Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2011-08-19
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER1
Tid	14-18
Kurskod	TDDD01
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Språkteknologi Skriftlig tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	11
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	ankn. 2422, 0703-18 24 22
Besöker salen ca kl.	ca. kl 14:45
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Anna Grabska Eklund, ankn. 2362, anna.grabska.eklund@liu.se
Tillåtna hjälpmedel	inga
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	
Antal exemplar i påsen	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD01 Språkteknologi

fredag 19 augusti 2011 kl 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 26.
13 poäng ger säkert godkänt.

Besvara alla frågor. Det går bra att besvara flera frågor på samma papper.

1. Vad menas med att (a) tokenisera, (b) lemmatisera en text? (2p)

2. Ett namnigenkänningssystem uppgavs ha en precision på 85% när det testats på en dokumentsamling innehållande 500 namnförekomster. Går det med ledning av dessa uppgifter att avgöra hur många namn systemet hittade? Svaret måste motiveras. (2p)

3. Transduktorer (eng. finite-state transducers) har användning inom språkteknologin t.ex. för att lagra lexikon och få ord i löpande text morfologiskt analyserade. Ange minst två fördelar med denna typ av representation, jämfört med att ha orden med deras analyser lagrade i en textfil. (2p)

4. Förklara (a) generellt vad som menas med en språkmodell, (b) specifikt att en sådan modell är trigrambaserad och använder sig av back-off. (3p)

5. Förklara vad som menas med redigeringsavstånd och beskriv någon algoritm för att beräkna redigeringsavståndet mellan två ord. Tillämpa algoritmen på exempelorden *pasta* och *potatis*. (3p)

6. En grundläggande modell för statistisk maskinöversättning är den brusiga kanalen (eng. The Noisy Channel). Beskriv denna modell och hur den är tänkt att ge oss en lösning för den bästa översättningen. (3p)

7. Automatisk syntaktisk analys av meningar kan göras t.ex. genom chunkning eler parsning. Förklara skillnaden och ange också för båda metoderna någon tillämpning där de är lämpliga att använda. (2p)

8. I en top-down chart-parser drivs parsningen framåt bland annat via prediceringar av hypoteser. (a) Vilken är den initiala prediktionen? (b) Hur prediceras tillstånd (eller bågar) därefter? Du kan ställa upp en allmän regel eller förklara i ord. (2p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post: lah@ida.liu.se

9. En central modul i ett frågebesvarande system är frågeanalysatorn som bestämmer svarstyp. (a) Förklara vad som menas med en svarstyp och ge några exempel. (b) Ge något exempel på att det inte räcker med att känna igen frågeord i en fråga för att kunna bestämma svarstyp. (2p)

10. I system för informationsextraktion ingår ofta en modul för koreferensbestämning av nominalfraser. Beskriv kortfattat vilken typ av information man använder för att avgöra om det föreligger koreferens mellan en nominalfras NP1 och en tidigare nominalfras NP2 om (a) NP1 består av ett efternamn, (b) NP1 består av ett pronomen som *han* eller *hon*. (2p)

11. 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)