Schriftliche Wiederholungsprüfung zur
Vorlesung Statistische Sprachverarbeitung
WS 2015/16
Dozent: Helmut Schmid

Aufgabe 1) Geben Sie an, wie bei einem Bigramm-HMM-Tagger die Wahrscheinlichkeit einer Wortfolge $w1, ..., w_n$ mit den Tags $t_1, ..., t_n$ definiert ist. Was ist an den Satzgrenzen zu beachten? (3 Punkte)

Aufgabe 2) Wie lauten die Formeln für die Berechnung der Wahrscheinlichkeiten $p(t_3|t_1, t_2)$, wenn Sie eine Maximum-Likelihood-Schätzung (= relative Häufigkeiten) verwenden? (2 Punkte)

Aufgabe 3) Wie lautet die Formel zur Berechnung der Wahrscheinlichkeit $p(t_3|t_1, t_2)$, wenn Sie eine interpolierte Backoff-Schätzung mit Absolute Discounting verwenden? (3 Punkte)

Aufgabe 4) Angenommen der Satz "I can can a can" soll mit einem Bigramm-Tagger und dem Viterbi-Algorithmus mit Wortarten annotiert werden und das Tagset umfasst nur die Tags PRO, D, N, und V. Bei jedem Wort soll jedes Tag erlaubt sein.

Wie wird die Tabelle mit den Viterbiwahrscheinlichkeiten zu Beginn initialisiert?

Wie berechnet der Algorithmus die Viterbi-Wahrscheinlichkeit des Tags D an Position 4, also beim Wort "a"? (Geben Sie den richtigen Ausdruck für diesen konkreten Fall ohne Variablen an.) (3 Punkte)

Aufgabe 5) Ein Naive-Bayes-Modell kann verwendet werden, um die wahrscheinlichste Bedeutung $\hat{s}$ eines ambigen Wortes gegeben die Folge seiner Nachbarwörter $w_1...w_n$ zu bestimmen. Erklären Sie für jeden Schritt der folgenden Herleitung des Modelles, wie er begründet werden kann. Welche Theoreme und Definitionen werden angewandt? Welche Unabhängigkeitsannahmen werden gemacht?

$$
\begin{aligned}
\hat{s} &= \arg\max_s p(s|w_1, ..., w_n) && (1)\\
&= \arg\max_s \frac{p(s)\ p(w_1, ..., w_n|s)}{p(w_1, ..., w_n)} && (2)\\
&= \arg\max_s p(s)\ p(w_1, ..., w_n|s) && (3)\\
&= \arg\max_s p(s) \prod_{i=1}^{n} p(w_i|s, w_1, ..., w_{i-1}) && (4)\\
&= \arg\max_s p(s) \prod_{i=1}^{n} p(w_i|s) && (5)\\
&= \arg\max_s \prod_{i=1}^{n} p(w_i|s) && (6)
\end{aligned}
$$

(4 Punkte)

Aufgabe 6) Erklären Sie, wie eine Pseudowort-Evaluierung funktioniert und wofür man Sie verwenden kann. (3 Punkte)

Aufgabe 7) Erklären Sie detailliert, wie Sie mit dem Binomialtest berechnen, ob ein Tagger signifikant besser ist als ein Vergleichstagger. (3 Punkte)

Aufgabe 8) Die Formel für die Berechnung des χ^2-Testes lautet:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Was verbirgt sich hinter den O_{ij} und E_{ij}? Wozu kann man die Formel benutzen? Was ist eine Kontingenztabelle und wie sieht sie aus? (3 Punkte)

Aufgabe 9) Wie funktioniert die Addiere-1-Glättung und warum ist sie nicht geeignet, um Wortwahrscheinlichkeiten zu glätten? Wann ist die Addiere-1-Glättung optimal? (3 Punkte)

Aufgabe 10) Erklären Sie das EM-Training am Beispiel des unüberwachten Trainings von Wortart-Taggern. Was ist das Grundprinzip? Welche Schritte umfasst das Verfahren? Was sollte gegeben sein? Mit welchem Algorithmus kann das EM-Training bei einem HMM-Tagger effizient implementiert werden?

(3 Punkte)

(30 Punkte insgesamt)