Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	*2010-03-19*
Sal	*TER1*
Tid	*8-12*
Kurskod	*729G17*
Provkod	*TEN1*
Kursnamn/benämning	*Språkteknologi*
Institution	*IDA*
Antal uppgifter som ingår i tentamen	*15*
Antal sidor på tentamen (inkl. försättsbladet)	*4*
Jour/Kursansvarig	*Lars Ahrenberg*
Telefon under skrivtid	*ankn 2422, 0703182422*
Besöker salen ca kl.	~~*09.99*~~ ⇒ ca 09.
Kursadministratör (namn + tfnnr + mailadress)	*Anna Grabska Eklund* *Ankn. 23 62, annek@ida.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

fredag 19 mars 2010 kl. 8-12

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 30. 15 poäng ger säkert godkänt.

Det går bra att besvara flera frågor på samma papper.

Del A

Besvara alla frågor i denna del

1. System för namnigenkänning utvärderas vanligen med både precision och recall. Förklara varför det inte är tillräckligt att bara ange hur många namn systemet hittat. (2p)

2. System för ordklasstaggning utnyttjar ords lokala kontext för att välja rätt tagg. Ange ett vanligt regelformat för sådan disambiguering och formulera någon eller några explicita regler som kan appliceras på ordet *vid* i dessa två meningar: (2p)

   Tomten lämnade säcken vid dörren.

   Klänningen var för vid i midjan.

3. I en text på sammanlagt 2000 ord finns 150 som börjar med stor bokstav. 80 av dessa inleder en mening och 100 är egennamn. Det finns 25 ordbigram, d.v.s. båda de ingående orden är egennamn. Använd Maximum Likelihood-uppskattning för att uppskatta följande två sannolikheter: (a) sannolikheten för att ett ord som börjar med stor bokstav är ett egennamn, (b) sannolikheten för att ett egennamn följs av ett egennamn. (2p)

4. Nominalfraser kan se ut på många olika sätt. Här är några exempel:

   hästen, en häst, en stor häst, en stor svart häst, en mycket stor häst, en mycket stor svart häst

   Ange ett reguljärt uttryck som sammanfattar strukturen i dessa nominalfraser med användning av alfabetet {A, D, N, R} där A står för adjektiv, D artikel/determinerare, N substantiv och R adverb. Uttrycket får inte matcha **en mycket häst, *stor en häst*. (2p)

5. Vad menas med redigeringsavstånd (Minimal edit distance, eller Levenshtein distance)? Beskriv kortfattat hur man räknar ut redigeringsavståndet och hur kan man använda det i samband med stavningskontroll. Illustrera med paret *skärna ~ stjärna.* (3p)

6. Förklara skillnaden mellan parsning top-down och parsning bottom-up. (2p)

7. Vad menas med partiell parsning (eller chunkning)? Varför föredras partiell parsning framför fullständig parsning i många applikationssystem som t.ex. informationsutvinning? (2p)

8. Vad är en statistisk språkmodell och vilken roll har sådana språkmodeller i statistisk maskinöversättning? (2p)

9. Språkteknologi använder sig ofta av rudimentära semantiska representationer (som många inte tycker förtjänar epitetet 'semantiska'), som t.ex. ordpåsar (eng. *bag-of-words*) eller mallar (eng. *templates*). Förklara vad dessa båda representationer står för och ange för var och en av dem någon typ av system som använder dem. (2p)

10. Ange två olika slags särdrag (eng. *features*) som är användbara i samband med ordbetydelsebestämning. (2p)

11. I modellen den brusiga kanalen (eng. *Noisy Channel*) används beslutsregeln

$$S^* = \underset{S}{\operatorname{argmax}}\, p(S|O)$$

(a) Förklara denna regel i ord. (b) Beräkningen av p(S|O) baseras vanligtvis på en omskrivning som leder till två olika sannolikhetsfördelningar. Visa hur omskrivningen ser ut och ange vad de resulterande sannolikheterna kallas. (3p)

Del B

Välj en av frågorna och besvara den utförligt. Varje fråga kan ge maximalt 6 poäng.

12. Förklara huvuddragen i Earley's algoritm, d.v.s. en chartparsningsalgoritm som använder grammatikreglerna top-down.

13. På amerikanska tangentbord saknas tangenter för å, ä och ö. Svenskar i USA skriver därför gärna mail utan dessa bokstäver varvid å och ä ersätts med a, och ö ersätts med o. Beskriv och motivera en design av ett system som återställer å, ä och ö i ett sådant mail.

14. Ett moment i informationsextraktionssystem är relationsbestämning. Förklara vad detta problem går ut på och beskriv någon datadriven metod att hitta olika uttryckssätt för en given relation.

15. Många taggningssystem, dvs system för att kategorisera ord i löpande text, använder transformationer. De nödvändiga transformationerna kan antingen konstrueras för hand eller läras in från uppmärkt text genom s.k. transformationsbaserad inlärning (TBL). Förklara hur TBL går till och ange minst två problem som det har tillämpats på.