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	729G17
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Språkteknologi Tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	10
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	ankn. 2422, 0703-18 24 22
Besöker salen ca kl.	ca. 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

729G17 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 32.
16 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 (a) tokenisera, (b) normalisera 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. Ange de tre viktigaste dataresurserna i ett transformationsbaserat system för ordklasstaggning (en s.k. Brilltaggare). (2p)

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

6. Vilket är redigeringsavståndet mellan orden *pasta* och *potatis?* Var noga med att ange vilken definition du använder. (2p)

7. Automatisk syntaktisk analys av meningar kan göras t.ex. genom chunkning eler parsning. Förklara skillnaden och ange också för vardera metoden, med motivering, någon tillämpning där de är mer lämplig än den andra att använda. (2p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

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)

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. (2p)

10. Vad är WordNet? (1p)

Del B

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

10. Beskriv den modell som kallas för den brusiga kanalen (eng. *Noisy Channel*). Ange och förklara dess beslutsregel och hur den kan omformas med hjälp av Bayes regel. Visa hur modellen tillämpas på översättning och någon annan språkteknologisk tillämpning.

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. Ett moment i informationsextraktionssystem är koreferenslösning av nominalfraser. Förklara vad detta problem går ut på och beskriv ett antal typfall av koreferens mellan nominalfraser som egennamn, pronomen, bestämda beskrivningar och deras antecedenter.

13. I svenska och många andra språk förekommer lexikaliserade fraser (eller idiom) i form av samordningar. Exempel är *hit och dit, smått och gott, i vått och torrt*. Föreslå utformning av ett system som dels extraherar sådana samordningar, dels klassificerar dem som idiomatiska eller ej.

14. Ett berömt citat av Noam Chomsky, även citerat i kursboken, är "But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term". Chomsky avsåg nog i första hand användning för teoretisk lingvistisk forskning när han skrev detta, men då sågs också språkteknologi närmast som beroende av lingvistisk forskning för sin utveckling. Diskutera påståendet mot bakgrund av de sannolikhetsmodeller som kursen tagit upp, antingen snävt i relation till språkteknologi, eller mer generellt för förståelsen av mänsklig språkförmåga.