frac{count(f_j, s)}{count(s)}$$

Vektorrommodellen

- Representerer dokumenter som punkter i et høydimensjonalt rom
- Representasjonen kan enkelt (og effektivt) beregnes: trenger bare å telle ord
- Likhet mellom dokumenter blir til distanse i rommet
- Likhet i vektorrommet forutsier likhet i tema
  - Dokumenter som inneholder like ord handler ofte om det samme!



Termvekting

- Alle ord er ikke like informative
- Hvordan kan vi vekte ord slik at informative ord teller mer og ikke-informative ord teller mindre?
- **tf-idf vekting:**

$$\mathrm{v}_{\mathrm{movie}} = \mathrm{tf}_{\mathrm{movie}} \times \mathrm{idf}_{\mathrm{movie}}$$

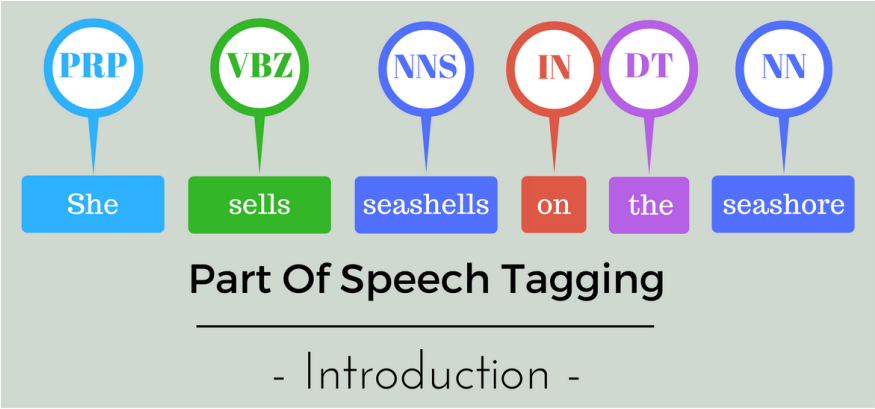
Språkteknologiske Oppgaver & Applikasjoner

Tokenisering

- Dele opp en tekst i løpende ord
- Første skritt i nesten alle språkteknologiske oppgaver
- Problematiske tilfeller:
  - forkortelser: *f.eks.*
  - bindestrek: *Oslo-borgeren*
  - mellomrom: *New York*
  - URL'er
  - ...

Ordklassetagging

- Input: streng av ord og en spesifisert mengde tagger (taggsett)
- Output: en tagg per ord



Flertydighet

Algoritmer for ordklassetagging

- To hovedkategorier:
  1. **Regelbaserte taggere**: stor database med håndskrevne regler. Eksempel: *book* er substantiv, og ikke verb, dersom etterfølger en determinativ
  2. **Probabilistiske taggere:** bruker et ordklassetagget korpus ("treningskorpus") til å beregne sannsynlighet for en gitt tagg i en gitt kontekst
     - Hidden Markov Models (HMM-taggere)

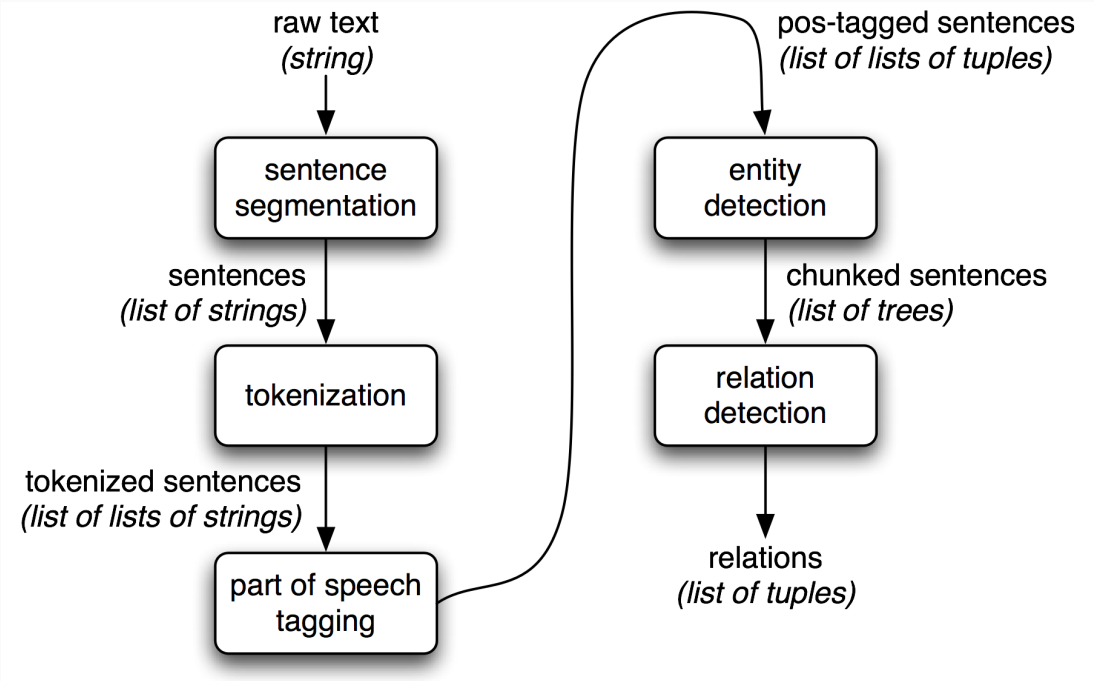
Syntaks i språkteknologi

- **Chunking:**
  - dele setningen inn i en sekvens **"chunks"**
  - en chunk inneholder et **hode**, muligens med noen funksjonsord/modifikatorer først
    [walk] [straight past] [the lake]
- forenklede konstituenter ("fram til hodet")
- ikke komplett syntaktisk beskrivelse, men tilstrekkelig for mange applikasjoner
- **ikke-rekursive**: en chunk kan ikke inneholde en chunk av samme kategori

Betydningsdisambiguering (Word Sense Disambiguation)

- Word Sense Disambiguation (WSD)
  - gitt en setning med et spesifikt ord ("target word") og en liste med betydninger (f.eks. fra WordNet)
  - angi den betydningen som passer best for målordet i den setningen
- Klassifisering fra annotert datasett: **supervised** klassifisering
  - hente ut trekk som er sentrale for disambiguering, f.eks. ord i konteksten, syntaktiske funksjoner, osv.

Informasjonsekstraksjon (IE)



IR

- Informasjonsgjenfinning (Information Retrieval – IR): lagring og gjenfinning av alle slags media (fokus her: tekst)
- Automatisk finne fram til dokumenter som er relevante for en søkestreng
- **Vektorrommodellen** brukes i de fleste moderne systemer, inkludert søkemotorer som Google

Maskinoversettelse

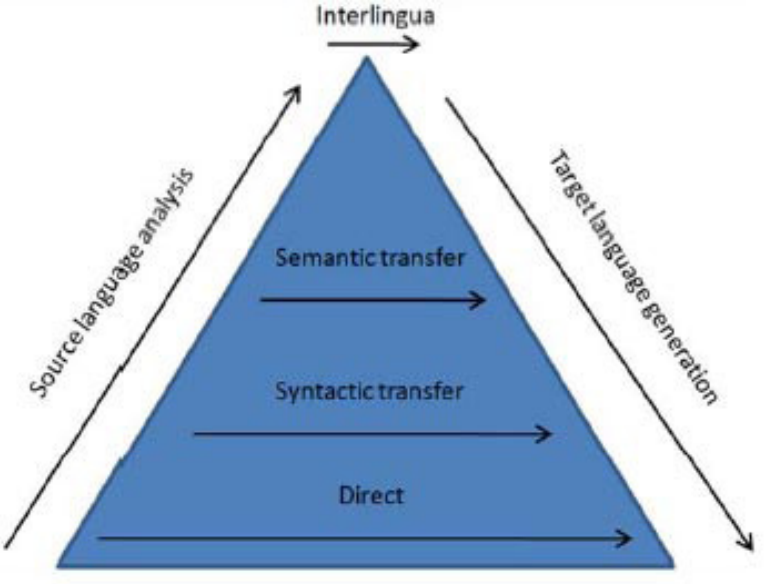


Figure 1: The Vauquois triangle

Statistisk MT