

Tentamen för DH2418, Språkteknologi, 2010-10-21, kl. 9-13

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 2010 kommer att adderas till dina tentapoäng för respektive del. 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 **13 poäng på del 1 inkl. teoretiska bonuspoäng, och minst 12 poäng på del 2 inkl. praktiska bonuspoäng**. Betygsgränserna är enligt följande på hela tentan:

A: 45 poäng och uppåt, 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, en riktlinje är max två sidor för hela del 1.

1. Flertydighet (4 poäng)

Följande mening är flertydig.

Hon såg inte bara män utan röda kläder på festen.

På vilka sätt? Argumentera utifrån en språklig analys.

2. Maskininlärning (4 poäng)

a) Vad menas med övervakad respektive oövervakad maskininlärning?

b) Ge exempel på en språklig träningsmängd för respektive metodik.

3. Informationsteori (4 poäng)

Definiera begreppet perplexitet och ge ett exempel på hur perplexiteten kan användas vid utvecklingen av något språkteknologiskt system.

4. Ordklasstagging (4 poäng)

En ordklasstaggar som bygger på Transformationsbaserad inläring (TBL) har följande information från en korpus:

$$P(\text{VB}|\text{sticka}) = .82$$

$$P(\text{NN}|\text{sticka}) = .18$$

Hur kan en TBL-taggar få rätt tolkning av ”sticka” i meningen ”Han fick en sticka i fingret.”? Beskriv de olika stegen.

5. Morfologi (4 poäng)

Vad är det för skillnad mellan stemming och lemmatisering? Vilka fördelar respektive nackdelar finns det med metoderna?

Del 2: Problemdel (30 poäng)

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

6. Stavningskontroll (6 poäng)

Det finns metoder för stavningskontroll som inte använder någon ordlista. Ett sätt är att använda en tredje ordningens markovmodell med hjälp av statistik för bokstavsfygram. Beskriv hur man kan gå tillväga för att bygga en engelsk stavningskontroll med en sådan modell.

7. Ordprediktion (6 poäng)

Sture arbetar på ett stort mjukvaruföretag där han utvecklar en trigrammodell för ordprediktion. När Sture studerar modellen upptäcker han att sannolikheten för en mängd rimliga trigram är noll. Varför är det så? På vilket sätt är detta ett problem? Beskriv en metod för att komma tillrätta med problemet.

8. Informationssökning (6 poäng)

Företaget Search'R'Us har implementerat sin sökmotor på följande sätt: För att jämföra en sökfråga Q med ett dokument D räknar man antal förekomster av termer i D som också förekommer i Q . Varje sådan förekomst ger D ett poäng. Som svar på en sökfråga returneras länkar till de dokument som har fått poäng större än 0 (i fallande poängordning). Kan du tipsa Search'R'Us om hur de kan förbättra sin algoritm?

9. Utvärdering och språkgranskning (6 poäng)

Eva arbetar på ett läromedelsförlag, och hennes uppdrag är att ansvara för upphandlingen av språkteknologiska system. Just nu skall förlaget köpa in en grammatikkontroll som skall användas för att undvika språkfel i nya läromedel. Hon har sett sig omkring efter leverantörer, och det visar sig finnas fem olika grammatikkontroller för svenska. Hur skall hon bära sig åt för att ta reda på vilken som är bäst? Hon behöver dessutom tips om hur hon kan göra arbetet så effektivt som möjligt.

10. Dialogsystem (6 poäng)

Stora Filmbolaget (SF) vill skapa en talstyrd tjänst där kunder kan ringa och reservera biobiljetter. För att kunna göra en reservation måste SF veta önskad film, biograf, föreställning, datum och antal biljetter. Om det finns biljetter kvar ska en reservationskod läsas upp. Förklara hur du skulle designa en sådan tjänst med avseende på taligenkänning och dialoghantering. Hur kan man utvärdera systemet när det väl har driftsatts?