Taggning av räkneord som årtal eller andra räkneord, Språkteknologi 2D1418, HT 01

Jonas Sjöbergh, 761029-0178, jsh@nada.kth.se

15 oktober 2001

1 Bakgrund

1.1 Kort om taggning

Taggning innebär att man ger varje ord i en text en ”tagg”, vilket är en samling värden som beskriver egenskaper hos ordet. En tagg kan t.ex. se ut som ”bilen <lemma=bil, wc=nn, num=sin, spec=def>”, vilket innebär att ordet ”bilen” är ett substantiv i bestämd from singular och att grundformen är ordet ”bil”.

Vid taggning av text finns i huvudsak två angreppssätt, regelbaserad taggning och statistisk taggning. Regelbaserad taggning innebär att man ställer upp ett antal regler för vilken tagg ett ord ska tilldelas, medan statistiska metoder innebär att man första analyserar stora mängder text och lagrar statistik, t.ex. över vilka ord som brukar få vilken tagg.

1.2 Problemformulering

Vid taggning kan det vara användbart att kunna skilja årtal i texten från andra räkneord. En regelbaserad metod för att göra detta beskrivs nedan. Ett exempel på då det är bra att skilja på årtal och andra räkneord är vid talsyntes (maskinuppläsning av text), då uttalet av årtal skiljer sig från andra räkneord (”år nittonhundranittionio” men ”ettusenniohundranittionio kronor”).

Vid talförståelse (maskintolkning av tal) är naturligtvis uttalet av räkneord på samma sätt en stor ledtråd till om det är ett årtal eller ej. I denna rapport används dock bara text som indata.

1.3 Kort om programmet Granska

Granska är ett program för automatisk grammatikkontroll som utvecklas vid Nada på KTH i Stockholm. Granska har en samling regler som definierar olika typer av fel (t.ex. ”den huset”) och ger rättningsförslag (”det huset”). Granska taggar texten och reglerna kan sedan beskriva feltyper med hjälp av den information som finns i taggar, t.ex. ”om ett adjektiv som föregår ett substantiv inte har samma genus (utrum/neutrum) som substantivet är det fel”.

Regler i Granska skrivs genom ett vänsterled som talar om hur texten ska se ut för att regeln ska vara applicerbar följt av ett högerled som talar om vad som ska göras då denna regel uppfylls. Ett litet exempel på en regel:

```
exempelregel@exempelkategori {
 X (wordcl=dt),
 Y (wordcl=jj),
 Z (wordcl=nn & gender != Y.gender)
-->
 mark(X Y Z)
 info("genusfel")
 action(scrutinizing)
}
```

Denna regel säger att då en determinerare följt av ett adjektiv följt av ett substantiv förekommer i texten och adjektivet inte har samma genus som substantivet (vänsterledet, före ”–>”) ska de tre orden markeras i texten och beskrivningen ”genusfel” ska skrivas ut. En introduktion till Granska finns på http://www.nada.kth.se/theory/projects/granska/popular.html

1.4 Intressanta webbplatser

På http://www.nada.kth.se/theory/projects/granska/scrutinizer-rules-demo.html kan man pröva Granska med egna regeltillägg.

På http://www.nada.kth.se/theory/projects/granska/rapporter/rulelang20010308.pdf finns dokumentation om Granskas regelspråk.

På http://www.student.nada.kth.se/~jsh/taltagg.txt finns de regler som använts i denna rapport.

2 Lösning

Här presenteras en regelbaserad lösning av problemet att skilja ut årtal från andra räkneord. Reglerna är skrivna i Granskas regelspråk. Reglerna beskrivs i avsnitt 2.1 och 2.2.

2.1 Regler

De regler som använts finns sammanfattade nedan och hela regelsamlingen i Granskaformat finns i appendix A. Prioriteringsordningen mellan reglerna är uppifrån och ned, och så fort en regel passar in så stannar man, dvs den översta regel som gäller för ett räkneord är den som bestämmer om det är ett årtal eller inte.

1. Om ett räkneord följs av ”e.Kr.”, ”f.Kr.” eller någon variant av stavning av dessa så är det ett årtal.
2. Om ett ord är på formen ”X-talet”(t.ex. 1700-talet) eller någon liknande variant så är X ett ”årtal” (det har åtminstone med årtal att göra).
3. Om ett räkneord föregås av ”år” eller någon variant (t.ex. ”åren”) är det ett årtal.
4. Om ett räkneord följs av ett godtyckligt antal adjektiv och sedan ett substantiv i pluralis är det inte ett årtal.
5. Om ett räkneord föregås av vissa nyckelord (t.ex. månadsnamn eller ”sommaren”) är det ett årtal.
6. Om ett räkneord är fyrsiffrigt är det ett årtal.
7. Om ett räkneord är på ”datumformat” (t.ex. 2001-10-16) består det i och för sig av bl.a. ett årtal, men klassas som ”datum” istället.
8. Om två räkneord förekommer på formen ”X - Y” (”åren 1998 - 2000”) är antingen båda årtal eller så är inget av orden årtal.
9. Räkneord som inte passar under någon av ovanstående regler är inte årtal.

En regel som sa att om ett räkneord föregås av ett substantiv i obestämd form singularis är det inte ett årtal testades också tidigare. Regeln var tänkt att ta hand om formuleringar som ”kapitel 5”, ”rum 1645” och ”rad 189”. Denna gav dock många fel, t.ex. för uttryck av typen ”Lanark dödade Wallaces fru 1297” och de uttryck den var tänkt att ta hand om var dels ovanliga och dels tenderade de att bli rätt ändå (genom regeln att räkneord vanligen inte är årtal), så den regeln togs bort.

2.2 Motivering av regler

- Regel 1 kräver inte så mycket förklaring, ”e.Kr.” används i princip bara efter årtal.
- Regel 2 innebär att ”1700-talet” betraktas som ett årtal, vilket det i strikt mening inte är, men det har med årtal att göra. Vill man inte betrakta ”1700-talet” som ett årtal är det bara att ta bort regel 2.

- Regel 3 säger att ”år 1976” och ”åren 1976 - 2000” betyder att ”1976” i dessa fall är ett årtal, vilket är det normala då ”år” används före ett räkneord.
- Regel 4 fångar text som ”1999 kronor”, ”50 år” och ”fem stora hus” och säger att detta inte är årtal utan en bestämning av antalet knutet till substantivet senare i meningen. Granskaregeln matchar även substantiv där numerus är obestämt, detta då enheter efter mätetal, såsom ”30 m”, oftast taggas som <nn>, utan några mer detaljer.
- Regel 5 är i princip likadan som regel 3, vanligtvis är ett räkneord som följer t.ex. ett månadsnamn ett årtal, ”sommaren 2001” eller ”januari 1900”. Man kan också tänka sig att lägga till fler nyckelord, vissa verb föregår ofta årtal (”han dog 1976”). Ett problem med sådana verb är dock att de även kan följas av andra räkneord (”av de insjuknade dog 30”).
- Regel 6 är en heuristisk metod som fungerar ganska bra i vanlig text, fyrsiffriga tal är ofta årtal. Två problem är att det förekommer även andra fyrasiffrorskombinationer samt att i vissa typer av texter, t.ex. historieböcker, förekommer många årtal som inte är fyrsiffriga.
- Regel 7 innebär att räkneord på särskilda datumformat specialtaggas som datum, trots att de innehåller årtal. Oftast är det intressantare att veta att ett uttryck är ett datum än att det finns ett årtal i närheten. Eventuellt kan man inkludera information om huruvida ett datum innehåller årtal eller ej.
- Regel 8 säger att två räkneord med ”-” mellan är av samma typ, som i ”åren 1976 - 2000” eller ”5 - 2 blir 3”, vilket är rimligt att anta.
- Regel 9 säger att om inget av ovanstående gäller så är det inte ett årtal, vilket oftast ger rätt beteende.

3 Utvärdering

3.1 Resultat

De enkla regler som använts ger mycket bra resultat. På de texter som testats blir de flesta (över 90%) räkneord rätt uppmärkta. Resultatet varierar dock beroende på texten, historiska texter där årtal skrivs som ”år 0” eller ”44 f.Kr.” fungerar bra, medan texter där sådana markörer inte används (t.ex. ”Julius Caesar (100 - 44)”) fungerar dåligt.

Texter med många språkliga fel fungerar också dåligt (”Jag köpte 1999 bingolott igår.”)

3.2 De bästa reglerna

Regel 4 om att räkneord som följs av ett substantiv i pluralis inte är årtal är mycket användbar, den har aldrig fel (i de test som gjorts) och den täcker många fall.

Reglerna 1, 2 och 3, med ”år”, ”e.Kr.” och ”-talet”, är mycket användbara för att täcka in årtal, och de ger inte heller några falska träffar (i de test som gjorts).

Regel 6, om att fyrsiffriga tal är årtal, täcker in väldigt många årtal som inte täcks av någon annan regel, men alla fyrsiffriga tal är inte årtal, så regeln ger upphov till en del falska träffar. Hur många falska träffar det blir beror mycket på den text som analyseras (även hur många riktiga årtal man hittar beror på texten, i historietexter är årtal ofta inte fyrsiffriga).

Grundregeln (regel 9) att de flesta räkneord (de som inte täckts av någon annan regel) inte är årtal fungerar ganska bra. Många räkneord hamnar under denna regel och de allra flesta är inte årtal, undantaget i vissa historietexter.

3.3 Problem

Ett Granskarelaterat problem är att regel 6 just nu inte kontrollerar om ett räkneord är fyrsiffrigt, eftersom det inte går att kontrollera på ett enkelt sätt i Granska. Istället kontrolleras om räkneordet består av fyra tecken. Detta leder till problem med ord som ”fyra” och ”40,9”, så därför görs ytterligare några test, bl.a. kontrolleras att sista tecknet inte är ”a” och att inget av de två tecknen i mitten är ”.”, ”,” eller ”:”. Ett annat problem är att ”1250-1400” taggas som ett enda räkneord, vilket gör att regeln för fyrsiffriga tal inte upptäcker dessa två räkneord, eftersom de betraktas som ett nio tecken långt räkneord.

I övrigt är det vanligt med fel där fyrsiffriga räkneord som inte är årtal upptäcks av regel 6 och då märks fel. Ett sådant exempel är ”rum 1645”. Ett annat vanligt fel är att årtal som inte är fyrsiffriga inte upptäcks av någon regel och därför inte märks som årtal. Exempel på uttryck där årtal inte upptäcks är ”Lanark dödade Wallaces fru 297”, ”600- och 500-talet” (där ”600-” märks fel) och ”720 grundades Birka”.

Räkneord skrivna med bokstäver blir också problematiskt om de särskrivs, ”åtta tusen f.Kr.”, eftersom de då taggas som två separata räkneord, talet ”åtta” (regel 9) och årtalet ”tusen” (regel 1).

3.4 Möjliga förbättringar

En förbättring vore att lägga till fler regler liknande den om att ”X - Y” innebär att räkneorden ”X” och ”Y” har samma typ, för andra typer av uttryck, t.ex. ”X och Y”. En snarlik förbättring är att om man har en följd av räkneord efter varandra ska de alla ha samma typ (”åtta tusen tre hundra kaniner”).

De Granskarelaterade problemen borde kunna rättas till ganska enkelt, t.ex. kontrollera om räkneord verkligen är fyrsiffriga och inte som nu fyra tecken långa.

A Regler i Granskas regelspråk

```
category taltagg {
  info("Taltaggning")
  link("http://www.nada.kth.se/~jsh/taltagg.html"
  "Jonas taltaggningstest")
}

datum@taltagg {
 X (wordcl=rg & token=TOKEN_DATE)
-->
 mark(X)
 info("Datum")
 jump(endlabel)
 action(scrutinizing)
}

nnrg1@taltagg {
 Y (wordcl=nn & real_text.length > 5
    & (real_text.substr(real_text.length-4, 4)="-tal"
       | real_text.substr(real_text.length-5, 5)="-tals"))
-->
 mark(Y)
```

```
 info("Årtal")
 jump(endlabel)
 action(scrutinizing)
}

nnrg2@taltagg {
 Y (wordcl=nn & real_text.length > 6
    & (real_text.substr(real_text.length-6, 5)="-tale"
       | real_text.substr(real_text.length-7, 5)="-tale"))
-->
 mark(Y)
 info("Årtal")
 jump(endlabel)
 action(scrutinizing)
}

efkr@taltagg {
 X (wordcl=rg),
 Y (tolower(text)="f.kr" | tolower(text)="e.kr"
    | tolower(text)="f. kr" | tolower(text)="e. kr"
    | tolower(text)="f.kr." | tolower(text)="e.kr."
    | tolower(text)="f. kr." | tolower(text)="e. kr."
    | tolower(text)="f kr." | tolower(text)="e kr."
    | tolower(text)="f kr" | tolower(text)="e kr")
-->
 mark(X)
 info("Årtal")
 jump(endlabel, 1)
 action(scrutinizing)
}

talet1@taltagg {
 Y (wordcl=rg & real_text.length > 5
    & (real_text.substr(real_text.length-5, 4)="tale"
       | real_text.substr(real_text.length-3, 3)="tal"
       | real_text.substr(real_text.length-4, 4)="tals"
       | real_text.substr(real_text.length-6, 4)="tale"))
-->
 mark(Y)
 info("Årtal")
 jump(endlabel)
 action(scrutinizing)
}

talet2@taltagg {
 Y (wordcl=rg),
 X (text="talet")
-->
 mark(Y)
 info("Årtal")
```

```
 jump(endlabel, 1)
 action(scrutinizing)
}

år1@taltagg {
 Y (wordcl=rg & real_text.substr(0,2)="år")
-->
 mark(Y)
 info("Årtal")
 jump(endlabel)
 action(scrutinizing)
}

år2@taltagg {
 X (text="år" | text="året" | text="åren"),
 Y (wordcl=rg)
-->
 mark(Y)
 info("Årtal")
 jump(endlabel, 1)
 action(scrutinizing)
}

nnplu@taltagg {
 X (wordcl=rg),
 Z (wordcl=jj)*,
 Y (wordcl=nn & (num=plu | num=undef))
-->
 mark(X)
 info("Tal")
 jump(endlabel, 2)
 action(scrutinizing)
}

nyckelord@taltagg {
 Y (text="januari" | text="februari" | text="mars" | text="april"
    | text="maj" | text="juni" | text="juli" | text="augusti"
    | text="september" | text="oktober" | text="november"
    | text="december" | text="hösten" | text="sommaren"
    | text="vintern" | text="våren"
    | text="halvåret" | text="kvartalet"),
 X (wordcl=rg)
-->
 mark(X)
 info("Årtal")
 jump(endlabel, 1)
 action(scrutinizing)
}

fyrsiffrig@taltagg {
```

```
 X (wordcl=rg & real_text.length=4
   & real_text.substr(real_text.length-1, 1)!="a"
   & real_text.substr(real_text.length-1, 1)!="v"
   & real_text.substr(real_text.length-2, 1)!=","
   & real_text.substr(real_text.length-3, 1)!=","
   & real_text.substr(real_text.length-2, 1)!=":"
   & real_text.substr(real_text.length-3, 1)!=":"
   & real_text.substr(real_text.length-2, 1)!="."
   & real_text.substr(real_text.length-3, 1)!=".")
-->
 mark(X)
 info("Årtal")
 jump(endlabel)
 action(scrutinizing)
}

basregel@taltagg {
 X (wordcl=rg)
-->
 mark(X)
 info("Tal")
 action(scrutinizing)
}

ordningstal@taltagg {
 X (wordcl=ro)
-->
 mark(X)
 info("Tal")
 action(scrutinizing)
}
```