Tentamen för DH2418, Språkteknologi
2007-10-26, 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. Eventuella bonuspoäng intjänade under 2007 kommer att adderas till dina tentapoäng. Lämna in del 1 i separat konvolut innan ni 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: E: 25-29 poäng, D: 30-34 poäng, C: 35-39 poäng, B: 40-44 poäng, A: 45-50 poäng.

Del 1: Teoretisk del (20 poäng)

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

1. Textsammanfattning (4 poäng)
Språkteknologiska system använder sig ofta av en kombination av lingvistiska, statistiska och heuristiska (genrespecifika) metoder. Ge ett exempel på vardera i kontexten automatisk textsammanfattning. Ge även ett exempel där en av dina exempelmetoder kompletterar en annan.

2. Ordklasstaggning (4 poäng)
I svenskan finns det oändligt många tänkbara sammansatta ord, vilket betyder att inget lexikon innehåller alla. Hur kan en ordklasstaggare gissa vilken tagg ett okänt sammansatt ord (dvs ett ord i meningen som inte är med i lexikonet) ska ha? Föreslå minst två olika och väl fungerande metoder!

3. Informationssökning (4 poäng)
I informationssökning (IR) används den så kallade vektorrumsmodellen.

a) I den representeras en textmängd av term-dokument-matrisen. Hur tolkar man den?
Vad innehåller den för värden? (2 poäng)

b) Hur beräknas likhet mellan två texter i vektorrumsmodellen? (2 poäng)

4. Morfologi (4 poäng)
Beskriv metoderna stemming och lemmatisering, ge exempel på deras effekt på något lämpligt ord samt i valfritt system.

5. Språkstatistik (4 poäng)
Vilken typ av ord hittar man i den översta toppen av s.k. frekvenslistor som framställs från stora balanserade korpusar. Ge exempel. Vad kan man använda frekvenslistor till?

Vänd!

Del 2: Problemdel (30 poäng)

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

6. Statistiska metoder (6 poäng)

Språkrådet (tidigare Svenska språknämnden) publicerar varje år en nyordlista som innehåller ord som börjat användas i svenska språket under det senaste året. Några exempel från senaste nyordlistan är bloggosfären, minnespinne, rondellhund, latteliberal, legga. Språkrådet sätter ihop listan genom att aktivt läsa tidningar och försöka uppmärksamma nya ord. Naturligtvis borde detta gå att automatisera med hjälp av statistiska metoder i språkteknologin. Föreslå hur man ska hitta förslag till nyord med hjälp av korpusar och metoder från kursen.

7. Utvärdering (6 poäng)

Du har ett språkteknologiskt system som bland annat innehåller en ordklasstaggningskomponent.

a) Varför är det viktigt att inte bara utvärdera hela systemet, utan även utvärdera de enskilda komponenterna (delsystemen) för sig? (2 poäng)

b) Hur skulle du gå till väga för att utvärdera taggaren? Resonera i termerna guldstandard, nedre gränser (baselines) och övre gräns (human ceiling) när du motiverar ditt val av utvärderingsmetod. (4 poäng)

8. Textkategorisering/klustring (8 poäng)

Snabba Cash AB har ett invecklat dokumenthanteringssystem. I det ingår att alla nyproducerade texter taggas med en del metadata. Ett exempel är texttyp, som kan vara pressrelease, bruksanvisning, kvartalsrapport, order, teknisk specifikation, etc. Dokumenten är skrivna på svenska.

Texter som skrevs innan man införde systemet har inte den informationen. Du har anställts för att bygga ett system som automatiskt lägger till texttypen på de äldre texterna. Hur går du tillväga? Beskriv och resonera kring textmängder, representation, algoritm och utvärdering av hela metoden.

9. Maskininlärning och skrivverktyg (5 poäng)

Företagsledningen för Snabba Cash AB är nöjd med språket (svenska) i de dokument som har producerats, men man har fått många klagomål från nya medarbetare om att det är svårt att anpassa sig till jargongen för respektive texttyp. Man vill därför att du konstruerar ett system som ger skribenter förslag på lämpliga synonymer för den aktuella texttypen. Hur konstruerar du ett sådant system? Vad använder du för metod och hur fungerar den? Vilka texter använder du? Tänk på att texterna inom en texttyp kanske inte innehåller de ord som en ny och ovan skribent använder.

10. Grammatikkontroll (5 poäng)

Svenska staten har bestämt att man skall satsa på de olika minoritetsspråken i Sverige. Ditt uppdrag är att redovisa två olika scenarier som i stort beror på hur mycket pengar som satsas på utveckling av en grammatikkontroll för respektive minoritetsspråk (en begränsad satsning och större mer långsiktig satsning). Du skall redovisa lämpliga metoder för varje satsning (men för ett språk), och vilka resultat som staten kan förvänta sig av respektive satsning. I ditt fall handlar det om språket sydsamiska som har mycket begränsade språkliga resurser (korpusar och lexikon) till sitt förfogande i nuläget. Grundläggande språkteknologiska verktyg saknas också. Vilka scenarier skulle du vilja redovisa? Vilka resurser och verktyg skulle du satsa på? Kom ihåg att beskriva de två olika fallen.