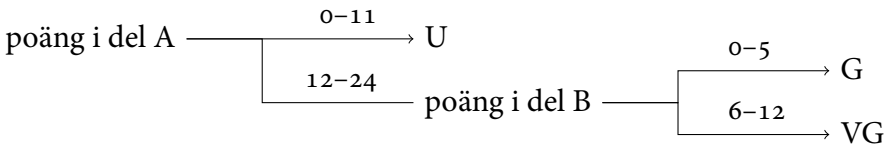
Tentamen 2016-03-16

Marco Kuhlmann

Tentamen består två delar, A och B. Varje del omfattar ett antal frågor à 3 poäng. **Del A** omfattar 8 frågor som kan besvaras kortfattat. Det krävs minst 12 poäng på denna del för att del B ska rättas. **Del B** omfattar 4 frågor som kräver utförliga svar med korrekt terminologi. Betyget sätts enligt följande:



Lycka till!

Del A

01 **Korrekthet, precision och täckning (recall).**

Vid utvärderingen av en ordklasstaggare fick man ut nedanstående förväxlingsmatris. Den markerade cellen anger antalet gånger systemet klassade ett ord som adjektiv (`JJ`) medan det enligt guldstandarden var ett substantiv (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

a) Ställ upp ett bråk för taggarens precision på adjektiv.

b) Ställ upp ett bråk för taggarens täckning (recall) på verb.

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men där respektive värden för a) och b) är 0%.

02 Textklassificering

Ett system för textklassificering baserat på metoden Naive Bayes ska avgöra om dokumentet ”Tokyo Tokyo Peking” är en nyhet om Japan (klass `J`) eller en nyhet om Kina (klass `K`).

a) För att predicera dokumentets klass använder systemet bl.a. sannolikheterna $P(\texttt{J})$ och $P(\text{Tokyo} \mid \texttt{J})$. Lista de övriga sannolikheter som är relevanta.

b) Skatta de relevanta sannolikheterna med Maximum Likelihood-metoden utifrån följande dokumentsamling. Ställ upp bråk.

	dokument	klass
1	Tokyo Tokyo	`J`
2	Tokyo Peking	`J`
3	Tokyo Seoul	`J`
4	Peking Tokyo	`K`

c) Utifrån de skattade sannolikheterna, vilken klass predicerar systemet? Redovisa hur du räknat. Visa tydligt att du förstått Naive Bayes-klassificeringsregeln.

03 Ordpredicering

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

a) Skatta unigramsannolikheten $P(\textit{är})$ och bigramsannolikheten $P(\textit{är} \mid \textit{det})$ med Maximum Likelihood-metoden. Ställ upp bråk.

b) Vad händer när man skattar bigramsannolikheten $P(\textit{nalkas} \mid \textit{det})$ med Maximum Likelihood-metoden? Varför kan detta vara ett problem?

c) Skatta bigramsannolikheten $P(\textit{nalkas} \mid \textit{det})$ med Maximum Likelihood-metoden och addera-k-utjämning med $k = 0{,}01$ (**inte** addera-1). Ställ upp bråk.

04 Ordklasstaggning

Följande matriser specificerar en Hidden Markov-modell. Istället för sannolikheter anges kostnader (negativa log-sannolikheter).

	PL	PN	PP	VB	EOS
BOS	11	2	3	4	19
PL	17	3	2	5	7
PN	5	4	3	1	8
PP	12	4	6	7	9
VB	3	2	3	3	7

	hen	vilar	ut
PL	17	17	4
PN	3	19	19
PP	19	19	3
VB	19	8	19

När man använder Viterbi-algoritmen för att beräkna den mest sannolika (minst kostsamma) taggsekvensen för meningen ”hen vilar ut” enligt denna modell får man ut följande matris. Notera att matrisen saknar tre värden (markerade celler).

		hen	vilar	ut
BOS	0			
PL		28	27	21
PN		A	28	35
PP		22	27	20
VB		23	B	36
EOS				C

a) Beräkna värdet A. Redovisa hur du räknat.

b) Beräkna värdena B och C. Redovisa hur du räknat.

c) Ange den mest sannolika taggsekvensen för meningen. Förklara hur den kan fås från (den fullständiga) matrisen.

05 Syntaktisk analys

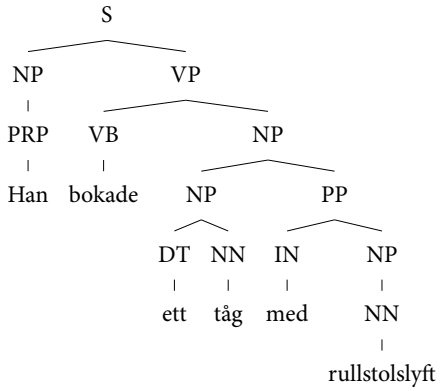
Nedan anges alla NP-regler och alla VP-regler i en viss probabilistisk kontextfri grammatik. Den sista NP-regeln saknar ett sannolikhetsvärde.

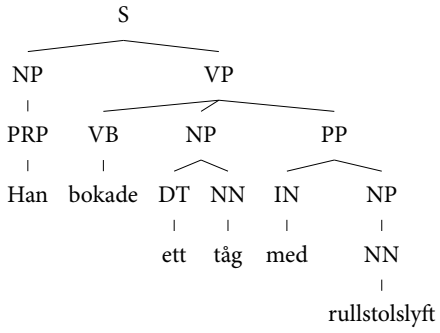
$$\text{NP} \to \text{PRP}\ \tfrac{2}{7} \qquad \text{NP} \to \text{NP PP}\ \tfrac{1}{7} \qquad \text{NP} \to \text{DT NN}\ \tfrac{2}{7} \qquad \text{NP} \to \text{NN}\ ?$$

$$\text{VP} \to \text{VB NP}\ \tfrac{1}{2} \qquad \text{VP} \to \text{VB NP PP}\ \tfrac{1}{2}$$

a) Vilket sannolikhetsvärde saknas?

b) Nedan anges två träd som genererats av grammatiken. Ställ upp bråk för deras sannolikhetsvärden. Antag att alla regler som inte anges ovan har sannolikhet 1.





c) Hur måste man ändra sannolikheterna för VP-reglerna så att det vänstra trädet får högre sannolikhet än det högra? (Du behöver inte ange de exakta värdena.)

06 Semantisk analys

En enkel metod för att bestämma ett ords betydelse är Lesks algoritm, som använder sig av semantiska lexikon. Följande betydelser hittar man när man slår upp det engelska ordet *course* i Wiktionary:

1 A normal or customary sequence. **2** A learning program, as in university. *I need to take a French course.* **3** The direction of movement of a vessel at any given moment. *The ship changed its course 15 degrees towards south.*

a) Vilken av betydelserna har ordet *course* i följande mening?

In the United States, the normal length of a course is one academic term.

b) Vilken av betydelserna förutsäger Lesks algoritm för denna mening med de angivna definitionerna som underlag? Ignorera ordet *course*, skiljetecken och stoppord. Motivera ditt svar.

c) Förändra meningen så att ordet *course* behåller sin betydelse men Lesks algoritm nu förutsäger en annan betydelse än innan.

07 Informationsextraktion

Ett system för informationsextraktion är tränat på att hitta tre typer av namngivna entiteter: personer (PER), organisationer (ORG) och svenska tätorter (LOC).

a) Vilken av dessa tre typer är lättast att hitta med hjälp av namnlistor?

b) Entitetsextraktion kan ses som uppgiften att tagga varje token i en mening med en så kallad BIO-tagg. Sätt ut BIO-taggarna för följande mening. (Observera att meningen består av 20 stycken token.)

Astrid Lindgren , född den 14 november 1907 i Vimmerby , utsågs till hedersdoktor vid Linköpings universitet år 2000 .

c) System som använder BIO-taggningen kan utvärderas på taggnivå eller entitetsnivå. Ändra en av dina taggar så att den nya taggningen har 95% korrekthet på taggnivå men 0% precision och recall med avseende på entitetstypen PER. Använd din ursprungliga taggning som guldstandard.

08 Frågebesvarande system

Följande frågor ställdes till ett frågebesvarande system:

A När föddes Einstein?

B Hur var han som människa?

C Var föddes Einsteins första fru?

a) Ange svarstyperna för A och C.

b) Vilka data behövs för att träna en Naive Bayes-klassificerare som automatiskt predicerar en frågas svarstyp?

c) Olika frågor är olika svåra för frågebesvarande system att hitta rätt svar på. Rangordna de tre frågorna från lättast till svårast. Motivera din rangordning. Antag att systemet inte har någon egen kunskapsdatabas utan bygger på dokumentsökning i svenska Wikipedia.

Del B

09 **Rättstavningskorrektur**

När man googlar på ett felstavat ord som t.ex. *rätstavning* så tar det endast några bråkdelar av en sekund och Google frågar:

> Menade du: *rättstavning*

I den här uppgiften ska du fundera på hur denna teknik kan implementeras.

a) Förklara hur en enkel algoritm skulle kunna fungera som tar ett ord w och beräknar alla ord vars Levenshtein-avstånd till w är exakt ett.

b) Hur skulle man kunna utöka denna algoritm till en algoritm som föreslår det mest sannolika rättstavade ordet?

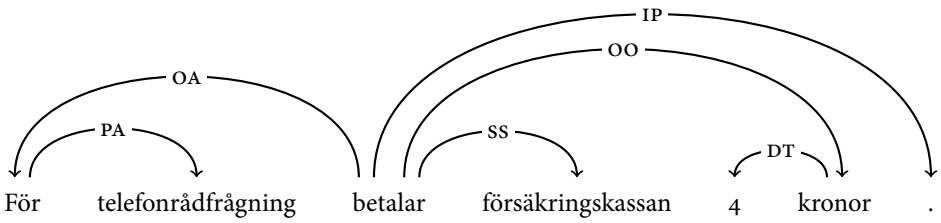
c) Hur skulle man kunna generalisera ansatsen till ord vars Levenshtein-avstånd till det felstavade ordet är större än ett? Vilket beräkningsmässigt problem uppstår?

10 **Ordklasstaggning**

Diskutera likheter och skillnader mellan Viterbi-algoritmen och den perceptron-baserade algoritmen för ordklasstaggning. Vilka fördelar och nackdelar finns? För vilka tillämpningar passar den ena bättre än den andra?

11 **Transitionsbaserad dependensparsning**

Förklara utförligt hur parsning med en girig dependensparser fungerar. Illustrera ditt svar med hänvisningar till nedanstående dependensträd. Ange en transitionssekvens som skapar bågarna i detta träd.



12 **Attitydanalys**

Ditt företag har utvecklat ett framgångsrikt system för klassificering av nyhetstexter baserat på modellen Naive Bayes. Nu blir ni kontaktade av en kund som undrar om systemet även kan användas för predicering av attityder gentemot kundens produkter utifrån yttranden på sociala medier som Twitter och Facebook.

Diskutera likheter och skillnader mellan de två tillämpningarna. Vilka utmaningar ser du med den nya tillämpningen? Vilka data och andra resurser skulle ni kunna ha nytta av om ni bestämde er för att utveckla en anpassad version av ert system?