



Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2011-06-07
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och <u>ringa in</u> vilken sal som avses	TER2
Tid	14-18
Kurskod	729G17
Provkod	TEN1
Kursnamn/benämning Provnamn/benämning	Språkteknologi Tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	14
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	ankn. 24 22, 070-31 82 422
Besöker salen ca kl.	
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	valfritt
Antal exemplar i påsen	

TENTAMEN

729G17 Språkteknologi
tisdag 7 juni 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. Skriv ett reguljärt uttryck som identifierar fyrsiffriga tal, dvs naturliga tal i intervallet 1000-9999. Skriv också ett reguljärt uttryck som begränsar träffarna till tal i intervallet 1000-2499. (2p)
2. (a) Vid språkteknologiskt korpusarbete tokeniserar man helst texter så att skiljetecken som komma, punkt etc. utgör egna tokens. Varför gör man det? (b) Varför resonerar man annorlunda när det gäller förkortningspunkter? (2p)
3. I förväxlingsmatrisen nedan redovisas utdata från ett morfologiskt taggningssystem i form av absoluta frekvenser för korrekta och felaktiga kategoriseringar av ett antal verbformer i svenska. Rader anger systemets kategoriseringar, kolumner kategorierna i facit. Siffran 36 i raden Imperativ anger alltså att i 36 fall taggade systemet infinitivformer som imperativ. (a) Vad är systemets precision för infinitiva verbformer? (b) Vilken verbkategori har systemet svårast att tagga rätt? Motivera svaret! (2p)

	Infinitiv	Imperativ	Preteritum	Perf. particip	Andra ordklasser
Infinitiv	720	5	6	0	69
Imperativ	36	45	4	1	4
Preteritum	0	2	804	44	10
Perf. particip	0	1	20	100	18
Andra ordklasser	50	8	40	15	4200

4. Ange Levenshteinavståndet mellan strängarna *krot* och *kort* och visa hur man räknar fram det algoritmiskt, t.ex. genom att ställa upp en matris för avstånden mellan delsträngar. (2p)
5. Förklara skillnaden mellan ett fullformslexikon och ett morfembaserat lexikon och ange minst två fördelar med morfembaserade lexikon. (2p)
6. (a) Förklara vad som menas med en språkmodell (eng. *language model*). (b) Vad innebär det att en språkmodell är baserad på trigram? (c) Nämn någon metod att i en trigrammodell hantera treordsekvenser som aldrig setts i träningsdata. (3p)
7. I modellen 'den brusiga kanalen' (eng. Noisy Channel) ser formeln för att välja mest sannolika sträng ut så här:
- $$S^* = \underset{S}{\operatorname{argmax}} p(S|O)$$
- (a) Förklara formeln i ord. (b) Ofta skriver man om sannolikheten $p(S|O)$ med hjälp av Bayes regel. Visa resultatet av omskrivningen och förklara varför man gör den. (3p)
8. System för partiell parsning följer ofta en s.k. kaskadmodell. Vad innebär det? Ange också minst fyra olika moduler som vanligen ingår i ett sådant parsningssystem. (2p)
9. Textsammanfattningssystem baseras ofta på extraktion av en delmängd av de meningar som ingår i texten. Ange tre egenskaper hos meningar som gör att de kan komma i fråga för att ingå i sammanfattningen. (2p)

Del B

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

10. Prestandamått som precision och recall är enkla att beräkna givet att det bara finns en möjlig kategori för varje enhet som systemet värderar och ett facit. I fallet översättning finns dock många översättningar som kan anses vara korrekta för en given mening, och i ännu högre grad för en hel text. Diskutera hur mått som precision och recall kan tillämpas i detta fall och hur man kan gå till väga för att beräkna dem med en viss grad av tillförlitlighet.
11. Beskriv hur topdown-chartparsning går till och förklara vad syftet är med att använda en chart. Illustrera gärna med en enkel kontextfri grammatik.

12. Sammansättning är en produktiv process i svenska vilket innebär att de ofta inte finns i lexikonet för ett textanalyssystem. Även om man kan känna igen delarna är ytterligare ett problem att bestämma betydelsen. Detta kan angripas genom att man först försöker bestämma betydelsen av de ingående leden. Diskutera hur detta kan göras algoritmiskt. Vi antar att de möjliga betydelserna är listade i ett lexikon med definitioner. Exempel: ordet *kurs* har i svenska tre huvudbetydelser, ordet *bok* antas ha två. Ordet *kursbok* däremot kan ganska säkert anses vara en sammansättning av det ord *kurs* som betyder 'kort (del av) utbildning' snarare än 'riktning' eller 'pris, värde', medan *bok* gäller en skrift och inte ett träd. Jämför *dollarkurs*, *växelkurs*, *kursändring*, *kursfall*, ... som också är tämligen entydiga i fråga om vilken betydelse av *kurs* som aktualiseras.
13. I informationsextraktionssystem är man ofta intresserad av specifika entiteter och relationer mellan dem, t.ex. mellan personer och deras födelseår. Ett problem är då att de relationer man behöver modellera uttrycks på många olika sätt i naturlig text, ofta på sätt som är svåra att komma på, även för experter. En metod som har föreslagits för att hitta en stor mängd uttryckssätt för en given relation är bootstrapping. Beskriv vad denna metod går ut på.
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.