

Schriftliche Prüfung zur Vorlesung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2022/23
Dozent: Helmut Schmid

Sie haben **90 Minuten** Zeit für die Bearbeitung der Aufgaben.

Aufgabe 1) Die folgenden Formeln definieren jeweils die Wahrscheinlichkeiten für ein bestimmtes statistisches Modell, das wir in der Vorlesung behandelt haben.

Achtung: Die Formeln wurden gegenüber den Formeln aus der Vorlesung abgewandelt durch Austausch von Variablennamen und andere Umformungen. Variablen in Fettschrift stehen für Listen/Folgen/Vektoren.

- a)
$$p(l_1, l_2, \dots, l_m) = \prod_{k=1}^{m+1} p(l_k | l_{k-2}, l_{k-1})$$
- b)
$$\log p(l_1, l_2, \dots, l_m | d_1, d_2, \dots, d_m) = -S(d_1, \dots, d_m) + \sum_{k=1}^{m+1} s(l_{k-1}, l_k, \mathbf{d}, k)$$
- c)
$$p(d_1, \dots, d_m) = \prod_{k=1}^m p(d_k) \quad \text{mit } d_k = (l, (r_1, \dots, r_{L_{d_k}}))$$
- d)
$$p(l, d_1, \dots, d_m) = p(l) \prod_{k=1}^{m+1} p(d_k | l)$$
- e)
$$p(l | \mathbf{d}) = \frac{\prod_{k=1}^m e^{a_k b_k(l, \mathbf{d})}}{S(\mathbf{d})}$$

Bearbeiten Sie für jede der Formeln a) bis e) die folgenden Teilaufgaben:

- I) Wie heißt das entsprechende **statistische Modell**?
- II) Nennen Sie eine konkrete computerlinguistische **Anwendung** für dieses Modell. Welche Bedeutung hat jede einzelne Variable (inklusive der Variablen m und k) bei dieser Anwendung? (Eine Anwendung genügt hier.)
- III) Geben Sie für die in II) gewählte Anwendung ein **Beispiel** für die Argumente “...” der Wahrscheinlichkeitsverteilung $p(\dots)$ bzw. $\log p(\dots)$ auf der linken Seite der Formel an.

Denken Sie sich ein **neues** Beispiel aus dem Themen-Bereich **Tourismus** aus. Wenden Sie die Formel auf dieses Beispiel an: Schreiben Sie also hin, wie nach der Formel die Wahrscheinlichkeit für Ihr Beispiel zu **berechnen** ist.

(10 Punkte)

Aufgabe 2) Gegeben sei die Buchstabenfolge *baaab*. Sie sollen damit die Parameter eines buchstabenbasierten **Markowmodelles** 1. Ordnung auf 3 Arten schätzen. (Es muss erkennbar sein, wie Sie jeweils das Ergebnis berechnet haben.)

Im Folgenden sind x und y Variablen, während a und b Buchstaben sind.

- Extrahieren Sie aus der Folge *baaab* die **Häufigkeiten** $f(x,y)$ aller Buchstabenpaare (x,y) . Berücksichtigen Sie dabei auch die Grenzsymbole.
- Berechnen Sie die **Kontexthäufigkeiten** $f_1(x)$ aus den Bigramm-Häufigkeiten $f(x,y)$.
- Berechnen Sie die **Unigramm-Häufigkeiten** $f_2(y)$ aus den Bigramm-Häufigkeiten $f(x,y)$. (Diese Häufigkeiten werden für die Schätzung der Backoff-Wahrscheinlichkeits-Verteilung gebraucht.)
- Berechnen Sie die Unigramm-Häufigkeiten $f_3(y)$ nach dem **Kneser-Ney**-Verfahren.
- Berechnen Sie die **Backoff-Wahrscheinlichkeit** $p(a)$ nach dem normalen Verfahren und nach dem Kneser-Ney-Verfahren.
- Berechnen Sie den **Backoff-Discount** δ für die Häufigkeiten $f(x,y)$.
- Berechnen Sie die (ungeglättete) **relative Häufigkeit** $p_0(a|a)$.
- Berechnen Sie die **relative Häufigkeit mit Discount** $r(a|a)$.
- Wie berechnen Sie den **Backoff-Faktor** $\alpha(a)$ für die interpolierte Backoff-Glättung aus den oben berechneten Werten?
(Den genauen Wert müssen Sie nicht berechnen.)
- Wie berechnen Sie die mit **interpoliertem Backoff** geglätteten Wahrscheinlichkeiten $p(\langle/s\rangle|a)$ und $p(b|a)$? (Den genauen Wert müssen Sie nicht berechnen.)

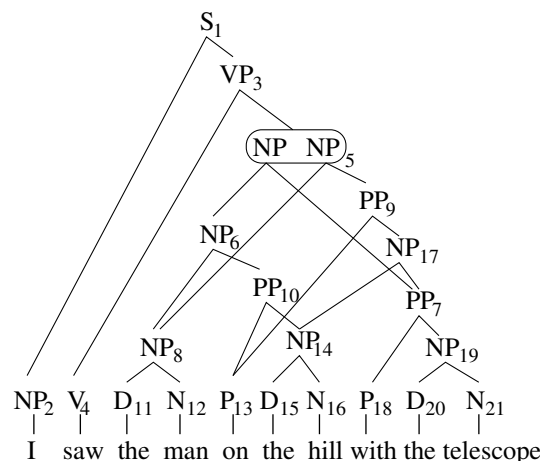
(10 Punkte)

Aufgabe 3) Die Inside-Wahrscheinlichkeiten α und Outside-Wahrscheinlichkeiten β eines Parsewaldes werden nach folgenden Formeln berechnet:

$$\begin{aligned}\alpha(a) &= 1 \quad \text{für jedes Terminalsymbol } a \\ \alpha(A \rightarrow X_1 \dots X_n) &= p(A \rightarrow X_1 \dots X_n) \prod_{i=1}^n \alpha(X_i) \\ \alpha(A) &= \sum_{A \rightarrow \gamma} \alpha(A \rightarrow \gamma)\end{aligned}$$

$$\begin{aligned}\beta(S) &= 1 \quad \text{für das Startsymbol } S \\ \beta(B \rightarrow X_1 \dots X_m \underline{A} X_{m+1} \dots X_n) &= \beta(B) p(B \rightarrow X_1 \dots X_m \underline{A} X_{m+1} \dots X_n) \prod_{i=1}^n \alpha(X_i) \\ \beta(A) &= \sum_{B \rightarrow \gamma \underline{A} \delta} \beta(B \rightarrow \gamma \underline{A} \delta)\end{aligned}$$

Wie berechnen Sie konkret im folgenden Parsewald die **Inside**-Wahrscheinlichkeiten $\alpha(NP_5)$ und $\alpha(P_{13})$ und die **Outside**-Wahrscheinlichkeiten $\beta(NP_5)$ und $\beta(P_{13})$ aus den anderen Inside- und Outside-Wahrscheinlichkeiten?



(4 Punkte)

Aufgabe 4) Ein Hidden-Markow-Modell ist gegeben durch die Tabelle

	A	B	$\langle s \rangle$	a	b	x	ϵ
A	0.5	0	0.5	0.5	0	0.5	0
B	0	0.5	0.5	0	0.5	0.5	0
$\langle s \rangle$	0.5	0.5	0	0	0	0	1

mit $p(\langle s \rangle | A) = 0.5$ und $p(x | A) = 0.5$. Start- und Ende-Tag sind hier identisch und ϵ ist das Endetoken.

Berechnen Sie für die Tokenfolge $x x a$ und das obige HMM die **Viterbi**-Wahrscheinlichkeiten $\delta_t(i)$ und die besten Vorgänger-Tags $\psi_t(i)$ nach den Formeln:

$$\begin{aligned}\delta_t(0) &= \begin{cases} 1 & \text{falls } t = \langle s \rangle \\ 0 & \text{sonst} \end{cases} \\ \delta_t(k) &= \max_{t'} \delta_{t'}(k-1) p(t|t') p(w_k|t) \quad \text{für } 0 < k \leq n+1 \\ \psi_t(k) &= \arg \max_{t'} \delta_{t'}(k-1) p(t|t') p(w_k|t) \quad \text{für } 0 < k \leq n+1\end{aligned}$$

Schreiben Sie nicht nur das Ergebnis hin, sondern zeigen Sie den Rechenweg.

Wie groß ist die Gesamtwahrscheinlichkeit aller Tagfolgen? (6 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!