Schriftliche Wiederholungsprüfung
Vorlesung Statistische Methoden in der Sprachverarbeitung
WS 2014/15
Dozent: Helmut Schmid

Aufgabe 1) Wie ist die Wahrscheinlichkeit der getaggten Wortfolge "The/DT man/NN slept/VBD" bei einem **HMM-Wortart-Tagger** 2. Ordnung (Trigramm-Tagger) definiert? (3 Punkte)

Aufgabe 2) Welche Art von Daten brauchen Sie, um die Parameter des HMMs **überwacht** zu trainieren? (1 Punkt)

Aufgabe 3) Wie lauten die Formeln zur Schätzung von **ungeglätteten Werten** für die Parameter eines HMM aus den Trainingsdaten (sog. Maximum-Likelihood-Estimate)? Geben Sie an, was mit den Variablen bezeichnet wird, die Sie verwenden. (2 Punkte)

Aufgabe 4) Wozu dient eine **Parameterglättung**? Welches Problem soll gelöst werden? (1 Punkt)

Aufgabe 5) Wie glätten Sie die **Kontextwahrscheinlichkeiten** eines HMMs am besten? Geben Sie die Formel für das Glättungsverfahren an. (2 Punkte)

Aufgabe 6) Wie geht ein HMM-Tagger am besten mit unbekannten Wörtern um? Beschreiben Sie die Methode genau, am besten auch mit Formeln. (2 Punkte)

Aufgabe 7) Wie kann ein Wortart-Tagger auf ungetaggten Daten trainiert werden? Welche Daten benötigen Sie dafür? Wie läuft das Training ab? Wie lauten die Formeln für den dabei verwendeten Algorithmus? (5 Punkte)

Aufgabe 8) Wie wird bei PCFGs die Wahrscheinlichkeit eines Parsebaumes, die Wahrscheinlichkeit eines Satzes und die Wahrscheinlichkeit einer Folge von Sätzen (=Korpus) definiert? (3 Punkte)

Aufgabe 9) Wie berechnet der Viterbi-Algorithmus den besten Parse für einen gegebenen Parsewald? Wie werden die Viterbi-Wahrscheinlichkeiten berechnet (mit Formeln)? (5 Punkte)

Aufgabe 10) Wie arbeitet der Inside-Outside-Algorithmus und wie trainiert man damit eine PCFG? (4 Punkte)

Aufgabe 11) Wie wird eine Grammatik **markowisiert**. Was ist der Vorteil der Markowisierung? (2 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!