

Tentamen för DH2418, Språkteknologi, 2009-10-22, kl. 14-18

Tentan består av 10 frågor. Totala poängsumman är 50 poäng. Tentan består av två delar. Inga hjälpmedel på del 1 av tentan. På del 2 får boken *Speech and Language Processing* av Daniel Jurafsky & James H. Martin användas. Utbytesstudenter får ta med sig lexikon, och använda engelska exempel om det behövs. Eventuella bonuspoäng intjänade under 2009 kommer att adderas till dina tentapoäng. Lämna in del 1 i separat konvolut innan du tar fram boken och fortsätter med del 2.

Betygsgränser

För att bli godkänd måste du ha minst 15 poäng på del 1 inkl. bonuspoäng. Betygsgränserna är enligt följande på hela tentan:

A: 45-50 poäng, B: 40-44 poäng, C: 35-39 poäng, D: 30-34 poäng, E: 25-29 poäng, Fx: 20-24 poäng, F: 0-19 poäng.

Del 1: Teoretisk del (20 poäng)

Skriv korta svar på dessa fem frågor, max 2 sidor totalt på denna del.

1. Klustering (4 poäng)

- a) Vad menas med en partitionerande respektive en hierarkisk klustringsalgoritm?
- b) Finns det en optimal klustering (resultatet av att köra en klustringsalgoritm) av en mängd texter, d.v.s. någon klustering som är bättre än alla andra. Motivera!

2. Utvärdering (4 poäng)

Förklara begreppen täckning och precision. Ange dessutom formlerna för hur du räknar ut täckning respektive precision för ett språkteknologiskt system.

3. Informationstäthet (4 poäng)

Olika språk har olika mängd information per tecken.

- a) Ange en formel för den förväntade informationstätheten för ett språk mätt i bitar/tecken.
- b) Vissa textgenrer är informationstätare än andra. Har SMS-språk (mobiltex-meddelanden) högre eller lägre informationstäthet än vanligt skrivet språk? Ge ett exempel på ett stildrag i SMS-språk och visa hur det påverkar formelns värde i a-uppgiften.

4. Syntax (4 poäng)

Syntaktisk parsning har historiskt dominerat det språkteknologiska fältet. Vilken nytta har man egentligen av denna analys? Beskriv åtminstone ett tillämpningsområde (applikation) där syntaktisk analys gör nytta. Förklara på vilket sätt den syntaktiska analysen gör nytta.

Vänd!

5. Datorstödd språkinlärning (4 poäng)

Inom området datorstödd språkinlärning finns det olika tankar om vilken roll datorn egentligen skall ha. Det som diskuteras är huruvida datorn är tillräckligt intelligent för den ena eller andra rollen. Beskriv minst två roller som datorn kan ta i datorstödd språkinlärning, och deras tänkbara för- respektive nackdelar.

Del 2: Problemdel (30 poäng)

Tänk på att även goda försök kan ge poäng. Skriv helst maximalt 1 sida per uppgift.

6. Statistisk lexikal semantik (4 poäng)

Lexikografen Lisa vill när hon utvecklar förlagets nya synonymordbok ha en lista med tänkbara synonymer och antonymer för varje uppslagsord hon arbetar med. Föreslå en metod som skulle kunna hjälpa Lisa att generera sådana listor. Vilket är det grundläggande lingvistiska antagandet som används av metoden? Kommer den föreslagna metoden att kunna skilja synonymer respektive antonymer från varandra?

7. Utvärdering (6 poäng)

För att åstadkomma en trovärdig utvärdering måste man som minst använda sig av en baseline (referenspunkt för lägsta gräns för acceptabilitet), och helst även av en s.k. human ceiling (mått på hur svår uppgiften är för en människa). Ge exempel på tre distinkt olika baseline-metoder och exemplifiera deras användning inom varsitt språkteknologiskt problem.

8. Informationssökning och maskininlärning (8 poäng)

Elektroniska patientjournaler innehåller en massa intressant information för den som har rätt att använda sig av den. Sjukhusen utnyttjar dem dock inte så mycket ännu.

Journalerna innehåller många olika sorters information, bl.a. en text, s.k. anamnes, som beskriver patientens hälsotillstånd i ord. Dessutom finns en diagnos (om läkaren lyckats ställa en sådan), t.ex. brutet ben, lunginflammation, etc. Det är lätt att se vilka journaler som har en diagnos och vilken diagnos de har.

Hur skulle du gå tillväga för att hjälpa ett sjukhus att utveckla följande tillämpningar som utnyttjar den i journaler lagrade informationen? Vilka datamängder skulle du skapa och hur? Vilka representationer skulle du använda?

a) Då en läkare inte lyckas ställa en diagnos vore det bra att få se de journaler som har en liknande anamnes som den aktuella journalen. Dessa kanske har en diagnos som hjälper läkaren. Hur tar du fram journaler med liknande anamnes?

b) En lat läkare vill ha hjälp att ta fram ett system som automatiskt sätter diagnos på nya patienter utifrån deras anamnes. Hur löser du det? Törs du säga att ditt system kommer att vara fullständigt säkert? Varför eller varför inte?

Vänd!

9. Språkgranskning (4 poäng)

Leif Pagrotsky iförd sin bästa söndagskostym satte parantes runt meningen om politiker som borstar tänderna utan tantkrämen.

Om man språkgranskar ovanstående text med Microsoft Word kommer stavningskontrollen att klaga på det rättstavade ordet "Pagrotsky" och grammatikkontrollen tycker att "runt" inte kongruerar med ordet "meningen" samtidigt som den släpper igenom de problematiska orden "parantes" och "tantkrämen". Förklara för vart och ett av dessa ord varför språkkontrollen gör dessa fel.

10. Morfologi (8 poäng)

Språkrådet har skannat in det avstavningslexikon man gett ut i bokform och tänker lägga upp det som en webbtjänst. Så här ser ett autentiskt utdrag ur den skannade texten ut:

ballad-en/-erna
bal-las-ten/-ter-na
allast-en/-erna
bal-le-ri-nan/-nor-na
ballerin-an/-orna
bal-long-em/-er-na
bal-sa-men/-mer-na
balsam-en/-erna
bal-sa-me-rar/era-de/-rat
balsam-er-ar/-ade/-at

För att så många skanningsfel som möjligt ska kunna upptäckas har man bett Viggo om hjälp. Viggo har tillgång till Stava på sin dator. Kan du råda Viggo hur han automatiskt ska kunna hitta så många fel som möjligt i lexikonet utan att felen dränks av falska alarm? Felen behöver inte rättas automatiskt.