Einführung in die Computerlinguistik

Hausaufgabe (CFG 3), Abgabe 14.06.2021, 25 Pkte

Laura Kallmeyer

Sommer 2021, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1

1. Geben Sie für die Sprache $\{a^n w w^R a^n \mid n \geq 0, w \in \{b, c\}^+\}$ eine CFG in Chomsky Normalform an.

2. Betrachten Sie folgende CFG:

$$G = \langle \{S, B\}, \{a, b\}, \{S \rightarrow aSBS \mid a, B \rightarrow bB \mid \varepsilon\}, S \rangle$$

Geben Sie eine äquivalente Grammatik in Chomsky Normalform an.

Tipp: Sie müssen G zunächst in eine äquivalente CFG ohne ε -Produktionen transformieren, anschließend kommt die eigentliche Umwandlung in CNF.

Lösung: (hier nur Produktionen, es sollte aber eigentlich das ganze Tupel hingeschrieben werden)

1. $S \rightarrow AT \mid BU \mid CV \mid BB \mid CC, T \rightarrow SA, U \rightarrow XB, V \rightarrow XC, X \rightarrow BU \mid CV \mid BB \mid CC,$
 $A \rightarrow a, B \rightarrow b, C \rightarrow c$

2. Nach Entfernen der ε -Produktion:

$$S \rightarrow aSBS \mid aSS \mid a, B \rightarrow bB \mid b$$

Nach Hinzufügen von Präterminalen:

$$S \rightarrow C_a SBS \mid C_a SS \mid a, B \rightarrow C_b B \mid b, C_a \rightarrow a, C_b \rightarrow b$$

Nach Binarisierung:

$$S \rightarrow C_a X \mid C_a Y \mid a, X \rightarrow SZ, Y \rightarrow SS, Z \rightarrow BS, B \rightarrow C_b B \mid b, C_a \rightarrow a, C_b \rightarrow b$$

Aufgabe 2

Aus der Vorlesung wissen Sie, dass $L_1 = \{a^n b^n c^n \mid n \geq 1\}$ nicht kontextfrei ist (s. Folie 18).

1. Zeigen Sie unter Bezugnahme auf die Abgeschlossenheit von CFLs unter Homomorphismen, dass auch $L_2 = \{(ae)^n b^n g^m c^n \mid n \geq 1, m > 2\}$ nicht kontextfrei ist.

2. Zeigen Sie, unter Bezugnahme auf die Abgeschlossenheit von CFLs unter Schnittbildung mit regulären Sprachen, dass auch $L_3 = \{x^n y^n z^n \mid x, y, z \in \{a, b, c\}\}$ nicht kontextfrei ist.

Tipp: Die Beweisführung in allen drei Fällen ist ein Widerspruchsbeweis: Man nimmt das Gegenteil von dem an, was man beweisen möchte, und zeigt, dass dies zu einem Widerspruch führt.

Konkret: Man nimmt an, dass die jeweilige Sprache L_i kontextfrei ist. Dann muss auch ihr Bild unter einem geeigneten Homomorphismus (bzw. ihr Schnitt mit einer geeigneten regulären Sprache) kontextfrei sein. Hier muss man Homomorphismus bzw. reguläre Sprache so wählen, dass man bei einer Sprache landet, von der man schon gezeigt hat, dass sie nicht kontextfrei ist. Dies ist ein Widerspruch zu dem, was aus der Annahme der Kontextfreiheit von L_i folgt, und somit ist diese Annahme falsch.

Lösung:

1. Annahme: L_2 ist kontextfrei. Dann ist auch $f(L_2)$ kontextfrei, wobei f ein Homomorphismus ist mit $f(a) = a, f(b) = b, f(c) = c, f(e) = \varepsilon, f(g) = \varepsilon$. $f(L_2) = L_1$. Wir wissen aber schon, dass L_1 nicht kontextfrei ist, es ergibt sich also ein Widerspruch. Damit kann L_2 auch nicht kontextfrei sein.

2. Annahme: L_3 ist kontextfrei. Dann ist auch der Schnitt von L_3 mit $L(a^+b^+c^+)$, also mit der von $a^+b^+c^+$ denotierten regulären Sprache, kontextfrei. $L_3 \cap L(a^+b^+c^+) = L_1$. Wir wissen aber schon, dass L_1 nicht kontextfrei ist, es ergibt sich also ein Widerspruch. Damit kann L_3 auch nicht kontextfrei sein.

Aufgabe 3

Betrachten Sie die folgende Sprache: $\{c^nabad^{n+1} \mid n \geq 2\}$. Es handelt sich um eine kontextfreie Sprache. Damit gilt das Pumping Lemma von Folie 17 für diese Sprache.

1. Wie ist das kleinstmögliche k , so dass es ein Wort der Länge k gibt und das Pumping Lemma für die Konstante k und diese Sprache gilt?
Tipp: Bedenken Sie, dass auch eine Iteration mit $i = 0$, also ein Weglassen der beiden iterierbaren Teile, zu einem Wort in der Sprache führen sollte.
2. Welches Wort w der Länge k ist dann das kürzeste, in dem man zwei Teilstrings v_1 und v_2 findet, die wie im Pumping Lemma parallel iteriert werden können (bei dieser Sprache gibt es nur eines), und welche Teilstrings sind v_1 und v_2 ? Geben Sie alle Möglichkeiten für v_1 , v_2 an, wobei jeweils spezifiziert werden soll, welche Teilstrings von w v_1 und v_2 sind (es gibt hier mehrere Möglichkeiten).

Lösung:

1. $k = 10$
2. $w = cccabadddd$ mit $v_1 = c$, $v_2 = d$. Es gibt 12 verschiedene Möglichkeiten bzgl. der jeweiligen Teilstrings, denn v_1 kann entweder das erste, das zweite oder das dritte c sein und v_2 kann das erste, das zweite, das dritte oder das vierte d sein.