Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2013-03-26
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER3
Tid	14-18
Kurskod	729G17
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Språkteknologi Tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	14 (varav 4 valbara)
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	013-282422
Besöker salen ca kl.	14.45-15.00
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	annelie.almquist@liu.se, 2934
Tillåtna hjälpmedel	inga
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	valfritt
Antal exemplar i påsen	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

729G17 Språkteknologi

tisdag 26 mars 2013 kl 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 34.
17 poäng ger säkert godkänt.

Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del.

1. Vad menas med att normalisera en text och varför gör man det? (2p)

2. Ett namnigenkänningssystem uppgavs ha en precision på 92% när det testats på en text innehållande 200 namnförekomster. Går det med ledning av dessa uppgifter att avgöra hur många namn systemet hittade? Motivera svaret. (2p)

3. (a) Skriv ett reguljärt uttryck som exakt definierar mängden av teckensekvenser {hål, håll, hålla, hållas}, utan att använda disjunktionsoperatorn '|'. (b) Ange ett bokstavsträd, dvs en tillståndsautomat (eng. finite state automaton, FSA), som definierar precis samma mängd. (2p)

4. Ange de tre viktigaste dataresurserna i ett transformationsbaserat system för ordklasstaggning (en s.k. Brilltaggare). (2p)

5. I statistiska språkmodeller har vi sannolikheter för händelser av typen 'ordet *en* kommer efter orden *det är*'. (a) Ange ett uttryck för denna sannolikhet om vi använder en modell baserad på trigramsannolikheter, (b) Ange ett uttryck för samma sannolikhet om den i stället baseras på unigramsannolikheter. (2p)

6. Vad är Levenshteinavståndet mellan orden *sträcka* och *styrka?* Motivera svaret. (2p)

7. Automatisk syntaktisk analys av meningar kan göras t.ex. genom chunkning eller parsning. Förklara skillnaden och ange också för vardera metoden, med motivering, någon tillämpning där den är mer lämplig än den andra att använda. (3p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

8. Beskriv den modell som kallas den brusiga kanalen (eng. *the noisy channel*) och var noga med att förklara dess s.k. beslutsregel. (2p)

9. System för betydelsebestämning av ord representerar ofta olika betydelser av ett ord med en kontextvektor. Förklara vilken information en sådan kontextvektor kan innehålla och beskriv någon metod baserad på kontextvektorer för att bestämma ett ords betydelse i löpande text. (3p)

10. Koreferenslösning innebär bland annat att hitta antecedenter till pronomen som han, hon, den eller det. För en viss förekomst av pronomenet *den* finns tre substantiv med n-genus som är möjliga antecedenter (beskrivna nedan). Avgör vilken eller vilka av dem som kan vara antecedent. Motivera valet. (2p)

    Ord1 är ett substantiv i bestämd form som står som subjekt i föregående mening

    Ord2 är ett substantiv i singularis obestämd form som står före pronomenet i samma mening men inte i samma sats

    Ord3 är ett substantiv i singularis bestämd form som står före pronomenet och är subjekt i samma sats.

Del B

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

11. Förväxlingsmatriser (eng. *confusion matrices*) är användbara verktyg i samband med utveckling av språkteknologiska system. Förklara vad en förväxlingsmatris är, hur värdena i den används för att uppskatta överensstämmelse och hur de kan användas i samband med ordklasstaggning.

12. IBM har visat att ett datorsystem kan besegra även en mästare i Jeopardy, men hittills har ingen lyckats utveckla ett system som kan översätta bättre än en mänsklig översättare. Diskutera skillnader mellan aktiviteterna 'spela Jeopardy' och 'översättning' som kan förklara detta faktum.

13. Beskriv någon vanlig arkitektur för ett frågebesvarande system med angivande av vilka tekniker som används i de olika komponenterna.

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