**Schriftliche Prüfung zur Übung**
**Statistische Methoden in der maschinellen Sprachverarbeitung**
**SS 2014**
**Helmut Schmid**

**Aufgabe 1)** Welche Wahrscheinlichkeitsformel verwendet ein Naive-Bayes-Klassifikator? Welche Unabhängigkeitsannahmen macht der Klassifikator? Wie wählt der Klassifikator die Ergebnisklasse aus? (5 Punkte)

**Aufgabe 2)** Die Formel für die Backoff-Glättung mit Absolute Discounting lautet in der interpolierten Version:

$$p(a_k|a_1^{k-1}) = \frac{f(a_1^k) - \delta_k}{f(a_1^{k-1})} + \alpha(a_1^{k-1})p(a_k|a_2^{k-1})$$

Leiten Sie die Formel zur Berechnung des Backoff-Faktors $\alpha(a_1^{k-1})$ her. (3 Punkte)

**Aufgabe 3)** Geben Sie die Formeln zur Berechnung des besten Parsebaumes aus einer Menge von Parsebäumen in Parsewaldrepräsentation an. Wie wird initialisiert? Wie wird des beste Parse am Ende ausgegeben?

Sie können davon ausgehen, dass $w_1, ..., w_n$ die Wörter des Eingabesatzes sind, N die Menge der nichtterminalen Knoten im Parsewald und P die Menge der Parsewaldregeln. Die Funktion $p(r)$ liefert die Wahrscheinlichkeit der PCFG-Regel, welche der Parsewaldregel $r \in P$ entspricht. (5 Punkte)

**Aufgabe 4)** Warum benötigt man bei der Anwendung von statistischen Verfahren (und anderen Lernverfahren) in der Regel drei Teilmengen von Daten? Wie werden Sie üblicherweise benannt und wozu dient jede der drei Teilmengen? (2 Punkte)

**Aufgabe 5)** Wie ist bei einem Trigramm-Sprachmodell die Wahrscheinlichkeit der Wortfolge $w_1, ..., w_n$ definiert? Erklären Sie welche Besonderheiten am Satzanfang und Satzende zu beachten sind. (3 Punkte)

**Aufgabe 6)** Schreiben Sie eine Programmfunktion "viterbi", welche die wahrscheinlichste Tagfolge für eine gegebene Wortfolge gemäß einem Bigramm-HMM berechnet. Der Funktion wird ein Array mit der Liste der Wörter als Argument übergeben.

Sie können annehmen, dass eine Funktion "lookup" bereits existiert, welche ein Wort als Argument nimmt und die Liste der Tags und ihre lexikalischen Wahrscheinlichkeiten $p(Wort|Tag)$ in einer von Ihnen gewünschten Datenstruktur zurückgibt. Sie können außerdem annehmen, dass eine Funktion "context_prob" existiert, welche zwei Wortart-Tags $t_1$ und $t_2$ als Argument nimmt und die Wahrscheinlichkeit $p(t_2|t_1)$ zurückliefert.

Überlegen Sie sich geeignete Datenstrukturen und schreiben Sie dann die Funktion. Sie können die Funktion in Perl, Python, C++ oder Java implementieren.

(12 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!