729G17 Språkteknologi (2015)

Tentamen 2015-03-16

Marco Kuhlmann

Denna tentamen består av två delar: del A, som innehåller frågor 1–8, och del B, som innehåller frågor 9–12. Varje fråga är värd 3 poäng. För G krävs minst 12 poäng i del A. För VG krävs utöver detta minst 6 poäng i del B.

Viktig information! Du som skriver denna tentamen enligt den provuppsättning som gällde innan kursomgången 2015 (TEN1) ska även bearbeta extrauppgifterna 13 och 14 (efter del B). För G krävs då sammanlagt minst 15 poäng i del A och extrauppgifterna.

Del A

1. Ett textklassificeringssystem baserat på metoden Naive Bayes ska klassificera engelska nyhetstexter som antingen ”texter som handlar om Kina” (K) eller ”texter som handlar om Japan” (J). Systemet ska tränas på nedanstående dokumentsamling:

	dokument	klass
1	Chinese Beijing Chinese	K
2	Chinese Chinese Shanghai	K
3	Chinese Tokyo	K
4	Tokyo Japan Chinese	J

   Antag att systemet ska predicera klassen för dokumentet ”Chinese Chinese Chinese Tokyo”. Skatta de för denna klassificering relevanta sannolikheterna med Maximum Likelihood-metoden. Ställ upp bråk.

2. Ge tre exempel på tekniker som används för normalisering av textdokument (efter tokenisering) och förklara kort hur dessa tekniker fungerar.

3. I en korpus innehållande 1 215 396 token och 105 436 unika ord hittas ordet *det* 13 694 gånger, ordet *är* 13 700 gånger, ordet *nalkas* 2 gånger, sekvensen *det är* 927 gånger, och sekvensen *det nalkas* 0 gånger.

(a) Ställ upp bråk för ML-skattningen (Maximum Likelihood) av unigramsannolikheten $P(det)$ och bigramsannolikheten $P(är \mid det)$.

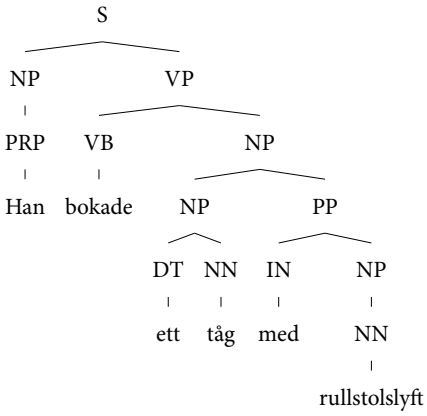
(b) Ställ upp ett bråk för ML-skattningen av bigramsannolikheten $P(nalkas \mid det)$ med Add One-utjämning. Antag att vokabulären består av alla unika ord.

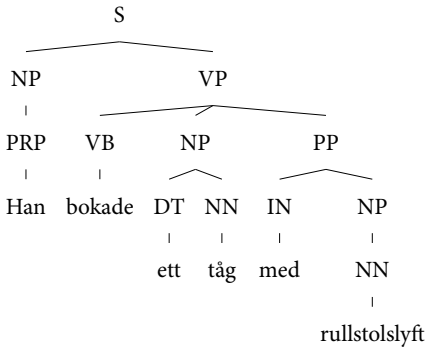
4. En Hidden Markov-modell för ordklasstaggning har genererat två taggsekvenser för meningen *Jag skrev på utan att tveka*:

	Jag	skrev	på	utan	att	tveka
sekvens 1	PN	VB	PP	PP	IE	VB
sekvens 2	PN	VB	PL	PP	IE	VB

Antag att du vet sannolikheten som modellen tilldelar sekvens 1. Hur kan du utifrån denna sannolikhet räkna ut sannolikheten för sekvens 2?

5. Nedanstående visas en liten trädbank bestående av två frasstrukturträd för meningen *Han bokade ett tåg med rullstolslyft*. Skatta sannolikheter för alla NP-regler och alla VP-regler med hjälp av Maximum Likelihood-metoden.





6. Följande synsets är hämtade från WordNet:

(1) *final examination*, *final exam*, *final* (sv. *tentamen*, ty. *Klausur*)
(2) *breakthrough, making an important discovery* (sv. *genombrott*, ty. *Durchbruch*)
(3) *communication*, *communicating* (sv. *kommunikation*, ty. *Kommunikation*)
(4) *act*, *deed*, *human activity* (sv. *mänsklig aktivitet*, ty. *menschliche Aktivität*)
(5) *examination*, *exam*, *test* (sv. *examination*, ty. *Prüfung*)
(6) *discovery* (sv. *upptäckt*, ty. *Entdeckung*)

Rita upp den partiella WordNet-hierarkin för dessa synsets. Vilken semantisk relation representerar en båge mellan två synsets? Hur stor är den semantiska likheten mellan (1) och (2) baserad på deras avstånd i hierarkin (*path length*)?

7. Ett namnigenkänningssystem testades på en samling testdata innehållande 800 namnförekomster. Av dessa namn bestod 500 av ett ord, 250 av två ord och 50 av tre ord. Tabellen nedan anger systemets resultat.

	korrekta	falska
Ettordsnamn	420	60
Tvåordsnamn	200	40
Treordsnamn	44	12

Ställ upp bråk för följande:

(a) systemets recall (täckning) på ettordsnamn
(b) systemets precision på treordsnamn
(c) systemets precision på samtliga namn

8. Förklara den standardarkitektur för frågebesvarande system som vi lärt känna under kursen och ge exempel på tekniker som kan användas för att lösa de olika deluppgifterna i denna arkitektur.

Del B

9. Förklara kortfattat hur Viterbi-algoritmen för ordklasstaggning fungerar. Vad gör den? På vilken grundidé bygger dess effektivitet?

10. I flera typer av språkteknologiska system kan täckning (recall) inte mätas på det vanligaste sättet, dvs. genom att dividera antalet fall där system och facit överensstämmer med det totala antalet fall i facit. Ange två typer av system där detta inte fungerar så bra, förklara varför, och beskriv de utvärderingsmått som används i stället.

11. Du är konsult inom ett forskningsprojekt som ska analysera texter i patientjournaler. För att få etikprövning krävs att texterna deidentifieras, dvs. att all information som kan användas för att spåra datan till enstaka patienter tas bort. Exempel på sådan känslig information är namn, personnummer, telefonnummer och adress. Beskriv hur deidentifiering skulle kunna implementeras med hjälp av tekniker från kursen. Föreslå och motivera även ett relevant utvärderingsmått.

12. Förklara modellen ”den brusiga kanalen” som används i samband med maskinöversättning. Hur kan man skatta de sannolikheter som ingår i denna modell?

Extrauppgifter

13. Inom språkteknologin används ordet *ord* på flera olika sätt. Ange tre olika betydelser och ge exempel på sammanhang där dessa olika betydelser är relevanta.

14. For att utvärdera maskinöversättningssystem används måttet BLEU som bygger på ett modifierat precisionsmått för n-gram. Förklara hur detta mått fungerar och varför man inte använder det vanliga precisionsmåttet. Illustrera dina svar med ett exempel.