Schriftliche Prüfung zur Vorlesung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2020/21
Dozent: Helmut Schmid

Sie haben **90 Minuten** Zeit plus 5 Minuten zum Absenden Ihrer Lösungen per Email. Sie können eine Textdatei oder einen Scan schicken.

Wenn Sie einen Fehler in einer der Aufgaben entdecken sollten, dann melden Sie sich bitte per Zoom (aber nicht Zoom-Chat) oder rufen Sie mich unter der Nummer 07121 44240 an.

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.

a)
$$p(y|x) = \frac{1}{N(x)} e^{\sum_k w_k m_k(x,y)}$$

b)
$$p(x_1, x_2, ..., x_m) = \prod_{k=1}^{m+1} p(x_k | x_{k-i}, ..., x_{k-1})$$

c)
$$p(y, x_1, ..., x_m) = p(y) \prod_{k=1}^{m} p(x_k|y)$$

d)
$$p(y_1, ..., y_m, x_1, ..., x_m) = \left[\prod_{k=1}^{m} p(y_k|y_{k-2}, y_{k-1})\ p(x_k|y_k)\right] p(\langle s\rangle | y_{m-1}, y_m)$$

e)
$$p(B) = p(\pi_1, ..., \pi_m) = \prod_{k=1}^{m} p(\pi_k) \quad \text{wobei } \pi_k \text{ die Form } A \to x_1...x_l \text{ hat}$$

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 i 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(...) auf der linken Seite der Formel an.

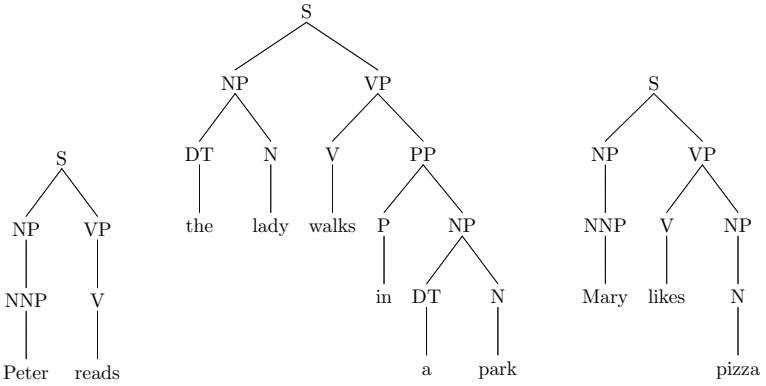
Denken Sie sich ein **neues** Beispiel aus dem Themen-Bereich **Medizin** aus. Nehmen Sie kein Beispiel, das im Kurs behandelt wurde.

IV) Wenden Sie die Formel auf das in III) gewählte Beispiel an: Schreiben Sie also hin, wie gemäß der Formel die Wahrscheinlichkeit für Ihr Beispiel zu **berechnen** ist.

Sie müssen hier insgesamt 5*4=20 Teilfragen beantworten. **Beispiel:** (allerdings kein statistisches Modell) Formel: $p(a\|b) = p(b\|a)p(a)/p(b)$ I) Name: Bayes'sches Theorem II) Anwendung: Transformation der lexikalischen Wahrscheinlichkeiten beim HMM-Tagger zur besseren Behandlung unbekannter Wörter; a ist bei dieser Anwendung ein Wort und b ein Tag. III) Beispiel: bedingte Wahrscheinlichkeit des Wortes "Haus" gegeben das Tag "N" IV) $p(Haus\|N) = p(N\|Haus)p(Haus)/p(N)$

(10 Punkte)

Aufgabe 2) Extrahieren Sie eine kontextfreie **Grammatik** inklusive Regelhäufigkeiten aus der folgenden Baumbank. Schätzen Sie dann die ungeglätteten **Regelwahrscheinlichkeiten**.

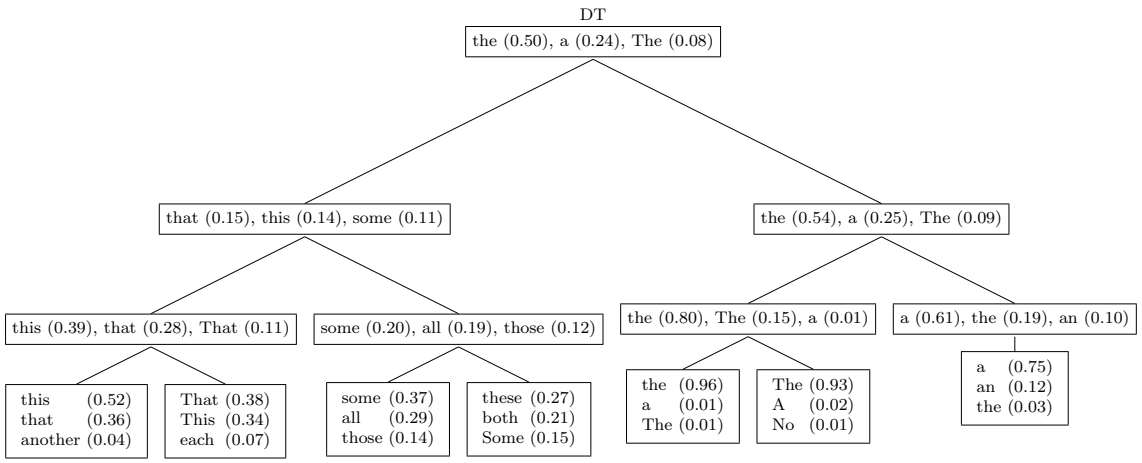


(4 Punkte)

Aufgabe 3) Die Wahrscheinlichkeit p(N → Haus) der lexikalischen PCFG-Regel N → Haus könnte man auch als $p(Haus|N)$ schreiben, also als Wahrscheinlichkeit der rechten Seite der Regel "Haus" gegeben die linke Seite der Regel "N".

Erklären Sie detailliert, wie man durch mathematische **Transformation** von $p(Haus|N)$ bessere Wahrscheinlichkeitsschätzungen für Wörter bekommt, die nicht in den Trainingsdaten aufgetaucht sind. (4 Punkte)

Aufgabe 4) Gegeben ist das folgende Schaubild aus der Vorlesung:



Erklären Sie, wie der zugehörige **Parser** trainiert wird (bspw. Vorverarbeitung der Daten, Trainingsschritte), und wie man dabei die im Schaubild gezeigte Struktur bekommt.

(3 Punkte)

Aufgabe 5) Ein HMM 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 $p(x|A) = 0.5$.

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

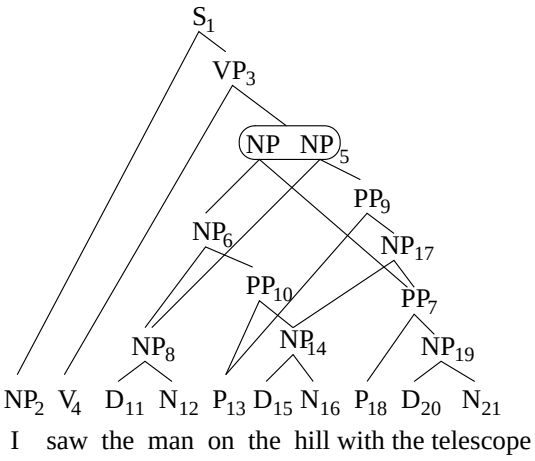
$$\begin{aligned}
\delta_t(0) &= \begin{cases} 1 & \text{falls } t = \langle s \rangle \\ 0 & \text{sonst} \end{cases} \\
\delta_t(k) &= \max_{t'} \; \delta_{t'}(k-1)\; p(t|t')\; p(w_k|t) \quad \text{für } 0 < k \leq n+1 \\
\psi_t(k) &= \arg\max_{t'} \; \delta_{t'}(k-1)\; p(t|t')\; p(w_k|t) \quad \text{für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

Schreiben Sie nicht nur das Ergebnis hin, sondern zeigen Sie den Rechenweg. Extrahieren Sie dann die beste **Tagfolge** nach den Formeln

$$\begin{aligned}
t_n &= \psi_{\langle s \rangle}(n+1) \\
t_k &= \psi_{t_{k+1}}(k+1) \quad \text{für } n > k > 0
\end{aligned}$$

(5 Punkte)

Aufgabe 6) Wie werden im folgenden Parsewald die **Inside**-Wahrscheinlichkeiten $\alpha(\text{NP}_5)$ und $\alpha(\text{NP}_8)$ und die **Outside**-Wahrscheinlichkeiten $\beta(\text{NP}_5)$ und $\beta(\text{NP}_8)$ aus den anderen Inside- und Outside-Wahrscheinlichkeiten berechnet? (Zeigen Sie auch die Zwischenschritte und nicht nur das Endergebnis.)



zugehörige Formeln:

$\alpha(a) = 1$ für jedes Terminalsymbol a
$\alpha(A \to X_1 ... X_n) = p(A \to X_1 ... X_n) \prod_{i=1}^{n} \alpha(X_i)$
$\alpha(A) = \sum_{A \to \gamma} \alpha(A \to \gamma)$

$\beta(S) = 1$
$\beta(B \to X_1 ... X_m \underline{A} X_{m+1} ... X_n) = \beta(B) \cdot$
$\quad p(B \to X_1 ... X_m A X_{m+1} ... X_n) \prod_{i=1}^{n} \alpha(X_i)$
$\beta(A) = \sum_{B \to \gamma A \delta} \beta(B \to \gamma \underline{A} \delta)$

(4 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!