Exempeltentamen för kursen DH2418, Språkteknologi 2007

Tentan består av 10 frågor. Totala poängsumman är 50 poäng. 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. Eventuella bonuspoäng intjänade under 2007 kommer att adderas till dina tentapoäng.

Betygsgränser

För att bli godkänd måste du ha minst 15 poäng på del 1. Betygsgränserna sätts enligt följande: E: 25-29 poäng, D: 30-34 poäng, C: 35-39 poäng, B: 40-44 poäng, A: 45-50 poäng på hela tentan.

Del 1: Teoretisk del (20 poäng)

Skriv korta svar på dessa frågor, max 1 sida.

1. Informationssökning (4 poäng)

a) Vad är term-dokument-matrisen?
b) Vad modellerar man med värdena som står i den?
c) Varför normaliserar man ofta kolumnerna i den?
d) Man normaliserar ofta något annat också. Vad? Vad menas med denna normalisering? Varför gör man den?

2. Analys och generering (4 poäng)

Att analysera text har länge varit en central del inom språkteknologin. Språkgenerering har däremot hamnat i skymundan. Beskriv två tillämpningar och två språkliga nivåer (t.ex. morfologi, syntax, semantik, diskurs, pragmatik) där språkgenerering är nödvändigt om systemet skall vara användbart.

3. Språkteknologins mognad (4 poäng)

En del problem inom språkteknologin är mer eller mindre lösta medan andra problems lösning är avlägsna. Beskriv kortfattat ett problem som kan anses löst till den grad att det finns verkliga program som löser problemet så att det kan anses vara användbart. Beskriv också på ett kortfattat vis ett problem som vore önskvärt att lösa men som inget program idag kan lösa helt automatiskt.

4. Grammatikkontroll (4 poäng)

Inom området automatisk språkgranskning är det en ständig dragkamp mellan att hitta så många fel som möjligt och undvika att signalera korrekta konstruktioner som felaktiga. Beskriv denna problematik utifrån begreppen täckning (recall) och precision.

5. Språkteknologiska metoder (4 poäng)

Det finns i princip två grundläggande angreppssätt inom språkteknologin: regelbaserade eller datadrivna (statistika/maskininlärnings) metoder. Beskriv två fall: ett när det ena är lämpligt och den andra olämpligt och viceversa.

Del 2: Problemdel (30 poäng)

6. Textklustring och textkategorisering (6 poäng)

Ett företag får massor av e-brev till sin kundtjänst. De har sparat alla manuellt i två mappar: en för positiva och en för negativa. Nu vill de ha en automatisk metod som lägger inkommande brev i rätt mapp.

a) Ska de använda klustring eller kategorisering? Varför inte den ena eller den andra? (2 poäng)

b) Förklara i stora drag hur det går till utgående från denna situation. (2 poäng)

c) Vad skulle de kunna ha för nytta av den andra metoden? (2 poäng)

7. Textsammanfattning (4 poäng)

Du ska konstruera en automatisk textsammanfattare. Beskriv kortfattat minst två språkteknologiska verktyg, två statistiska metoder och en heuristisk tumregel (genrespecifik observation) som du skulle vilja inkludera i ditt system, samt vad de skulle vara användbara till i just textsammanfattning. Det står dig fritt att välja om du vill göra textextraktion eller textabstraktion, så länge du beskriver hur du skulle göra det automatiskt med en dator.

8. Utvärdering (6 poäng)

Du har en automatisk textsammanfattare som du vill utvärdera. För enkelhetens skull kan vi anta att du enbart ska utvärdera den på en text, som då följaktligen skall sammanfattas med ditt system. Hur skulle du gå till väga för att utvärdera denna textsammanfattare på denna enda text? Du ska alltså inte beskriva hur den automatiska sammanfattaren ifråga är uppbyggd eller hur den fungerar, bara hur själva utvärderingen av densamma ska gå tillväga. Tänk i termerna guldstandard, nedre gränser (baselines), övre gräns (human ceiling) och utvärderingsmetod(er). Tänk också på att det är svårt även för en människa att sammanfatta en text och hur detta påverkar din utvärdering.

9. Stavningskontroll utan ordlista (8 poäng)

a) Beskriv hur programmet Stava kan kontrollera om ord är felstavade genom att bara använda information om vilka bokstavsfyrgram som används i svenska språket (ordbörjan/ordslut räknas som en egen bokstav). Hur lagras lämpligen tabellen över bokstavsfyrgram? (3 poäng)

b) Om man har tillgång till statistik för hur vanliga de olika bokstavsfyrgrammen är i svensk text så kan man bygga en tredje ordningens markovmodell (inte gömd) för svenska ord. Beskriv hur modellen fungerar och hur den praktiskt kan användas för att ge en bättre stavningskontroll än den i a-uppgiften. (5 poäng)

10. Maskininlärning (6 poäng)

Avdelningen för genusstudier vill ha en maskininlärningslösning för att hitta alla förekomster av ord som syftar på kvinnor respektive män i texter. Dessutom vill man veta vilka som är vad, så man kan färglägga texterna automatiskt och markera kvinnor med rött och män med blått, så det blir lättare att se hur olika de beskrivs i t.ex. nyheter. Hitta på och beskriv en metod för att lösa problemet.

Använd många maskininlärningstermer på ett initierat sätt, t.ex. träningsdata, testdata, optimeringsdata, (un)supervised learning, överinlärning, särdrag, m.m.