729G17 **Språkteknologi**

Tentamen

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

25 april 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 frågor i denna del. Varje fråga ger 2 poäng.

1. I kursboken (Jurafsky och Martin, 2009) på sida 86 står det:

   Shakespeare's complete works have 29,066 word form types (from 884,647 word form tokens).

   (a) Förklara skillnaden mellan begreppen *type* och *token*. (b) Ange exempel på situationer där man är intresserad i ord som *types* och situationer där man är intresserad i ord som *tokens*.

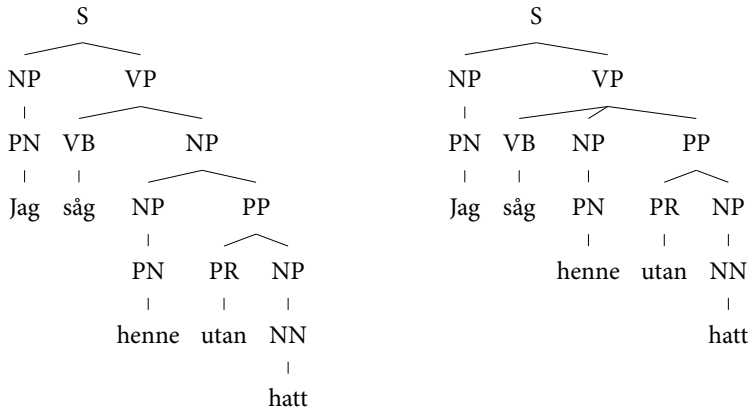
2. I en svensk korpus finner vi följande frekvenser för några utvalda ord och ordsekvenser:

   på: 10 000; *grund*: 120; *av*: 8 000; *på grund*: 100; *grund av*: 90; *på grund av*: 80

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

3. (a) Vilka typer av sannolikheter ingår i en Hidden Markov-modell för ordklasstaggning? (b) Vilka konkreta sannolikheter måste man ha skattat för att i en sådan modell kunna räkna ut den kombinerade sannolikheten för följande taggade mening: *Björnen/NN sover/VB*

4. Nedanstående visas två frasstrukturträd för meningen *Jag såg henne utan hatt.* Extrahera en probabilistisk kontextfri grammatik utifrån dessa träd. Uppskatta regelsannolikheterna med hjälp av Maximum Likelihood Estimation. Du behöver bara ange de regler som har NP eller VP som vänsterled.



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 *kurs* i svenska Wiktionary:

   1 *riktning. "Kaptenen satte kurs mot hemmahamnen."* ***2*** *avgränsat moment inom en utbildning. "Hon tog kurser i latin och astronomi för att bättra på sin allmänbildning."* ***3*** *pris för vara eller värdepapper vid en given tidpunkt. "Kursen för de flesta IT-aktier störtdök under eftermiddagen."*

   Välj ut en av betydelserna och formulera två meningar som innehåller ordet *kurs* 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 pos (positivt).

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 klassificerarens beslutsregel och förklara den.

8. (a) Vad innebär entitetsextraktion (eng. *named entity recognition*)? (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 för följande exempel. (b) Förklara varför BLEU inte använder det vanliga precisionsmåttet på n-gram.

Systemets översättning:	the	the	the	the	the	the	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 använder sig av textdokument som dataresurs.

A. Varför sjönk Titanic? *B. Vilket år sjönk Titanic?* *C. När började Titanic spelas in?*

Del B

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

11. 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.

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

13. 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.

14. Du jobbar på ett företag som har utvecklat ett automatiskt system som kan klassificera filmrecensioner som antingen positiva eller negativa. Nu har ni blivit kontaktade av ett politiskt parti som vill använda ert system för att få fram attityder gentemot partiet som de reflekteras i Twitterinlägg. Ange skillnader mellan uppgifterna ”klassificera produktrecensioner” och ”analysera attityder i Twitterinlägg”. Beskriv sedan några konkreta tekniska problem som ni skulle behöva lösa för att anpassa ert system till den nya uppgiften.