Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2013-06-03
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	13
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	013-282422
Besöker salen ca kl.	14.45
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	annelie.almquist@liu.se, tel 282934
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

måndag 3 juni 2013 kl. 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 32. 16 poäng ger säkert godkänt. Planerad VG-gräns: 24.

Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del.

1. Vad skiljer lemmatisering från stemming? (2p)

2. (a) Vad menas med att en ändlig tillståndsautomat (eng. finite-state automaton) är deterministisk? (b) Ange en sådan automat som genererar samma strängmängd som det reguljära uttrycket 'b(a|c)b*' (2p)

3. I en korpus hittades 120 ord som slutar på 'bar', bl.a. *snobbar* och *sökbar*. Av dessa befanns 40 vara substantiv, 65 vara adjektiv och 15 vara verb. (a) Ange MLE-uppskattningen med användningen av denna korpus för sannolikheten att ett ord som slutar på 'bar' är ett substantiv; (b) Hur kan vi ändra denna uppskattning om vi vill ge utrymme för möjligheten att egennamn också skulle kunna sluta på 'bar', trots att inga sådana egennamn setts i korpusen? Svara med en beskrivning eller en lämplig omräkning. (3p)

4. Ett namnigenkänningsprogram kritiserades för att generera för många 'falska positiva'. (a) Vad betyder det? (b) Vilket eller vilka av måtten precision eller recall påverkas av antalet falska positiva? (2p)

5. Vid ordprediktion är man intresserad av sannolikheten $p(w_i| w_1 w_2 \ldots w_{i-1})$. (a) Vad innebär det att använda en bigrammodell för denna sannolikhet? (b) Vilket är det vanligaste måttet för att mäta prestanda för en ordprediktionsmodell? (2p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

6. Modellen den brusiga kanalen (eng. *Noisy channel*) kan t.ex. tillämpas på problemet översättning med beslutsregeln "Givet en mening F på ett främmande språk, välj den engelska mening E* som har störst sannolikhet att ge upphov till F". Formulera beslutsregln matematiskt och illustrera modellen grafiskt. (3p)

7. (a) Förklara begreppen anafor och antecedent med exempel. (b) Varför är det viktigt att kunna känna igen anaforer och antecedenter exempelvis i en tillämpning som informationsextraktion? (2p)

8. Vad innebär extraktionsmodellen för textsammanfattning och vad menas med en sammanfattnings kompressionsgrad? (2p)

9. I en chart-parser som gavs meningen *vanlig mjölk är godast* som indata fanns i ett visst läge bland annat följande tillstånd/bågar:

   $S \rightarrow . \ NP \ VP$ [0,0]

   $NP \rightarrow A \ . \ N$ [0,1]

   Ange två andra tillstånd/bågar som också måste finnas i charten. Motivera ditt svar. (2p)

Del B

Välj två av frågorna 10-13 och besvara dem utförligt. Varje fråga kan ge maximalt 6 poäng.

10. Beskriv problemet ordbetydelsebestämning (eng. *word sense disambiguation*) och någon metod som har föreslagits för att lösa det. Förklara också varför metoder som fungerar bra för ordklasstaggning inte fungerar lika bra för ordbetydelsebestämning.

11. Relationer, som den mellan en vara och ett pris kan uttryckas på många sätt i naturligt språk, t.ex med uttryck som X *kostar Y, jag vill ha Y för X, priset för X är Y.* Beskriv en metod att hitta en stor mängd sådana relationsangivande uttryck i en större textkorpus.

12. Beskriv vad som menas med transformationsbaserad inlärning (TBL) och hur det kan tillämpas på problemet chunkning. Ange förutsättningar i form av resurser och data och hur träningen går till.

13. Språkrådet brukar mot slutet av varje år presentera en lista med s.k. nyord, dvs innehållsord som blivit populära under året eller fått nya betydelser. Skissa på ett system med angivande av arkitektur och nödvändiga resurser som kan generera, så precist som möjligt, nyordskandidater för ett givet år, t.ex. 2012.