2. Schriftliche Prüfung zur Vorlesung
Statistische Sprachverarbeitung
WS 2013/14
Helmut Schmid

Aufgabe 1) Geben Sie die allgemeine Formel an, mit der bei einem Hidden-Markow-Modell 2. Ordnung (d.h. einem Trigramm-HMM) die Wahrscheinlichkeit $p(w_1^n, t_1^n)$ einer Wortfolge w_1^n mit der Tagfolge t_1^n berechnet wird. Was ist bzgl. Satzanfang und Satzende zu beachten?

Welche Wahrscheinlichkeiten müssen konkret für die Wortfolge "Es regnet" und die Tagfolge "PPER VVFIN" multipliziert werden? (5 Punkte)

Aufgabe 2) Warum sind unbekannte Wörter beim Wortart-Taggen problematisch? Wie kann ein HMM-Tagger mit unbekannten Wörtern umgehen? Wie müssen Sie hierfür die "HMM-Formel" anpassen? (4 Punkte)

Aufgabe 3) Ein Naive-Bayes-Modell kann zur Wortbedeutungsdesambiguierung verwendet werden. Wie lautet die Formel zur Berechnung der wahrscheinlichsten Lesart eines Wortes und wie werden die Wahrscheinlichkeits-Parameter geschätzt? (4 Punkte)

Aufgabe 4) Wie wird die Maximum-Likelihood-Schätzung der bedingten Wahrscheinlichkeit $p(b|a)$ aus Häufigkeiten $f(a, b)$ berechnet? Welche Probleme kann es bei diesem Schätzverfahren geben? (3 Punkte)

Aufgabe 5) Wie lautet die Formel für die Backoff-Glättung mit Absolute Discounting? Welche Art von Wahrscheinlichkeiten können mit Backoff-Glättung geglättet werden? (4 Punkte)

Aufgabe 6) Wie ist die Wahrscheinlichkeit eines Parsebaumes bei einer probabilistischen kontextfreien Grammatik (PCFG) definiert? (2 Punkte)

Aufgabe 7) Erläutern Sie den EM-Algorithmus am Beispiel des unüberwachten Trainings von PCFGs (also Training auf Rohtexten). Welche Daten benötigen Sie? Welche Berechnungsschritte führt der EM-Algorithmus aus? (4 Punkte)

Aufgabe 8) Angenommen Sie trainieren einen HMM-Tagger mit dem Forward-Backward-Algorithmus. Wie können Sie die erwartete Häufigkeit (= Aposteriori-Wahrscheinlichkeit) des Tags t an der Position des Wortes w_k aus den Forward- und Backward-Wahrscheinlichkeiten berechnen?

Wie berechnen Sie die erwartete Häufigkeit des Tagpaares t und t' an den Positionen der Wörter w_k und w_{k+1}. (4 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!