Schriftliche Wiederholungsprüfung zur Übung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2017/18
Dozent: Helmut Schmid

Aufgabe 1) In einem Korpus mit 500 000 Adjektiv-Nomen-Paaren taucht das Adjektiv **bare** 18-mal, das Nomen **Münze** 43-mal und das Paar **bare Münze** 14-mal auf.

Wie berechnen Sie Schätzwerte für die Wahrscheinlichkeiten p(bare), p(Münze), p(bare, Münze), p_1(bare|Münze), p_2(Münze|bare), wobei p_1(x|y) die Wahrscheinlichkeit von x **vor** y und p_2(x|y) die Wahrscheinlichkeit von x **nach** y ist. (2 Punkte)

Aufgabe 2) Wie können Sie zeigen, dass das Wortpaar **bare Münze** in Aufgabe 1 signifikant häufiger ist, als bei statistischer Unabhängigkeit der Wörter zu erwarten wäre? Was müssen Sie hierfür konkret berechnen? (2 Punkte)

Aufgabe 3) Zeigen Sie durch Umformen, dass die Formel

$$p(w) = \mu \frac{f(w)}{N} + (1-\mu)\frac{1}{B} \quad \text{mit } \mu = \frac{N}{N+B\lambda}$$

äquivalent ist zur Addiere-λ-Glättungs-Formel

$$p(w) = \frac{f(w)+\lambda}{N+B\lambda}$$

(2 Punkte)

Aufgabe 4) Wie lautet die Formel für die Berechnung der Entropie $H(X)$ einer Zufallsvariablen X?

Wie hoch ist die Entropie einer Zufallsvariablen mit einer uniformen Verteilung über 4 Werte (also wenn alle Werte die gleiche Wahrscheinlichkeit haben)? Wie hoch ist die Entropie, wenn einer der Werte die Wahrscheinlichkeit 1 besitzt? (2 Punkte)

Aufgabe 5) Zeigen Sie, dass der Backoff-Faktor bei der interpolierten Backoffglättung

$$p(y|x) = \phi(y|x) + \alpha(x)p(y) \quad \text{mit } \phi(y|x) = max(0, \frac{f(x,y)-\delta}{f(x)})$$

gegeben ist durch

$$\alpha(x) = 1 - \sum_y \phi(y|x)$$

Startpunkt: Da $p(y|x)$ eine Wahrscheinlichkeitsverteilung ist, muss für alle x gelten:

$$\sum_y p(y|x) = 1$$

Setzen Sie in diese Gleichung die Definition von $p(y|x)$ ein und lösen Sie die Gleichung nach $\alpha(x)$ auf. (2 Punkte)

Aufgabe 6) Angenommen Sie haben das Wort w_1 10-mal in einem Korpus der Größe 10 000 gesehen. In welchem Bereich liegt die erwartete Häufigkeit desselben Wortes in einem neuen Korpus derselben Größe aus derselben Quelle? Zur Auswahl stehen 0-9 Mal, 9-10 Mal, 10-11 Mal, mehr als 11 Mal. (1 Punkt)

>>>>>>>>>>>>>> weiter auf der nächsten Seite >>>>>>>>>>>>>>>

Aufgabe 7) Bei der Wortbedeutungsdesambiguierung geht es darum, die wahrscheinlichste Bedeutung s eines ambigen Wortes w auf Basis der Kontextwörter $C = w_1^n = w_1, ..., w_n =$ zu berechnen, also

$$\arg\max_s p(s|w_1^n)$$

Zeigen Sie, wie man durch Umformen dieses Ausdrucks die Naive-Bayes-Formel herleitet:

$$\arg\max_s p(s) \prod_{i=1}^{n} p(w_i|s)$$

Welche vereinfachenden Annahmen müssen Sie dabei machen? (2 Punkte)

Aufgabe 8) Wie wird bei einer probabilistischen kontextfreien Grammatik (PCFG) die Wahrscheinlichkeit eines Parsebaumes/Satzes/Korpus definiert?

(Ein Parsebaum wird repräsentiert durch die Folge der Linksableitungs-Regeln $r_1, ..., r_n$ mit den Regelwahrscheinlichkeiten $p(r_i)$.) (2 Punkte)

Aufgabe 9) Programmieraufgabe 1

Sie sollen ein Programm schreiben, das ein einfaches Hidden-Markow-Modell auf Trainingsdaten trainiert. Die Daten befinden sich in der Datei **data**, wobei jede Zeile genau einen Satz enthält. Die Zeile beginnt mit der Wortfolge. Dann folgt ein Tabulatorzeichen und dann die Tagfolge. Einzelne Wörter (bzw. Tags) sind durch Leerzeichen getrennt.

Schritte:

- Lesen Sie die Datei Zeile für Zeile ein. Extrahieren Sie die Häufigkeiten aller Wort-Tag-Paare, aller Wörter, aller Tagpaare und aller (Vorgänger-)Tags. Für die Extraktion der Tagpaar-Häufigkeiten müssen Sie Grenztags hinzufügen.
- Schätzen Sie die Kontext-Wahrscheinlichkeiten $p(t|t')$ und die lexikalischen Wahrscheinlichkeiten $p(w|t)$ mit relativen Häufigkeiten (ohne Glättung).

(8 Punkte)

Aufgabe 10) Programmieraufgabe 2

Schreiben Sie eine Funktion **forward(words)**, welche eine Wortfolge (ohne Endesymbol) als Argument erhält und den Forward-Algorithmus für einen Bigramm-Tagger implementiert und die Tabelle mit den Forward-Wahrscheinlichkeiten zurückgibt.

$$\begin{aligned} \alpha_t(0) &= \begin{cases} 1 & \text{falls t das Startsymbol ist} \\ 0 & \text{sonst} \end{cases} \\ \alpha_t(i) &= \sum_{t'} \alpha_{t'}(i-1)\ p(t|t')\ p(w_i|t) \text{ für } 1 \leq i \leq n+1 \end{aligned}$$

Die Funktion besteht aus drei Schleifen über die Wortpositionen, Tags und Vorgängertags. Gegeben ist die Funktion **contextprob(t1,t2)**, welche die Wahrscheinlichkeit von *t2* gegeben *t1* zurückgibt, und die Funktion **lexprobs(w)**, welche ein Dictionary zurückliefert, welches die möglichen Tags des Wortes w als Keys und die Wahrscheinlichkeiten $p(w|tag)$ als Values hat.

(Vergessen Sie nicht das Endesymbol. Die Funktion **lexprobs(.)** liefert für das Endesymbol das Endetag mit der Wahrscheinlichkeit 1.) (7 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!