

STATISTISKA METODER 3

ORDKLASSTAGGNING

- TAGGNINGSPROBLEMET
- TAGGNING AV ENSTAKA ORD
- TAGGDISAMBIGUERING
- GISSNING AV TAGGAR FÖR OKÄNDA ORD
- HUR BRA KAN EN TAGGARE BLI?

EXEMPEL PÅ TAGGNING

VID EN VÄG , PÅ EN
pp dt. utr. sin. ind nn. utr. sin. ind. nom mid pp dt. utr. sin. ind
STEN , SATT EN LITEN
nn. utr. sin. ind. nom mid vb. prt. akt dt. utr. sin. ind jj. pos. utr. sin. ind. nom
FLUCKA .
nn. utr. sin. ind. nom mad

EXEMPEL PÅ SVÄRTAGGADE MENINGAR:

HON VAR RINGARE.

pn. utr. sin. def. sub vb. prt. akt. kop nn. utr. sin. ind. nom
nn. utr. sin. det. nom jj. kom. utr. neu. sin. plu. ind/ det. nom

DET VAR EN GRÖN EN.

BUSHS PLOTSLIGA BESLUT KOM OVÄNTAT.

SAS FÖRSTA PLAN KOMMER SNART.

ETT VÄLDIGT GENOMARBETAT VERK.

TILLÄMPNINGAR AV TAGGNING

- GRAMMATIKKONTROLL
- FÖRBÄTTRAD STAVNINGSKONTROLL
- MASKINÖVERSÄTTNING
- INFORMATIONSSÖKNING
- TALIGENKÄNNING
- SOM FÖRBEREDELSE TILL PARSNING I DIVERSE TILLÄMPNINGAR

TAGGNING AV ISOLERADE ORD

VILKA TAGGAR KAN ETT ORD HA?

METOD 1: MANUELL TAGGNING

SKAPA ETT LEXIKON FÖR HAND GENOM ATT
(LÅTA EN LINGVIST) TAGGA EN MASSA ORD.

METOD 2: ANVÄNDNING AV KORPUS

NOTERA HUR VARJE ORD HAR TAGGATS I
(TILL EXEMPEL) SVC OCH SUKA DÄRIGENOM ETT LEXIKON.

METOD 3: AUTOMATISK TAGGNING

ANVÄND MORFOLOGI FÖR ATT AUTOMATISKT TAGGA
ALLA REGELBUNDNA ORD.

IDÉ: UTVIDGA STAVAS SUFFIXREGLER MED TAGGAR.

EXEMPEL: ornas a, an, or # nn. utr. plu. gen

UTGÅENDE FRÅN STAVAS ORDLISTA FRÅS 700000 TAGGADE ORD.

PRÖVA GÄRNA TAGGSTAVA PÅ

<http://www.nada.kth.se/theory/projects/gramska/>

TAGGDISAMBIGUERING

HUR SKA VARJE ORD I EN TEXT (MENING) TAGGAS?

METOD 1: REGELBASERAD DISAMBIGUERING

ORD SOM KAN HA FLERA TAGGAR DISAMBIGUERAS MED HJÄLP AV REGLER SOM TAR HÄNSYN TILL SAMMANHÅNGET, TILL EXEMPEL

ETT ORD SOM FÖREGÅS AV EN ARTIKEL OCH FÖLJS AV ETT SUBSTANTIV SKA TAGGAS SOM ETT ADJEKTIV, DVS
DT X NN $\Rightarrow$ X/JJ

ATT SKAPA EN VÄL FUNGERANDE UPPSÄTTNING REGLER ÄR MYCKET JOBBIGT, MEN RESULTATET KAN BLI MYCKET BRA.

FÖRSÖK HAR GJORTS ATT AUTOMATISKT GENERERA DISAMBIGUERINGSREGLER.

GISSNING AV TAGGAR FÖR OKÄNDA ORD

METOD 1: GISSA PÅ ALLA TAGGAR SOM FINNS OCH LÅT DISAMBIGUERAREN VÄLJA.

FÖRBÄTTRING: GISSA INTE PÅ SLUTNA ORDKLASSER (T EX PRONOMEN, PREPOSITIONER, KONJUNKTIONER)

METOD 2: OM ORDET KAN BILDAS SOM EN SAMMANSÄTTNING AV KÄNDA ORD (SOM I STAVA) SÅ TAGGAR VI ORDET PÅ SAMMA SÄTT SOM EFTERLEDET.

METOD 3: ANALYSERA HUR KÄNDA ORD SOM SLUTAR PÅ SAMMA BOKSTÄVER SOM DET OKÄNDA ORDET TAGGAS OCH TAGGA DET LIKADANT

OVANSTÄENDE METODER KLARAR AV ATT TAGGA OKÄNDA ORD RÄTT TILL 92%

METOD 2: PROBABILISTISK DISAMBIGUERING

ANVÄND STATISTIK FÖR ATT TAGGDISAMBIGUERA MENINGAR.

VANLIGASTE MODELLEN: N-TE ORDNINGENS MARKOVMODELL

DÄR $N=1$ ELLER $N=2$.

ANTAGANDEN:

1. VARJE ORD ÄR OBEROENDE AV ANDRA ORD OCH DERAS TAGGAR.
2. VARJE TAGGNING BEROR ENDAST PÅ NÄRMAST N FÖREGÅENDE TAGGAR I MENINGEN.

TAGGNINGSPROBLEMFORMULERING FÖR $N=2$:

BÄSTA TAGGNINGEN AV $w_1, w_2, \dots, w_n$ ÄR DEN TAGGNING $t_1, t_2, \dots, t_n$ SOM MAXIMERAR

$$\prod_{i=1}^n P(t_i | t_{i-2}, t_{i-1}) \cdot P(w_i | t_i)$$

GIVET UPPSKATTNINGAR AV SANNOLIKHETERNA KAN TAGGNINGEN EFFEKTIVT BERÄKNAS MED DYNAMISK PROGRAMMERING MED HJÄLP AV VITERBIS ALGORITM.

NADAS NYA TAGGARE

GJORD 1998 AV JOWAN CARLBERGER OCH VIGGO KANN
KAN PROVAS PÅ

<http://www.nada.kth.se/theory/projects/granska/>

ARBETSGÅNG VID UTVECKLINGEN:

1. ANVÄND 86% AV SVC FÖR BERÄKNING AV ORD- OCH TAGGSTATISTIK (TAGGTRIGRAM).
2. ANVÄND 7% AV SVC SOM TRÄNINGSTEXT FÖR OPTIMERING AV PARAMETRAR I MODELLEN.
3. ANVÄND RESTERANDE 7% AV SVC FÖR UTVÄRDERING AV TAGGAREN.

RESULTAT:

TAGGAREN TAGGAR RIKTIGT TILL 96,5%