Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

Sommersemester 2018, Prof. Dr. Udo Hahn, Tobias Kolditz
Übungsblatt 2 vom 14.05.2018
Abgabe bis 21.05.2018 per E-Mail (PDF) an
tobias.kolditz@uni-jena.de

Aufgabe 1 : "Recap" – Reguläre Sprachen **5 p**

Gegeben sei eine reguläre Sprache $L = \{a^n bc^m \mid n \in \mathbb{N}_0, m \in \mathbb{N}^+\}$.

a) Regulärer Ausdruck 1 p

Geben Sie einen regulären Ausdruck an, der L **genau** beschreibt.

b) Endlicher Automat 2 p

Geben Sie einen (möglichst einfachen) endlichen Automaten **ohne** Epsilon-Übergänge an, der alle Wörter aus L akzeptiert. Hier reicht das entsprechende Diagramm, wie in der Übung (achten Sie auf die Kennzeichnung von Anfangszustand und finalem Zustand/finalen Zuständen).

c) Reguläre Grammatik 2 p

Geben Sie eine (möglichst einfache) reguläre Grammatik an, welche L generiert. Es reicht eine Liste von Regeln mit gesondert markiertem Startsymbol. Epsilon-Regeln sind erlaubt.

Aufgabe 2 : Abschätzung von Wahrscheinlichkeiten durch relative Häufigkeiten **5 pt**

Gegeben sei ein Korpus in Form einer Liste von N_T Tokens. Schreiben Sie in Pseudocode einen Algorithmus, der für zwei beliebige Wörter a und b aus dem Vokabular des Korpus die konditionale Wahrscheinlichkeit $P(b|a)$[1] berechnet. Die Wahrscheinlichkeit soll dabei allein auf im Korpus beobachteten Häufigkeiten beruhen:

$$P(b|a) = \frac{C_2(ab)}{\sum_w C_2(aw)} \tag{1}$$

wobei $C_2(ab)$ für die absolute Häufigkeit des Bigramms ab im gegebenen Korpus steht. Wenn wir unter $C_1(a)$ die abolute Häufigkeit des Unigramms a in den ersten $N_T - 1$ Tokens des Korpus verstehen, können wir den Nenner folgendermaßen vereinfachen:

$$P(b|a) = \frac{C_2(ab)}{C_1(a)} \tag{2}$$

a) Unigramm-Häufigkeiten 1 pt

Schreiben Sie eine Funktion, welche die absoluten Häufigkeiten C_1 der in den ersten $N_T - 1$ Tokens des Korpus beobachteten Unigramme ausgibt.

b) Bigramm-Häufigkeiten 1 pt

Schreiben Sie eine Funktion, welche die absoluten Häufigkeiten C_2 der im gesamten Korpus beobachteten Bigramme ausgibt.

c) Konditionale Wahrscheinlichkeiten 3 pt

Nutzen Sie die Funktionen aus den ersten beiden Teilaufgaben, um in einer weiteren Funktion die konditionale Wahrscheinlichkeit $P(b|a)$ nach Gleichung 2 zu berechnen.

[1] Die Wahrscheinlichkeit, Wort b zu sehen, unter der Bedingung, dass wir zuletzt Wort a gesehen haben.