



Introduktion till språkteknologi

Orduppdelning och ordklassanalys

Mattias Nilsson
mattias.nilsson@lingfil.uu.se



Schemaöversikt – återstående tillfällen (MN)

- 4/11: **Orduppdelning och ordklassanalys** (kl. 10-12)
- 4/11: Labbtillfälle (korpusbeh.) (kl. 13-15)
- 17/11: **Språk och kognitiv modellering** (kl. 12-14)
- 19/11: Labbtillfälle (ordklasstagning) (kl. 13-15)
- 25/11: Labbtillfälle (ordklasstagning) (kl. 10-12)
- 25/11: Labbtillfälle (ordklasstagning) kl. 13-15



Innehåll

- Textnormalisering och tokenisering
- Ordklasstagning
- Utvärdering av ordklasstagare
- Kort om parsning
- Kort om labbuppgift



Textnormalisering

Att ge en text ett enhetligt och strukturerat format som kan användas för vidare databehandling

Inbegriper främst:

- Meningssegmentering
 - Identifiera meningslut
- Tokenisering
 - Dela upp texten i dess *tokens*, d.v.s. löpord

Inledande steg i nästan all språkteknologisk programutveckling; t.ex. för ordklasstagning, parsning och maskinöversättning



Tokenisering

Vad räknas som ett ord?

- Standarddefinition:

"a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks" (Kučera & Francis 1967)

Finns dock flera problematiska fall...



Tokenisering – problematiska fall

Punkt

- Kan utgöra del av förkortning: *Lex.*
- Kan utgöra både förkortningspunkt och meningsavslutande punkt på samma gång:
Syftet är att minska samhällets utgifter för färdtjänst, sjukresor etc.

Apostrof

- Kan fungera som citationstecken eller som del av ordet:
'the girls' vs. the girls' toys
isn't, won't, I'll



Tokenisering – problematiska fall

Bindestreck

- ett eller flera ord?
- co-operate, 26-year-old*

Mellanslag

- särskrivna sammansättningar: *cat food*
- egennamn: *New York*
- fraser: *i alla fall, i dag*
- sifferuttryck: *100 000*



Ordklasstagning

- eng: Part-of-Speech (PoS) tagging
- mål: tilldela varje ord i en text dess korrekta ordklass
- delproblem i många tillämpningar
 - syntaktisk analys (parsning)
 - språkgranskning (t.ex. grammatikkontroll)
 - maskinöversättning
 - informationssökning
 - talsyntes



Ordklasstagning

Kan göras med olika grad av granularitet:

- Endast ordklass:
Det/**PRON** är/**VB** ingen/**DT** svår/**AJ** uppgift/**NN**
- Ordklass samt morfologisk information:

Det/**PRON_NEUT_SING**
 är/**VB_PRESENT_AKT**
 ingen/**DT_INDEF_UTR_SING**
 svår/**AJ_UTR_SING_INDEF**
 uppgift/**NN_UTR_SING_INDEF**



Metoder för ordklasstagning

- Manuell (för hand)
 - tids- och resurskrävande
 - Stor risk för inkonsekvenser, otillättna taggar, mm
- Automatisk
 - snabb, effektiv
 - förutsätter ett automatiskt taggningsprogram, en s.k. ordklasstagare



Ordklasstagning

- Enkla fall
 - Icke-tvetydiga ord
- Problematiska fall:
 - tvetydiga ord eller fraser: *modern såg*



Ordklasstagning

- Problematiska fall:
 - tvetydiga ord
 - modern såg
 - modern/**NN** såg/*v* *alt.* modern/**AJ** såg/**NN** ?
 - modern såg flickan
 - Kontexten avgör ordklass
- I Brown Corpus är 40% av alla löpord tvetydiga, dvs kan tilldelas mer än en ordklasstag



Ordklasstagningens delsteg

1. **Tokenisering** – dela upp texten i dess tokens
lära|ren/ var/ i alla fall/ snäll/
2. **Tilldelning** – tilldela varje ord dess *möjliga* ordklasser
lära|ren/NN var/NN|VB|ADV|PRON i alla fall/ADV snäll/AJ
3. **Disambiguering** – välj en unik tagg för varje ord
lära|ren/NN var/VB i alla fall/ADV snäll/AJ



Typer av taggningsprogram

- Regelbaserade
- Datadrivna (korusbaserade)
 - Transformationsbaserad taggning
 - Statistiskt baserad taggning



Regelbaserad taggning

- Flera system utvecklade på 80-talet
- Manuellt definierade disambigueringsregler
- Tidsödande, kräver lingvistisk expertkunskap
- Exempel: SWETWOL
 - testa här: <http://www2.lingsoft.fi/cgi-bin/swetwol>
 - används bl.a. i grammatikkontrollen i Microsoft Word



Transformationsbaserad taggning

- Eric Brill 1992, 1995
- En av de första datadrivna ordklasstaggarna
- *Transformation-Based Error-Driven Learning*
- Bygger på regler, eller transformationer, men dessa härleds direkt från en färdigtaggad träningskorpus
- Systemet lär sig reglmallar genom att detektera och ändra felaktiga taggningar



Transformationsbaserad taggning

- Grundprincip
 - **lexikonuppslagning**: välj den mest frekventa taggen för varje ord enligt en träningskorpus, annars använd heuristik
 - **disambiguering**: ändra den initiala taggningen m.h.a. kontexten (taggar & ord)
 - **trigger**: lexikala och kontextuella särdrag som utlöser transformationsregler
 - **transformationer**: omskrivningsregler som förändrar en tagg vid en viss kontext



Transformationsbaserad taggning

- Exempel på en transformationsregel:
"Om nuvarande tagg är adjektiv och nästföljande tagg är adjektiv, ändra då nuvarande tagg till adverb":
 - Ursprungstagg: AJ
 - Ersättningstagg: AB
 - Trigger: Nästa tagg är AJ
 - Transformation: relativt/AJ svårt/AJ → relativt/AB svårt/AJ



Statistiskt baserad taggning

- Taggningen baseras helt på beräkning av sannolikheter. Dessa beräknas utifrån relativa frekvenser i en (manuellt) färdigtaggad *träningsskorpus*
 - Vi antar att nya data liknar träningsdata
- Väljer för varje ord den mest sannolika taggen enligt kontexten (föregående taggar/ord)
- Ger normalt hög precision (ca 95-97% korrekta taggar)



Statistiskt baserad taggning


- Ur träningsdata beräknas två typer av sannolikheter
 - Lexikala sannolikheter:
 - Sannolikheten att ord w_i realiserar taggen t_i :
 - $P(\text{ord}|\text{tagg})$
 - Uppskattas med: $\text{Cnt}(\text{ord}, \text{tagg}) / \text{Cnt}(\text{tagg})$
 - Kontextuella sannolikheter:
 - Sannolikheten att tagg_i följer omedelbart efter tagg_{i-1}:
 - $P(\text{tagg}_i|\text{tagg}_{i-1})$
 - Uppskattas med: $\text{Cnt}(\text{tagg}_{i-1}, \text{tagg}_i) / \text{tagg}_{i-1}$
- Välj den taggsekvens som maximerar sannolikheten för:
 - $P(\text{ord}|\text{tagg}) * P(\text{tagg}_i|\text{tagg}_{i-1})$



Statistiskt baserad taggning - exempel

1. He/PRON wants/VB to/TO race/?
2. He/PRON won/VB the/DT race/?


- Vilken tagg skall vi välja för *race* i 1?
- VB eller NN?
 - Vi beräknar följande:
 - $P(\text{race}|\text{VB}) * P(\text{VB}|\text{TO})$
 - $P(\text{race}|\text{NN}) * P(\text{NN}|\text{TO})$



to/TO race/?

$P(\text{race} \text{VB}) \cdot P(\text{VB} \text{TO})$ $P(\text{race} \text{VB}) = .00003$ $P(\text{VB} \text{TO}) = .34$ $P(\text{race} \text{VB}) \cdot P(\text{VB} \text{TO}) = .00001$	$P(\text{race} \text{NN}) \cdot P(\text{NN} \text{TO})$ $P(\text{race} \text{NN}) = .00041$ $P(\text{NN} \text{TO}) = .021$ $P(\text{race} \text{NN}) \cdot P(\text{NN} \text{TO}) = .000007$
--	--

to/TO race/VB




Statistiskt baserad taggning

N-gram och kontext

- Vi kan variera kontextlängden då vi beräknar kontextuella sannolikheter
- I vårt exempel bestod kontexten endast av taggen för föregående ord, en sk. bigram modell.
 - > Taggar för de två föregående orden: trigram modell
 - > 4-gram, 5-gram ...
- Trigrammodeller ger bäst resultat

Ju längre N-gram, desto färre exempel i träningsdata
Längre sekvenser än trigram kräver väldigt mycket träningsdata



Statistiskt baserad taggning

- TnT - *Trigrams'n'Tags*, Brants, 2000
 - HMM-baserad (Hidden Markov Model)
 - Kontextmodell: trigram
- Resurs på institutionen:
 - TnT tränad för svenska (tränad på SUC-korpusen)
 - Bäst i test för svenska (Megyesi, 2001): 93.5%



Utvärdering av ordklasstagare

Ett taggningssystem utvärderas främst med avseende på korrekthet (*accuracy*). Den mäts vi genom att beräkna andelen korrekt taggade ord:

$$\text{korrekthet} = \frac{\text{Antalet korrekt taggade löpord}}{\text{Totala antalet löpord}}$$

Förutsätter att vi har tillgång till en manuellt taggad version av den korpus vi testar taggaren på, en s.k. *guldstandard*



Utvärdering av ordklasstagare

- Ordklasstagarens resultat jämförs med:
 - Guldstandard, ca. 96-97%
 - "Baseline": t.ex. andelen korrekt taggade ord då vi väljer den mest frekventa taggen för ett ord, utan hänsyn till kontexten, ca 90-91%
 - De bästa ordklasstagarna, ca 95-97%
- Ett taggningssystem kan också utvärderas med avseende på andra aspekter än korrekthet, t.ex. snabbhet.



Faktorer som påverkar resultatet

- Storleken på träningskorpus
 - Allmänt gäller: ju större träningskorpus desto bättre resultat
- Tagguppsättning
 - ju fler specifika taggar desto fler fall av tvetydighet
- Källan till tränings- och testdata (likheten mellan dem)
- Mängden okända ord (*sparse data*)



Kort om parsning

Problem: Att tilldela en mening dess syntaktiska struktur

- Förutsätter vanligtvis att orden har tilldelats ordklass
- Ett svårare problem än ordklasstagning
- Vid parsning kan en mening tilldelas väldigt många olika syntaktiska analyser (parse-träd)
 - Hur skall vi bära oss åt för att välja den rätta analysen?



Statistisk parsning

Reglerna i en frasstrukturgrammatik kan utökas med sannolikheter för respektive regel som beräknas utifrån en träningskorpus.

Givet en sådan *probabilistisk frasstrukturgrammatik* kan vi beräkna den mest sannolika syntaktiska analysen för en mening.

- S → NP VP (1.0)
- NP → Det Noun (0.5)
- NP → Noun (0.3)
- NP → Det AJ N (0.2)
- VP → Verb NP (0.6)
- VP → Verb (0.4)



Kort om Labb 3 – Ordklasstagning

- Syfte: praktisk erfarenhet av att köra ett befintligt ordklasstagningsprogram
- Uppgift: ordklasstaggat texter hämtade från EuroParl
 - för svenska och engelska
 - kvalitativ utvärdering och analys av resultat
- Lab-pm tillgängligt från kurskansliet på onsdag 5/11
- Första gemensamma labbtillfälle: onsdag 19/11
- Rapport inlämnad senast 3/12
