Schriftliche Prüfung
Seminar Statistische Sprachverarbeitung
WS 2014/15
Helmut Schmid

Aufgabe 1) Wie lautet das Zipf'sche Gesetz und was sagt es aus? (2 Punkte)

Aufgabe 2) Wie ist die Wahrscheinlichkeit einer Wortfolge bei einem Markowmodell 2. Ordung (Trigrammmodell) definiert? Geben Sie die Formel an. Welche Unabhängigkeitsannahmen macht das Modell? Was ist wichtig, damit das Markowmodell eine wohldefinierte Wahrscheinlichkeitsverteilung über alle möglichen Wortfolgen liefert? (4 Punkte)

Aufgabe 3) Erklären Sie, wie man auf Basis von Markowmodellen einen Sprachidentifizierer realisieren kann. Welche Daten braucht man dazu? (3 Punkte)

Aufgabe 4) Eine geglättete Wahrscheinlichkeitsverteilung sei wie folgt definiert:

$$p(w|w') = p^*(w|w') + \alpha(w')p(w) \quad \text{mit } p^*(w|w') = \frac{f(w',w) - \delta}{\sum_{w'} f(w',w')}$$

Leiten Sie die Formel für die Berechnung des Backoff-Faktors $\alpha(w')$ her. (3 Punkte)

Aufgabe 5) Erklären Sie detailliert und mit Formeln, wie ein Bigramm-Tagger auf Basis von HMMs effizient die beste Tagfolge berechnet. (4 Punkte)

Aufgabe 6) Erklären Sie, wie Sie berechnen können, ob der Performanzunterschied zwischen zwei Taggern statistisch signifikant ist oder nicht. Beschreiben Sie das genaue Vorgehen. (3 Punkte)

Aufgabe 7) Erklären Sie, wie der Perzeptron-Algorithmus die Merkmalsgewichte eines linearen Modelles lernt. Formulieren Sie den Perzeptronalgorithmus zusätzlich als Pseudo-Code. Welche Ideen kennen Sie, mit denen der Perzeptronalgorithmus verbessert werden kann? (5 Punkte)

Aufgabe 8) Ein Naive-Bayes-Modell soll benutzt werden, um einen Zeitungsartikel einem Themengebiet (Politik, Wirtschaft etc.) zuzuweisen. Wie trainieren Sie das Modell? Welche Art von Daten benötigen Sie dazu? Wie berechnen Sie die wahrscheinlichste Kategorie eines Artikels (mit Formel)?

Anmerkung: Wir haben diese Anwendung nicht in der Vorlesung besprochen. Sie müssen hier Ihr Wissen auf eine neue Anwendung übertragen. (4 Punkte)

Aufgabe 9) Erklären Sie die Grundidee der Berkeley-Parsers. Mit welcher Methode wird er trainiert? (2 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!