

**Schriftliche Wiederholungsprüfung 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 in der Vorlesung behandelt wurde.

*Achtung: Die Formeln wurden gegenüber den Formeln aus der Vorlesung abgewandelt durch Austausch von Variablennamen und andere Umformungen.*

a)

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

b)

$$p(y | \mathbf{x}) = \frac{\prod_k e^{w_k m_k(\mathbf{x}, y)}}{N(\mathbf{x})}$$

c)

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

d)

$$p(x_1, \dots, x_m, y_1, \dots, y_m) = p(\langle /s \rangle | y_m) \prod_{k=1}^m p(y_k | y_{k-1}) p(x_k | y_k)$$

e)

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

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 Variable bei dieser Anwendung?
- III) Geben Sie für die in II) gewählte Anwendung ein **Beispiel** für die Argumente “...” der Wahrscheinlichkeitsverteilung  $p(\dots)$  auf der linken Seite der Formel an. Nehmen Sie ein Beispiel aus dem Themen-Bereich **Tiere**.

Wenden Sie die Formel auf dieses Beispiel an und schreiben Sie hin, wie die Wahrscheinlichkeit für Ihr Beispiel gemäß der Formel zu **berechnen** ist.

(10 Punkte)

### Aufgabe 2) Schätzung von Wahrscheinlichkeiten

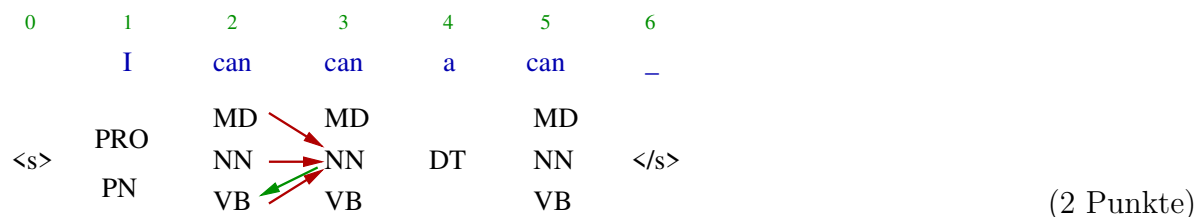
- Wie schätzen Sie die **ungeglätteten Wahrscheinlichkeiten** von  $p(x)$  aus den Häufigkeiten  $f(x)$ ?
- Wie schätzen Sie die ungeglätteten **bedingten Wahrscheinlichkeiten** (relative Häufigkeiten) von  $p(y|x)$  aus den Häufigkeiten  $f(x, y)$ ?
- Wie berechnen Sie **relative Häufigkeiten mit Discount**  $r(y|x)$  aus den Häufigkeiten  $f(x, y)$  und dem Discount  $\delta$ ? (*Relative Häufigkeiten mit Discount werden für die Backoff-Glättung gebraucht.*) (3 Punkte)

**Aufgabe 3)** Erklären Sie, warum ein **buchstabenbasiertes Markowmodell** 3. Ordnung, das auf deutschen Texten trainiert wurde, dem Satz “Er liest ein Buch.” eine höhere Wahrscheinlichkeit gibt als dem kürzeren Satz “Er liest ein Buc.” Worin unterscheiden sich die Wahrscheinlichkeiten der beiden Buchstabenfolgen? (2 Punkte)

**Aufgabe 4)** Was ist bei der Parameterglättung der Unterschied zwischen **Relative Discounting** und **Absolute Discounting**? Welche Form der Parameterglättung ist in der Computerlinguistik besser geeignet? (2 Punkte)

**Aufgabe 5)** Wie müssen Sie die lexikalische Wahrscheinlichkeit  $p(word|tag)$  eines HMM-Wortart-Taggers transformieren, damit Sie eine Variante des HMM erhalten, welche **unbekannte Wörter** besser verarbeiten kann? Wie schätzen Sie die geglätteten Wahrscheinlichkeiten? (2 Punkte)

**Aufgabe 6)** Wie berechnen Sie konkret im Beispiel unten die **Viterbiwahrscheinlichkeit**  $\delta_{NN}(3)$  und das beste Vorgängertag  $\psi_{NN}(3)$  bei einem HMM-Tagger erster Ordnung?



**Aufgabe 7)** Erklären Sie detailliert, wie Sie mit einem **Vorzeichentest** prüfen, ob TaggerB signifikant genauer ist als TaggerA. (2 Punkte)

**Aufgabe 8)** Wie ist bei einer **probabilistischen kontextfreien Grammatik** die Wahrscheinlichkeit eines Parsebaumes/eines Satzes/einer Folge von Sätzen jeweils definiert? Sie können hier Formeln angeben oder eine Textantwort geben. (3 Punkte)

**Aufgabe 9)** Wie gehen Sie beim **unüberwachten Training** eines HMM-Taggers vor? Welche Daten brauchen Sie? Wie initialisieren Sie das HMM? Wie trainieren Sie das Modell? Wann beenden Sie das Training? (4 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!