Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

729G17 Språkteknologi

fredag 17 augusti 2012 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. Definiera begreppen precision och recall. (2p)

2. Ange de tre viktigaste dataresurserna i ett transformationsbaserat system för ordklasstaggning (en s.k. Brilltaggare). (2p)

3. (a) Ange ett reguljärt uttryck som identifierar svenska ord som slutar på bokstavssekvensen 'rna'; (b) Många sådana ord (t.ex. *skorna*) är substantiv i pluralis bestämd form, dock inte alla. Ange minst två andra klasser av svenska ord som kan matcha uttrycket; (c) Ange ett mer specifikt reguljärt uttryck som med bättre precision företrädesvis hittar just substantiv i pluralis bestämd form. Obs att texten inte förutsätts vara ordklasstaggad. (3p)

4. Förklara vad som menas med ett Levenshteinavstånd (eng. Levenshtein distance) och beräkna detta för de två orden *sommar* och *romans*. (2p)

5. Vid s.k. top-down chart-parsning av kontextfri grammatik används tre grundläggande operationer för att generera och bekräfta hypoteser: predicering, scanning och kombinering (eng. *completion*). Beskriv vad dessa operationer innebär och illustrera med den kontextfria grammatiken nedan: (3p)

S → NP VP	N → Jan, Jonna, katt, katten, …
NP → N	DET → en, ett, …
NP → DET N	V → har, ser, ...
VP → V NP	

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

6. I frågebesvarande system (eller QA-system) är begreppet svarstyp centralt. Förklara vad som menas med en svarstyp och ge några exempel. (2p)

7. Ange fyra centrala komponenter i ett informationsextraktionssysten givet att indata är en ordklasstaggad text som antas höra till systemets domän. (2p)

8. Enligt kursboken har ett extraktionsbaserat sammanfattningssystem (för s.k. single-document summarization) tre problem att lösa. Vilka? (2p)

9. I en svensk korpus omfattande ca 2 miljoner ord hittas följande frekvenser för några utvalda ord och ordsekvenser:

   mer eller mindre: 921
   mer eller: 934
   eller mindre: 1008
   mer: 26,159
   eller: 45,272
   mindre: 8,625

   Förklara varför dessa frekvenser talar för att ordsekvensen *mer eller mindre* är en lexikaliserad fras (ett fast uttryck) och för att sekvensen *mer eller* inte är det. (2p)

Del B

Välj två av frågorna 10-14 och besvara dem utförligt. Varje fråga kan ge maximalt 6 poäng.

10. I flera språkteknologisa tillämpningar kan det vara svårt att mäta recall på det vanliga sättet, dvs utifrån ett facit med rätta svar som systemets utdata direkt kan jämföras med. Ange minst två sådana tillämpningar och för var och en av dem, något sätt man använt för att utnyttja befintliga korrekta utdata på andra sätt.

11. Utjämning (eng. *smoothing*) är en teknik som används när man bygger statistiska språkmodeller (eng. *language models*). Förklara vad det innebär och varför man använder det. Beskriv därefter utförligt den utjämningsteknik som kallas Add-1 (eller Laplace).

12. 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.

13. Negativa omdömen kan ibland uttryckas med hjälp av ord som är positivt laddade. Exempel (i t.ex. en restaurang- eller hotelldomän):

    Maten var inte bra.

    Maten gjorde ingen människa glad.

    Maten kan knappast få något bra betyg.

    Diskutera utformningen av ett system som kan känna igen negativa omdömen av detta slag och föreslå också hur systemet bäst utvärderas.

14. Anta att man i en insamlad textkorpus vill märka upp verje förekomst av ett antal flertydiga ord med avseende på vilken betydelse de har där de står. Beskriv någon metod att göra detta.