729G17 **Språkteknologi**

Tentamen

Marco Kuhlmann, Institutionen för datavetenskap, Linköpings universitet
marco.kuhlmann@liu.se

22 augusti 2014

Inga hjälpmedel är tillåtna. Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentamen är 32; 16 poäng ger säkert godkänt. Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla (10) frågor i denna del. Varje fråga ger 2 poäng.

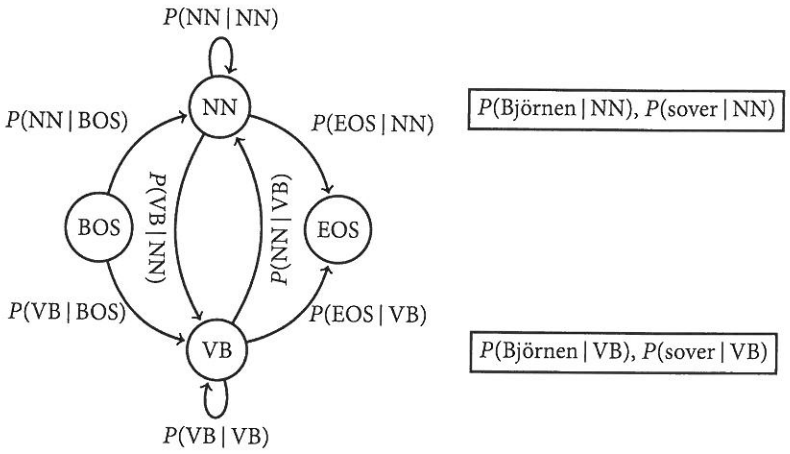
1. En texts *type–token ratio* är antalet *types* i texten delat med antalet *tokens*. (a) Gertrude Steins dikt *Sacred Emily* är känd för meningen "Rose is a rose is a rose is a rose." Räkna ut type–token ratio på denna mening. (b) Hur resonerar man när man använder type–token ratio för att skilja mellan svårlästa och lättlästa texter?

2. I en svensk korpus finner vi följande frekvenser för några utvalda ord och ordsekvenser:

   med: 10 000; *tanke*: 120; *på*: 8 000; *med tanke*: 100; *tanke på*: 90; *med tanke på*: 80

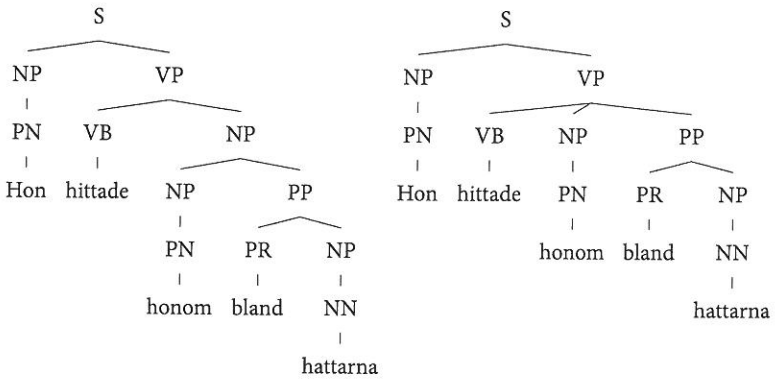
   Vad är den Maximum Likelihood-uppskattade sannolikheten $P(\text{på} \mid \text{med tanke})$ om vi använder (a) trigramsannolikheter, (b) en omskrivning till bigramsannolikheter?

3. Följande bild illustrerar en Hidden Markov-modell för ordklasstaggning där det endast finns två ord, *Björnen* och *sover*, och två möjliga taggar, NN och VB.



(a) Förklara vad de olika sannolikheterna i denna modell betyder. (b) Vilka konkreta sannolikheter måste man ha skattat för att i modellen kunna räkna ut den kombinerade sannolikheten för följande taggade mening: *Björnen/NN sover/VB*

4. Nedanstående visas två frasstrukturträd för meningen *Hon hittade honom bland hattarna.* Extrahera en probabilistisk kontextfri grammatik utifrån dessa träd. Uppskatta regelsannolikheterna med hjälp av Maximum Likelihood Estimation. Ange bara de regler som expanderar NP eller VP.



5. En enkel metod för att bestämma en ords betydelse är algoritmen Simplified Lesk, som använder sig av lexikon. Följande betydelser hittar man när man slår upp ordet *papper* i svenska Wiktionary:

   1 material, primärt format i tunna ark avsedda för skrift, vanligen tillverkat av växtfibrer. 2 en bit av ovanstående material, i allmänhet använt för att skriva eller rita på. 3 dokument, handling. "Om man ska åka utomlands bör man se till att inte glömma sina papper." 4 artikel publicerad i en vetenskaplig tidskrift.

   Välj ut en av betydelserna. Formulera sedan två meningar som innehåller ordet *papper* i den utvalda betydelsen: (a) en mening där Simplified Lesk räknar ut den avsedda betydelsen, (b) en mening där Simplified Lesk räknar ut fel betydelse.

6. Följande tabell visar resultaten av en utvärdering av ett automatiskt system (s) för attitydanalys på en guldstandard (g). Räkna ut systemets precision och recall med avseende på uppgiften att identifiera dokument med polaritet neg (negativt).

g \ s	pos	neu	neg
pos	72	15	3
neu	20	32	9
neg	8	9	69

7. En enkel metod för att tagga filmrecensioner med polariteter är att använda en Naive Bayes-klassificerare. Ange en formel som beskriver klassificerarens beslutsregel och förklara alla delar i denna formel.

8. Entitetsextraktion (eng. *named entity recognition*) går ut på att identifiera och klassificera ord eller andra textenheter som representerar semantiska enheter såsom personer, organisationer och platser. (a) Förklara varför entitesextraktion är viktig. (b) Ge exempel på två metoder som kan användas för entitetsextraktion.

9. För att utvärdera maskinöversättningssystem används måttet BLEU som bygger på ett modifierat precisionsmått på n-gram. (a) Räkna ut det modifierade precisionsvärdet på bigram för följande exempel. (b) Förklara varför BLEU inte använder det vanliga precisionsmåttet på n-gram.

Systemets översättning:	the	cat	the	cat	the	cat	the
Referensöversättning 1:	the	cat	is	on	the	mat	
Referensöversättning 2:	there	is	a	cat	on	the	mat

10. Olika frågor är olika svåra för automatiska frågebesvarande system att hitta rätt svar på. Rangordna följande tre frågor från lättast till svårast för ett frågebesvarande system och motivera rangordningen. Systemet antas inte ha någon egen kunskapsbas utan använda sig av textdokument som dataresurs.

    A. Vilket år föddes Einstein? *B. Hur var han som människa?* *C. Var föddes Einsteins första fru?*

Del B

Välj två (2) frågor och besvara dem utförligt. Varje fråga kan ge maximalt 6 poäng.

11. En probabilistisk parser ska räkna ut den mest sannolika syntaktiska analysen för en given mening. Förklara varför denna uppgift är beräkningsmässigt utmanande. Beskriv därefter två metoder för att bemöta denna utmaning.

12. Beskriv någon vanlig arkitektur för ett frågebesvarande system. Ange vilka tekniker som kan användas i de olika komponenterna.

13. I flera typer av system kan recall inte mätas på det vanligaste sättet, dvs. genom att dela antalet fall där systemet och guldstandard överensstämmer med det totala antalet fall i guldstandarden. Ange två applikationer där detta inte fungerar så bra, förklara varför, och beskriv något eller några mått som används i stället.

14. Ett beslutsstödsystem är ett datorsystem som hjälper användaren att fatta beslut genom att sammanställa, analysera och presentera information om tidigare, liknande situationer. Sådana system är intressanta t.ex. inom sjukvården där de kan hjälpa en läkare att ställa diagnoser. Ge exempel på komponenter i ett beslutsstödsystem som skulle kunna använda sig av språkteknologi. Beskriv därefter några av de specifika utmaningar som språkteknologin ställs inför när den ska integreras i beslutsstödsystem.