
Grundlagen der theoretischen Informatik

Kurt Sieber

Fachbereich Mathematik/Theoretische Informatik
Universität Siegen

Vorlesung vom 30.11.2004 (Stand: 3.12.2004)

Kontextfreie Sprachen

Bevor wir $L(M) = L(G)$ beweisen, wollen wir uns die Arbeitsweise dieses Kellerautomaten M an einem kleinen **Beispiel** verdeutlichen:

Sei $G = (\{a, b\}, \{S\}, S, P)$ mit $P = \{S \rightarrow aSb, S \rightarrow \varepsilon\}$.

Dann ist $M = (\{a, b\}, \{a, b, S\}, \{s, f\}, s, \Delta)$

mit

$\Delta = \{ ((s, \varepsilon, \varepsilon), (f, S)),$

$((f, \varepsilon, S), (f, aSb)),$

$((f, \varepsilon, S), (f, \varepsilon)),$

$((f, a, a), (f, \varepsilon)),$

$((f, b, b), (f, \varepsilon))\}$

also gilt z.B. bei Eingabe $aabb$:

$(s, aabb, \varepsilon) \vdash_M (f, aabb, S)$

$\vdash_M (f, aabb, aSb)$

$\vdash_M (f, abb, Sb)$

$\vdash_M (f, abb, aSbb)$

$\vdash_M (f, bb, Sbb)$

$\vdash_M (f, bb, bb)$

$\vdash_M (f, b, b)$

$\vdash_M (f, \varepsilon, \varepsilon)$

Kontextfreie Sprachen

Am Beispiel sieht man, dass in einer erfolgreichen Berechnung des Kellerautomaten das aktuelle Eingabewort stets aus dem aktuellen Kellerwort ableitbar ist. Diese Beobachtung bringen wir durch folgende Äquivalenz zum Ausdruck: Für alle $\alpha \in \Gamma^*$ und $w \in \Sigma^*$ gilt

$$(f, w, \alpha) \vdash_M^* (f, \varepsilon, \varepsilon) \Leftrightarrow \alpha \xRightarrow{*}_G w \quad (*)$$

Aus (*) folgt $L(M) = L(G)$, denn:

$$w \in L(M) \Leftrightarrow (s, w, \varepsilon) \vdash_M^* (f, \varepsilon, \varepsilon)$$

$$\Leftrightarrow (f, w, S) \vdash_M^* (f, \varepsilon, \varepsilon)$$

weil als erster Übergangsschritt nur

$$(s, w, \varepsilon) \vdash_M (f, w, S) \text{ in Frage kommt}$$

$$\Leftrightarrow S \xRightarrow{*}_G w$$

wegen (*) mit $\alpha = S$

$$\Leftrightarrow w \in L(G)$$

Kontextfreie Sprachen

Es bleibt also (*) zu beweisen

$$(f, w, \alpha) \vdash_M^* (f, \varepsilon, \varepsilon) \Leftrightarrow \alpha \xrightarrow{*}_G w \quad (*)$$

' \Rightarrow ': Induktion über die Anzahl n der Übergangsschritte in \vdash_M^*

$n = 0$, d.h. $w = \alpha = \varepsilon$:

Dann gilt $\alpha \xrightarrow{*}_G w$.

$n > 0$, d.h. $(f, w, \alpha) \vdash_M (f, v, \beta) \vdash_M^{n-1} (f, \varepsilon, \varepsilon)$ mit $v \in \Sigma^*$ und $\beta \in \Gamma^*$

Dann gilt nach Induktionsannahme $\beta \xrightarrow{*}_G v$.

Wenn der erste Übergangsschritt mit (3) erfolgt, so existiert ein $a \in \Sigma$ mit $w = av$ und $\alpha = a\beta$, also gilt $\alpha = a\beta \xrightarrow{*}_G av = w$.

Wenn der erste Übergangsschritt mit (2) erfolgt, so gilt $w = v$ und β entsteht aus α durch Anwendung einer Produktion $(A \rightarrow \gamma) \in P$.
Dann gilt $\alpha \xrightarrow{*}_G \beta \xrightarrow{*}_G v = w$.

Kontextfreie Sprachen

$$(f, w, \alpha) \vdash_M^* (f, \varepsilon, \varepsilon) \Leftrightarrow \alpha \xRightarrow{*}_G w \quad (*)$$

‘ \Leftarrow ’: Induktion über die Länge n einer Linksableitung $\alpha \xRightarrow{n}_G w$

$n = 0$, d.h. $\alpha = w \in \Sigma^*$:

Dann gilt $(f, w, \alpha) = (f, w, w) \vdash_M^{|w|} (f, \varepsilon, \varepsilon)$ mit (3).

$n > 0$, d.h. es existiert ein $\alpha_1 \in \Gamma^*$ mit $\alpha \Rightarrow_G \alpha_1 \xRightarrow{n-1}_G w$

Sei $A \rightarrow \gamma$ die Produktion im Linksableitungsschritt $\alpha \Rightarrow_G \alpha_1$, d.h. es existieren $u \in \Sigma^*, \beta \in \Gamma^*$ mit $\alpha = uA\beta$ und $\alpha_1 = u\gamma\beta$. Wegen $u \in \Sigma^*$ existiert dann ein $v \in \Sigma^*$ mit $w = uv$ und $\gamma\beta \xRightarrow{n-1}_G v$, also

$$(f, w, \alpha) = (f, uv, uA\beta) \vdash_M^{|u|} (f, v, A\beta) \text{ mit (3)}$$

$$\vdash_M (f, v, \gamma\beta) \text{ mit (2)}$$

$$\vdash_M^* (f, \varepsilon, \varepsilon) \text{ nach Induktionsannahme } \square$$

Kontextfreie Sprachen

Die Umkehrung von Satz 2.19 gilt ebenfalls.

Satz 2.20 *Zu jedem Kellerautomaten M lässt sich eine kontextfreie Grammatik G konstruieren mit $L(G) = L(M)$.*

Beweis: s. Literatur

□

Also erhalten wir eine zweite Charakterisierung für die Klasse der kontextfreien Sprachen.

Satz 2.21 *Eine Sprache $L \subseteq \Sigma^*$ ist genau dann kontextfrei, wenn es einen Kellerautomaten M gibt mit $L = L(M)$.*

Beweis: Das folgt unmittelbar aus den Sätzen 2.19 und 2.20.

□

Kontextfreie Sprachen

Abschlusseigenschaften

Satz 2.22 *Die Klasse \mathcal{L}_{kf} der kontextfreien Sprachen ist abgeschlossen unter den Operationen $\cup, \circ, *$ und $+$, d.h. wenn $L_1, L_2 \subseteq \Sigma^*$ kontextfrei sind, dann sind auch die Sprachen*

1. $L_1 \cup L_2$

2. $L_1 \circ L_2$

3. L_1^*

4. L_1^+

kontextfrei. Darüber hinaus gibt es Algorithmen, um Grammatiken für diese Sprachen aus den Grammatiken für L_1 und L_2 zu konstruieren.

Kontextfreie Sprachen

Beweis:

Seien $G_i = (\Sigma, N_i, S_i, P_i)$ ($i = 1, 2$) kontextfreie Grammatiken mit $L(G_i) = L_i$. Wir dürfen annehmen, dass $N_1 \cap N_2 = \emptyset$.

1. siehe Übung 11, Aufgabe 4
2. Sei $G = (\Sigma, N, S, P)$, wobei
 - S ein neues Zeichen ist, d.h. $S \notin N_1 \cup N_2 \cup \Sigma$,
 - $N = N_1 \cup N_2 \cup \{S\}$,
 - $P = \{S \rightarrow S_1 S_2\} \cup P_1 \cup P_2$.

Dann gilt $L(G) = L_1 \circ L_2$.

‘ \supseteq ’: Sei $w \in L_1 \circ L_2$, d.h. $w = w_1 w_2$ mit $w_i \in L_i$ für $i = 1, 2$.

Dann gilt $S_i \xrightarrow{*}_{G_i} w_i$ und damit auch $S_i \xrightarrow{*}_G w_i$ für $i = 1, 2$, und es folgt $S \xrightarrow{*}_G S_1 S_2 \xrightarrow{*}_G w_1 S_2 \xrightarrow{*}_G w_1 w_2 = w$, d.h. $w \in L(G)$.

Kontextfreie Sprachen

' \subseteq ': Sei $w \in L(G)$, d.h. es gibt eine Linksableitung $S \xRightarrow{*}_G w$.

Da S ein neues Zeichen ist, kann der erste Ableitungsschritt nur $S \Rightarrow S_1 S_2$ sein, also hat die gesamte Ableitung die Form $S \Rightarrow S_1 S_2 \xRightarrow{*}_G w_1 S_2 \xRightarrow{*}_G w_1 w_2 = w$, wobei $S_1 \xRightarrow{*}_G w_1$ und $S_2 \xRightarrow{*}_G w_2$.

Wegen $N_1 \cap N_2 = \emptyset$ können in $S_1 \xRightarrow{*}_G w_1$ nur Ableitungsschritte für G_1 vorkommen und in $S_2 \xRightarrow{*}_G w_2$ nur solche für G_2 .

Also gilt $S_1 \xRightarrow{*}_{G_1} w_1$ und $S_2 \xRightarrow{*}_{G_2} w_2$, d.h. $w_1 \in L_1, w_2 \in L_2$ und damit $w \in L_1 \circ L_2$.

3. Sei $G = (\Sigma, N, S, P)$, wobei

$$S \notin N_1 \cup \Sigma,$$

$$N = N_1 \cup \{S\},$$

$$P = \{S \rightarrow S_1 S, S \rightarrow \varepsilon\} \cup P_1.$$

Kontextfreie Sprachen

Dann gilt $L(G) = L_1^*$.

‘ \supseteq ’: Sei $w \in L_1^*$, d.h. es existieren $n \geq 0$ und $w_1, \dots, w_n \in L_1$ mit $w = w_1 \dots w_n$.

Dann gilt $S \Rightarrow S_1 S \Rightarrow \dots \Rightarrow S_1^n S \Rightarrow S_1^n$, und wegen $S_1 \xrightarrow{*}_G w_i$ für $i = 1, \dots, n$ folgt $S \xrightarrow{*}_G w_1 \dots w_n = w$, also $w \in L(G)$.

‘ \subseteq ’: Sei $w \in L(G)$. Durch Induktion über die Länge einer Linksableitung $S \xrightarrow{*}_G w$ beweisen wir $w \in L_1^*$.

Da S ein neues Zeichen ist, kann der erste Ableitungsschritt von $S \xrightarrow{*}_G w$ nur $S \Rightarrow S_1 S$ sein, also hat die gesamte Ableitung die Form $S \Rightarrow S_1 S \xrightarrow{*}_G w_1 S \xrightarrow{*}_G w_1 w' = w$, wobei $S_1 \xrightarrow{*}_G w_1$ und $S \xrightarrow{*}_G w'$.

Da $S_1 \xrightarrow{*}_G w_1$ nur Ableitungsschritte für G_1 enthalten kann, gilt $w_1 \in L(G_1) = L_1$, und nach Induktionssannahme gilt $w' \in L_1^*$.

Also folgt $w = w_1 w' \in L_1 \circ L_1^* \subseteq L_1^*$.

Kontextfreie Sprachen

4. Die Abgeschlossenheit unter \dagger folgt aus der Abgeschlossenheit unter \circ und $*$, weil $L_1^\dagger = L_1 \circ L_1^*$. \square

Es ist kein Zufall, dass Durchschnitt und Komplement in Satz 2.22 fehlen. Wir werden später sehen, dass der Durchschnitt zweier kontextfreier Sprachen im allgemeinen *nicht* kontextfrei ist.

Daraus folgt sofort, dass auch das Komplement einer kontextfreien Sprache im allgemeinen *nicht* kontextfrei ist, denn aus der Abgeschlossenheit unter Vereinigung und Komplement würde sich die Abgeschlossenheit unter Durchschnitt ergeben.

Für den Durchschnitt gilt eine schwächere Aussage, die sich besser mit Automaten (als mit Grammatiken) beweisen lässt:

Kontextfreie Sprachen

Satz 2.23 *Wenn $L_1 \subseteq \Sigma^*$ kontextfrei und $L_2 \subseteq \Sigma^*$ regulär ist, dann ist $L_1 \cap L_2$ kontextfrei.*

Beweis:

Sei $M_1 = (\Sigma, \Gamma, Q_1, s_1, F_1, \Delta_1)$ ein Kellerautomat mit $L(M_1) = L_1$ und sei $A_2 = (\Sigma, Q_2, s_2, F_2, \Delta_2)$ ein NDEA mit $L(A_2) = L_2$.

Wir konstruieren einen Kellerautomaten M mit $L(M) = L_1 \cap L_2$.

Idee: M arbeitet wie M_1 und simuliert “parallel dazu” noch die Übergänge des endlichen Automaten A_2 .

Diese “Parallelverarbeitung” bringt man dadurch zum Ausdruck, dass man auf der Zustandsmenge $Q_1 \times Q_2$ arbeitet: In der ersten Komponente merkt man sich den aktuellen Zustand von M_1 , in der zweiten Komponente den von A_2 .

Kontextfreie Sprachen

Sei also $M = (\Sigma, \Gamma, Q_1 \times Q_2, (s_1, s_2), F_1 \times F_2, \Delta)$ mit

$$\Delta = \{((p_1, p_2), u, \beta), ((q_1, q_2), \gamma) \mid ((p_1, u, \beta), (q_1, \gamma)) \in \Delta_1 \\ \text{und } (p_2, u) \vdash_{A_2}^* (q_2, \varepsilon)\}$$

Δ ist gerade so konstruiert, dass in jedem Übergangsschritt von M ein Übergangsschritt von M_1 und eine entsprechende Folge von Übergangsschritten von A_2 simuliert werden, d.h. es gilt stets:

$$((p_1, p_2), u, \alpha) \vdash_M ((q_1, q_2), \varepsilon, \alpha') \Leftrightarrow (p_1, u, \alpha) \vdash_{M_1} (q_1, \varepsilon, \alpha') \\ \text{und } (p_2, u) \vdash_{A_2}^* (q_2, \varepsilon)$$

Das ergibt sich unmittelbar aus der Definition von Δ und der Definition der möglichen Übergangsschritte eines Kellerautomaten.

Kontextfreie Sprachen

Der gleiche Zusammenhang gilt dann auch für *Folgen* von Übergangsschritten:

$$\begin{aligned} ((p_1, p_2), u, \alpha) \vdash_M^* ((q_1, q_2), \varepsilon, \alpha') &\Leftrightarrow (p_1, u, \alpha) \vdash_{M_1}^* (q_1, \varepsilon, \alpha') \\ &\text{und } (p_2, u) \vdash_{A_2}^* (q_2, \varepsilon) \end{aligned}$$

Das ergibt sich leicht durch Induktion über die Länge der Folge.

Daraus erhält man schließlich das gewünschte Ergebnis:

$$\begin{aligned} w \in L(M) &\Leftrightarrow \text{es existiert ein } (q_1, q_2) \in F \text{ mit} \\ &((s_1, s_2), w, \varepsilon) \vdash_M^* ((q_1, q_2), \varepsilon, \varepsilon) \\ &\Leftrightarrow \text{es existieren } q_1 \in F_1, q_2 \in F_2 \text{ mit} \\ &(s_1, w, \varepsilon) \vdash_{M_1}^* (q_1, \varepsilon, \varepsilon) \\ &\text{und } (s_2, w) \vdash_{A_2}^* (q_2, \varepsilon) \\ &\Leftrightarrow w \in L(M_1) \cap L(A_2) = L_1 \cap L_2 \quad \square \end{aligned}$$

Kontextfreie Sprachen

Entscheidbarkeitsfragen

Wir haben gesehen: Viele Fragestellungen über reguläre Sprachen sind entscheidbar, d.h. es gibt (einfache, manchmal auch effiziente) Algorithmen, die stets die richtige Antwort liefern.

Gilt das auch für kontextfreie Sprachen?

Wichtigste Fragestellung ist das *Wortproblem*:

1. Das *spezielle* Wortproblem für *eine* kontextfreie Grammatik G :

Eingabe: Ein Wort $w \in \Sigma^*$.

Frage: Ist $w \in L(G)$?

2. Das *allgemeine* Wortproblem für kontextfreie Grammatiken:

Eingabe: Eine kontextfreie Grammatik G und ein Wort $w \in \Sigma^*$.

Frage: Ist $w \in L(G)$?

Kontextfreie Sprachen

Für die Praxis ist das spezielle Wortproblem wichtig, denn das ist die Frage, die ein Parser für die Sprache $L(G)$ beantworten muss: Ist die Zeichenreihe, die der Programmierer eingibt, ein syntaktisch korrektes Programm? Diese Frage muss nicht nur entscheidbar sein, sondern sie muss sich *effizient* lösen lassen.

Wir zeigen hier, dass sogar das allgemeine Wortproblem entscheidbar ist (allerdings nicht sehr effizient).

Lösungsansatz:

Um $w \in L(G)$ zu überprüfen, sucht man systematisch nach einer Ableitung für w aus dem Startsymbol S . Man kann z.B. erst alle Wörter $u \in (N \cup \Sigma)^*$ bestimmen, die in einem Schritt aus S ableitbar sind, dann die, die in zwei Schritten ableitbar sind usw.

Wenn w dabei irgendwann auftaucht, ist die Antwort "ja". Aber wann kann man die Antwort "nein" geben, d.h. wann kann man sicher sein, dass w nicht mehr auftaucht?

Kontextfreie Sprachen

Wenn wir wüssten, dass bei jedem Ableitungsschritt mindestens ein Zeichen hinzukommt, dann hätten wir ein Abbruchkriterium: Dann bräuchten wir nur Ableitungen zu untersuchen, die höchstens die Länge $|w|$ haben.

Probleme bereiten also Ableitungsschritte, bei denen das Wort *nicht* länger wird. Solche Ableitungsschritte entstehen, wenn man Produktionen $A \rightarrow \gamma$ mit $|\gamma| \leq 1$ anwendet.

Definition 2.24

- Eine Produktion der Form $A \rightarrow \varepsilon$ (mit $A \in N$) heißt ε -Produktion.
- Eine Produktion der Form $A \rightarrow B$ (mit $A, B \in N$) heißt Einheitsproduktion.

Wir werden zeigen, dass man solche Produktionen (bis auf eine) aus einer kontextfreien Grammatik entfernen kann, ohne dass sich die von der Grammatik erzeugte Sprache verändert.
