Chomsky Normalform

Wolfgang Mulzer

Wir illustrieren den Algorithmus, der eine kontextfreie Grammatik in Chomsky Normalform überführt. Ausgangspunkt ist die Grammatik mit $V = \{S, A, B, C, D\}$, $\Sigma = \{a, b\}$ und den folgenden Produktionen:

$$\begin{aligned}
S &\to aAa \mid bBb \\
A &\to C \mid a \\
B &\to C \mid b \\
C &\to CD \mid \varepsilon \\
D &\to A \mid B \mid ab
\end{aligned}$$

Schritt 1: Trennen der Terminalsymbole. Für jedes Terminalsymbol $\sigma \in \Sigma$ führen wir eine neue Variable V_σ ein und ersetzen jedes Vorkommen von σ durch V_σ. Danach fügen wir für jedes $\sigma \in \Sigma$ die neue Produktion $V_\sigma \to \sigma$ hinzu.

$$\begin{aligned}
S &\to V_a A V_a \mid V_b B V_b \\
A &\to C \mid V_a \\
B &\to C \mid V_b \\
C &\to CD \mid \varepsilon \\
D &\to A \mid B \mid V_a V_b \\
V_a &\to a \qquad V_b \to b
\end{aligned}$$

Schritt 2: Auflösen von langen Produktionen. Für jede Produktion der Form $X \to Y_1 Y_2 \dots Y_k$ mit $k \geq 3$ führen wir neue Variablen $Z_2, Z_3, \dots, Z_{k-1}$ ein und ersetzen $X \to Y_1 Y_2 \dots Y_k$ durch $X \to Y_1 Z_2$, $Z_2 \to Y_2 Z_3$, $\dots$, $Z_{k-1} \to Y_{k-1} Y_k$.

$$\begin{aligned}
S &\to V_a E \mid V_b F \\
A &\to C \mid V_a \\
B &\to C \mid V_b \\
C &\to CD \mid \varepsilon \\
D &\to A \mid B \mid V_a V_b \\
E &\to A V_a \\
F &\to B V_b \\
V_a &\to a \qquad V_b \to b
\end{aligned}$$

Schritt 3: Beseitigen der ε-Produktionen. Zunächst bestimmen wir alle Variablen, aus denen sich ε Ableiten lässt. Dazu markieren wir zunächst alle Variablen X, für die es eine Regel $X \to \varepsilon$ gibt. In unserem Beispiel wird C markiert.

Dann wiederholen wir die folgende Regel, solange es geht: wenn eine unmarkierte Variable X existiert, so dass auf der rechten Seite einer Produktion $X \to \alpha$ nur markierte Variablen vorkommen, markiere X.

In unserem Beispiel markieren wir A und B (wegen $A \to C$ und $B \to C$) und dann D (wegen $D \to A$). Zum Schluss sind A, B, C und D markiert, und das sind genau die Variablen, aus denen man ε ableiten kann.

Schließlich betrachten wir jede Produktion der Form $X \to YZ$. Wenn Y markiert ist, fügen wir die Produktion $X \to Z$ hinzu, wenn Z markiert ist, fügen wir die Produktion $X \to Y$ hinzu. Dann löschen wir die Produktionen $X \to \varepsilon$.

$$
\begin{aligned}
S &\to V_aE \mid V_bF \\
A &\to C \mid V_a \\
B &\to C \mid V_b \\
C &\to CD \mid C \mid D \\
D &\to A \mid B \mid V_aV_b \\
E &\to AV_a \mid V_a \\
F &\to BV_b \mid V_b \\
V_a &\to a \qquad V_b \to b
\end{aligned}
$$

Schritt 4: Beseitigen der Kettenregeln. Zunächst müssen wir Kreise der Form $X_1 \to X_2 \to X_3 \to \cdots \to X_1$ beseitigen. Die aktuelle Grammatik enthält den Kreis $A \to C \to D \to B \to C \to D \to A$. Wir führen eine neue Variable G ein und ersetzen alle Vorkommen von A, B, C, D durch G.

$$
\begin{aligned}
S &\to V_aE \mid V_bF \\
G &\to G \mid V_a \\
G &\to G \mid V_b \\
G &\to GG \mid G \mid G \\
G &\to G \mid G \mid V_aV_b \\
E &\to GV_a \mid V_a \\
F &\to GV_b \mid V_b \\
V_a &\to a \qquad V_b \to b
\end{aligned}
$$

Dann beseitigen wir alle Regeln der Form $G \to G$.

$$
\begin{aligned}
S &\to V_aE \mid V_bF \\
G &\to GG \mid V_a \mid V_b \mid V_aV_b \\
E &\to GV_a \mid V_a \\
F &\to GV_b \mid V_b \\
V_a &\to a \qquad V_b \to b
\end{aligned}
$$

Jetzt enthält die Grammatik keine Kreise mehr. Wir ordnen die Variablen so, dass bei jeder Kettenregel $X \to Y$ die Variable X vor der Variablen Y kommt. Eine solche Ordnung ist zum Beispiel S, G, E, F, V_a, V_b. (Eine andere gültige Ordnung wäre S, F, E, G, V_a, V_b, aber nicht S, E, G, V_a, V_b, F, da es die Regel $F \to V_b$ gibt).

Dann gehen wir die Variablen in umgekehrter Reihenfolge durch. Für jede Variable X ersetzen wir jede Kettenregel $X \to Y$ durch neue Produktionen der Form $X \to \alpha$, für alle Produktionen $Y \to \alpha$.

In unserem Beispiel tun wir also folgendes: Wir betrachten zunächst V_b. Es ist nichts zu tun, weil es keine Produktion der Form $V_b \to X$, X Variable, gibt.

Dann betrachten wir V_a, wo es genauso aussieht.

Dann betrachten wir F. Es gibt die Produktion $F \to V_b$. Wir ersetzen sie durch $F \to b$, da es die Produktion $V_b \to b$ gibt.

Dann kommt E, wo wir $E \to V_a$ durch $E \to a$ ersetzen.

Danach ist G an der Reihe. Wir ersetzen $G \to V_a$ durch $G \to a$ und $G \to V_b$ durch $G \to b$.

Zum Schluss kommt S, wo nichts zu tun ist.

Die resultierende Grammatik ist in Chomsky Normalform:

$$
\begin{aligned}
S &\to V_aE \mid V_bF \\
G &\to GG \mid a \mid b \mid V_aV_b \\
E &\to GV_a \mid a \\
F &\to GV_b \mid b \\
V_a &\to a \qquad V_b \to b
\end{aligned}
$$