Lösningsförslag för tentamen i Språkteknologi, DH2418 2009-10-22

1. Klustring (4 poäng)

a) Resultatet av att applicera en partitionerande algoritm är en platt gruppering, en partition, av objekten.

Resultatet av en hierarkisk är en hierarki där kluster består av kluster. Denna hierarki kan vara från enskilda objekt ända upp till ett kluster bestående av alla objekt, eller någon del av denna fullständiga struktur.

b) Nej, det beror på vad klustringen ska användas till och vem som ska använda den.

2. Utvärdering (4 poäng)

Täckning och precision behöver tänkas igenom för varje enskilt tillämpningsområde. För grammatikkontroll ser det ut såhär:

Täckningen ger ett mått på hur många fel som hittas i en text. Precisionen ger ett mått på hur många av de alarm som ges som är korrekta.

a= antal korrekt detekterade fel = antal korrekta alarm
b = falska alarm
c = missade fel
täckning/recall R = a/(a+c)
precision P = a/(a+b)

3. Informationstäthet (4 poäng)

a) Entropin/ordlängden: -sum(p_i log(p_i))/medelordlängden där p_i är sannolikheten för ord i.
Även denna form godkänns: -sum(p_i log(p_i)/(antal bokstäver i ord i))

b) SMS-språk har högre informationstäthet än vanligt språk, eftersom skribenterna vill få fram budskapet med så få bokstäver som möjligt. Två exempel:
Ofta används förkortningar, vilket gör nämnaren i formeln mindre och därför värdet större.
Funktionsord som inte har någon större betydelse hoppas ofta över vilket gör att sannolikheten för innehållsorden ökar, samtidigt som meddelandets längd minskar. Eftersom entropin är störst då alla ord är lika sannolika så närmar man sej maximum när sannolikheten för de vanliga orden minskar och sannolikheten för de ovanliga orden ökar.

4. Syntax (4 poäng)

Syntaktisk parsning ger information om satsers och meningars strukturella uppbyggnad. Denna analys är ofta viktig för analyser på högre språkliga nivåer.

Informationsextrahering brukar ofta lyftas fram där till exempel ytparsning gör nytta. Den viktiga informationen finns t.ex. i nominalfraserna (NP). I dessa kan man hitta t.ex. ”Direktör Anders Johansson”, och i en annan fras står det ”nybliven pensionär”. Om informationsextraheringsprogrammet letar efter mönster som t.ex. NP + ”är” + NP så har programmet fått fram ett intressant samband. Genom frasgruppering får man fram att det är just direktör Anders Johansson som hittats, man får också veta att han är nybliven pensionär, och inte bara pensionär. Andra områden där syntaktisk parsning gör nytta är t.ex. maskinöversättning och grammatikkontroll.

5. Datorstödd språkinlärning (4 poäng)

De två mest intressant rollerna handlar om systemets ”intelligens”. Kan datorn verkligen ta rollen som lärare? Eller är det mer lämpligt att daton fungerar mer som ett smart verktyg. De roller som ligger längst ifrån varandra är när datorn används mer eller mindre endast som en bok, och de fall då datorn går in och styr lärande mer detalj, och försöker basera återkopplingen på studentens språk. I det första fallet används ingen språkteknologi, medan de i det andra fallet krävs språkteknologi och avancerad användarmodellering.

6. Statistisk lexikal semantik (4 poäng)

Metoden Random indexing skulle kunna hjälpa Lisa med arbetet genom att ta fram t.ex. de 20 ord som ligger närmast uppslagsordet som hon jobbar med. Antagandet handlar om att ord som förekommer i likartade kontexter (har en paradigmatisk relation) har likartade betydelser. Metoden ser ingen skillnad på synonymer och antonymer till ordet om de förekommer i likartade kontexter.

7. Utvärdering (6 poäng)

Det finns flera rimliga baselines att välja på, och givetvis många språkteknologiska problem. Här är några bra förslag på baselines:

1) textsammanfattning: linjärt urval av de inledande meningarna i en text
2) ordklasstaggning: den vanligaste taggen givet en annoterad referenskorpus
3) rättstavning (endast larm): slumpvis val mellan kategorierna "rättstavad"/ "felstvavd"

8. Informationssökning och maskininlärning (8 poäng)

a) En sökmotor. Indexera på alla anamneser i alla tidigare journaler. Ställ den nya journalens anamnes som sökfråga.

Bygg alltså en term-dokument-matris för alla tidigare journaler. Använd t.ex. cosinusmåttet som likhetsmått mellan den nya journalen (representerad på samma sätt) och alla andra. Lista de tidigare journalerna i likhetsordning.

b) En textkategoriserare. Träna på alla tidigare journaler med diagnos.

Representation: t.ex. term-dokument-matris.

Algoritm: t.ex. kategori-centroid-baserad. Skapa en centroid för varje diagnos-kategori och jämför nya journaler med dem. Anta att en journal ska ha den kategori vars centroid den är mest lik.

Skapa tränings-, optimerings- och test-mängder från alla journaler. Träna på träningsmängden och testa på optimeringsdata tills goda resultat erhålls. Testa sedan på testdata för att se att dessa resultat inte försämras avsevärt. Applicera sedan på nya exempel.

Denna kategorisering kommer inte alltid eftersom:

* texter inte är enkla objekt, de innehåller mycket brus
* journaler kan vara ännu mer "brusiga", vissa kan t.ex. vara väldigt ofullständiga.

9. Språkgranskning (4 poäng)

Pagrotsky – ett namn som inte finns med i lexikonet, förmodligen för att lexikonet är för gammalt. Stavas lexikon är modernare och innehåller Pagrotsky, och när Granska använder Stava så stavningskontrolleras inte egennamn.

runt – tolkas som ett adjektiv istället för preposition. Tolkningen av ordet parantes (se nedan) stör också tolkningen av runt.

parantes – tolkas som genitiv maskulin form av adjektivet parant och godkänns därför. Stava har en undantagsordlista där ordet ingår, och därför godkänns det inte av Stava.

tantkrämen – tolkas som en sammansättning av tant och krämen. Det gör även Stava. Det är dessutom svårt att upptäcka att det i det här fallet borde vara obestämd form också.

10. Morfologi (8 poäng)

För varje rad görs följande.

Dela upp raden vid snedstrecken. Hitta sista bindestrecket i första delen och lägg till det som föregår det som prefix före övriga delar.
Ta bort alla bindestreck och skriv ut delarna med mellanslag emellan.
Uppgiftens exempel blir då:

balladen balladerna
ballasten ballasterna
allasten allasterna
ballerinan ballerinorna
ballerinan ballerinorna
ballongem ballongerna
balsamen balsamerna

balsamen balsamerna
balsamerar balsamerarerade balsamerat

En körning genom Stava hittar enkelt felen ballongem och balsamerarerade. Gör dessutom en kontroll av bokstavsordningen så hittas även felet allasten.