TDDD01 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 3 krävs minst 12 poäng i del A. För 4 krävs utöver detta minst 4 poäng i del B, för 5 minst 8 poäng.

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.

Rättning: +1 p om det framgår att man vet vilka sannolikheter som är relevanta, +1 p om klassannolikheterna är korrekta, +1 p om ordsannolikheterna är korrekta.

$$P(\mathsf{K}) = 3/4 \qquad P(\text{Chinese} \mid \mathsf{K}) = 5/8 \qquad P(\text{Tokyo} \mid \mathsf{K}) = 1/8$$
$$P(\mathsf{J}) = 1/4 \qquad P(\text{Chinese} \mid \mathsf{J}) = 1/3 \qquad P(\text{Tokyo} \mid \mathsf{J}) = 1/3$$

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

Rättning: +1 p per teknik inkl. tillräckligt utförlig förklaring. Exempel: göra om versaler till gemener, eliminera stoppord (högfrekventa ord som *och*, *det*, *att*), lemmatisera ord (reducera ord till deras uppslagsformer).

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.

Rättning: +1 p per korrekt bråk: $\frac{13694}{1215396}$, $\frac{927}{13694}$, $\frac{0+1}{13694+105436}$

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

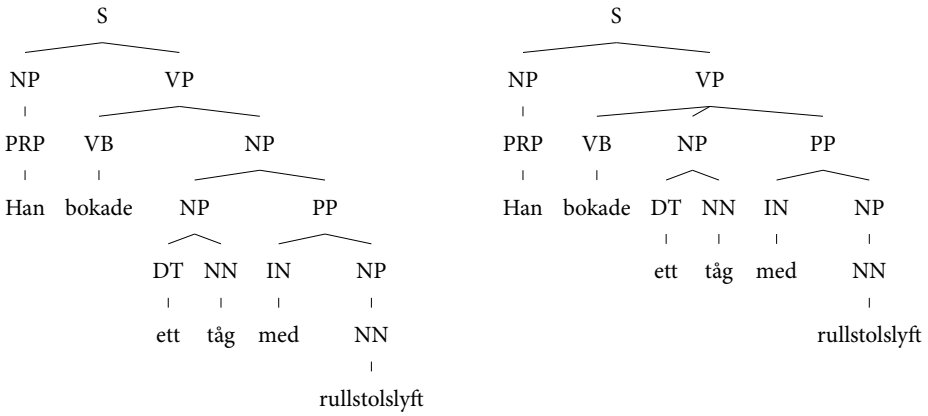
Antag att du vet sannolikheten som modellen tilldelar sekvens 1. Hur kan du utifrån denna sannolikhet räkna ut sannolikheten för sekvens 2?

Rättning: 3 p om man vet att man måste multiplicera sannolikheten för sekvens 1 med

$$\frac{P(\text{PL} \mid \text{VB}) \cdot P(\text{på} \mid \text{PL}) \cdot P(\text{PP} \mid \text{PL})}{P(\text{PP} \mid \text{VB}) \cdot P(\text{på} \mid \text{PP}) \cdot P(\text{PP} \mid \text{PP})}.$$

2 p om man har identifierat de sannolikheter som sekvenserna skiljer sig i. 1 p om det ur svaret framgår att man vet vilka typer av sannolikheter som ingår i en Hidden Markov-modell (tagg–tagg och tagg–ord).

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.



Rättning: 3 p om man har alla regler inkl. deras sannolikheter rätt:

$$\text{NP} \to \text{PRP}\ \tfrac{2}{7} \quad \text{NP} \to \text{NP PP}\ \tfrac{1}{7} \quad \text{NP} \to \text{DT NN}\ \tfrac{2}{7} \quad \text{NP} \to \text{NN}\ \tfrac{2}{7}$$
$$\text{VP} \to \text{VB NP}\ \tfrac{1}{2} \quad \text{VP} \to \text{VB NP PP}\ \tfrac{1}{2}$$

Poängavdrag: −1 p per regelgrupp om man har enstaka fel i den gruppen. 0 p om man inte angivit en väldefinierad probabilistisk kontextfri grammatik.

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*)?

Rättning: +1 p för korrekt nätstruktur (1 ← 5 ← 3 ← 4 → 6 → 2; alternativt kan man sätta 3 under 4). +1 p för korrekt angiven semantisk relation (hyponymi/hyperonymi). +1 p för korrekt likhet (1/6; för den alternativa strukturen får man 1/5).

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

Rättning: +1 p per korrekt svar. (a) 420/500 (b) 44/56 (c) 664/776

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.

Rättning: Poäng ges utifrån svarets omfattning och kvalitet: 1 p för en fragmentarisk beskrivning innehållande minst ett delproblem; 2 p för en sammanhängande beskrivning av flera delproblem; 3 p för en fullständig beskrivning av standardarkitekturen inklusive en förklaring av det centrala begreppet ”svarstyp”.

Del B

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

Rättning: Poäng ges utifrån svarets omfattning och kvalitet. Viterbi-algoritmen tar en Hidden Markov-modell och en ordsekvens och räknar ut den mest sannolika taggsekvensen för ordsekvensen enligt modellen. Grundidén är att bryta ner problemet i mindre delproblem. I varje delproblem frågar man: Vad är den mest sannolika taggsekvensen för orden $w_1, \ldots, w_i$ som slutar med tagg t? När man redan räknat ut den mest sannolika taggsekvensen för orden $w_1, \ldots, w_{i-1}$ som slutar med tagg t', för alla möjliga taggar t', kan man lösa det nya delproblemet genom att multiplicera dessa sannolikheter med övergångssannolikheten $P(t \mid t')$ och ordsannolikheten $P(w_n \mid t)$ och välja den taggsekvens som ger det största värdet.

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.

Rättning: Poäng ges utifrån svarets omfattning och kvalitet. Två exempel är textsammanfattning (alternativt mått: ROUGE) och maskinöversättning (alternativt mått: BLEU). I båda fall är problemet att det inte finns en guldstandard (ingen absolut korrekt textsammanfattning resp. översättning).

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.

Rättning: Poäng ges utifrån svarets omfattning och kvalitet. Ett sätt att lösa den här uppgiften skulle vara att träna en klassificerare som ska avgöra om ett token eller en tokensekvens innehåller känslig information. Ett relevant utvärderingsmått vore recall. För den specifika tillämpningen är recall betydligt viktigare än precision eftersom deidentifieringen måste vara mycket pålitlig.

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

Rättning: Poäng ges utifrån svarets omfattning och kvalitet. Modellen formaliserar översättning från en text skriven på språket R (”ryska”) till en text skriven på språket E (”engelska”) som beräkningen av uttrycket

$$\operatorname{argmax}_E P(E \mid R) = \operatorname{argmax}_E P(R \mid E)P(E)$$

Sannolikheterna $P(E)$ kan fås från en n-gram-modell för språket E; den i sin tur kan skattas från stora textmängder med hjälp av Maximum Likelihood-metoden. Sannolikheterna $P(R \mid E)$ kan fås genom att skatta dem från ordlänkade parallelkorpusar.