Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2011-03-14
Sal	TER4
Tid	14-18
Kurskod	729G17
Provkod	TEN1
Kursnamn/benämning	Språkteknologi
Institution	IDA
Antal uppgifter som ingår i tentamen	*15 (varav 5 icke-obligatoriska)*
Antal sidor på tentamen (inkl. försättsbladet)	*4*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*013-282422*
Besöker salen ca kl.	*14.40*
Kursadministratör (namn + tfnnr + mailadress)	*Helene Meisinger 281868, helene.meisinger@liu.se*
Tillåtna hjälpmedel	*Inga*
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
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 14 mars 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.

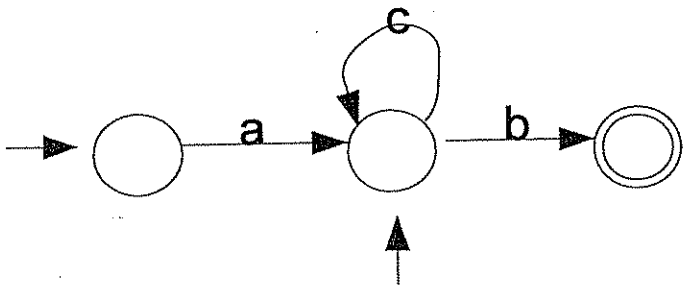
Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del.

1. Förklara vad som menas med att lemmatisera en text och ange något syfte, eller någon tillämpning, där det är motiverat att man gör det. (2p)

2. Den finita automaten här definierar ett språk över alfabetet {a,b,c}. Skriv automatens språk med ett reguljärt uttryck. (1p)



3. Många adjektiv i svenska som slutar på -a kan också vara substantiv (*glada, visa, raka, ...*) eller verb (*visa, viga, räta, ...*). Ange kontextvillkor av den typ som kan användas i system för ordklasstaggning som entydigt eller nästan entydigt visar om en viss förekomst av ett sådant ord är (a) ett substantiv, (b) ett verb, (c) ett adjektiv. (Du ska alltså ange ett villkor per delfråga.) (3p)

4. Definiera måtten recall och precision och förklara varför båda måtten behövs. (3p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

5. Vad menas med redigeringsavstånd (Minimal edit distance, eller Levenshtein distance)? Beskriv kortfattat hur Levenshteinavståndet definieras och avgör vilket eller vilka av följande ord teckensekvensen *svältkost* har kortast avstånd till: *sveciaost*, *frukost* eller *svältkatastrof.* (2p)

6. I en korpus hittades följande frekvenser för några olika ordsekvenser:

många	100
inte	200
är	400
inte många	10
inte är	20
inte många är	2
inte är många	0
många är	80
är många	20

(a) Beräkna Maximum Likelihood-uppskattningen av sannolikheten för att 'många' kommer efter 'är' dvs p(många | är) utifrån denna korpus. (b) Beskriv något sätt att uppskatta trigramsannolikheten p(många | inte är) till något annat än 0 givet dessa data. (Du behöver inte ge ett värde, det räcker med en beskrivning.) (2p)

7. Anta att vi har givet en bigrambaserad språkmodell. Vilken av följande två ordsekvenser ges störst sannolikhet av en sådan modell: *Sverige kämpade hårt - Sverige kämpade hårt men förlorade* Svaret ska motiveras. (1p)

8. En top-down chart-parser drivs framåt genom successiva prediceringar av hypoteser och matchningar av hypoteserna mot orden i indata och tidigare verifierade hypoteser. Låt grammatiken G vara given enligt nedan, där S står för kategorin meningar. (a) Vilken är den initiala hypotesen och hur representeras den som en punkterad regel? (b) Vilka prediceringar genereras utifrån denna initiala hypotes och reglerna i G oberoende av indata? (2p)

$S \rightarrow NP\ VP$

$NP \rightarrow DET\ N$

$NP \rightarrow PN$

$VP \rightarrow IV$

$VP \rightarrow TV\ NP$

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 ordbetydelsebestämning representeras kontexten för ett ord som skall betydelsebestämmas ofta med en s.k. kontextvektor. Förklara vad en kontextvektor är och hur den används för att bestämma betydelse. Du kan illustrera med ordet *panna* i meningen nedan och anta att ordet har tre olika betydelser (ansiktsdel, kärl eller drivanordning).

Han hade glömt pannan på spisen och det gjorde honom riktigt bekymrad. (2p)

Del B

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

11. Beskriv den modell som kallas för den brusiga kanalen (eng. *Noisy Channel*) inklusive beslutsregel och härledning av modellens parametrar (sannolikhetsfördelningar). Diskutera hur modellen tillämpas på översättning.

12. Utvärdering av system som ska producera texter, t.ex. översättningssystem eller sammanfattningssystem baseras ibland på ngram-jämförelser mellan systemets utdata och en eller flera referensöversättningar. Välj något sådant mått, t.ex. BLEU eller ROUGE och förklara hur man beräknar det. Diskutera för- och nackdelar med den här typen av mått.

13. Ett moment i informationsextraktionssystem är referenslösning av pronomen. Förklara vad detta problem går ut på och beskriv någon metod att hitta antecedenter till givna pronomen. Du kan begränsa uppgiften till pronomina *hon* och *han*.

14. Ett sätt att identifiera s.k. chunkar, till exempel icke-rekursiva nominalfraser i löpande text baseras på taggar som markerar positionen för ett ord i frasen. Beskriv hur detta går till och vilka resurser som behövs för att bygga ett sådant system.

15. Systemet Watson har visat sig kunna besegra människor i frågeleken Jeopardy. Argumentera utifrån detta faktum för eller emot påståendet "Watson gör sannolikt att datorsystem i framtiden kommer att kunna delta i samtal på samma nivå som en människa."