Theoretische Informatik 1

WiSe 23/24

Prof. Dr. Sebastian Siebertz
AG Theoretische Informatik
MZH, Raum 3160
siebertz@uni-bremen.de

Universität Bremen

Grammatiken

Definition Grammatik

Eine *Grammatik* ist von der Form $G = (N, \Sigma, P, S)$, wobei

- N und Σ endliche, disjunkte Alphabete von *Nichtterminalsymbolen* bzw. *Terminalsymbolen* sind,
- $S \in N$ das *Startsymbol* ist,
- $P \subseteq (N \cup \Sigma)^+ \times (N \cup \Sigma)^*$ eine endliche Menge von Ersetzungsregeln (*Produktionen*) ist.

- Wir schreiben Produktionen $(u, v) \in P$ gewöhnlich als $u \longrightarrow v$.
- Wir schreiben Elemente von N mit Großbuchstaben und Elemente von Σ mit Kleinbuchstaben.

Weitere Konvention

- Wenn es mehrere Regeln $A \longrightarrow w_1, \ldots, A \longrightarrow w_n$ gibt, so schreiben wir dies in eine Zeile als

$$A \longrightarrow w_1 \mid w_2 \mid \ldots \mid w_n.$$

- Beispiel:
  - $$\begin{aligned} P = \{S &\longrightarrow aSBc, \\ S &\longrightarrow abc, \\ cB &\longrightarrow Bc, \\ bB &\longrightarrow bb\} \end{aligned} \quad \Rightarrow \quad \begin{aligned} P = \{S &\longrightarrow aSBc \mid abc \\ cB &\longrightarrow Bc, \\ bB &\longrightarrow bb\}. \end{aligned}$$

Ableitbarkeit, erzeugte Sprache

Definition Ableitbarkeit, erzeugte Sprache

Sei $G = (N, \Sigma, P, S)$ eine Grammatik und x, y Wörter aus $(N \cup \Sigma)^*$.

1) *y aus x direkt ableitbar:*
   $x \vdash_G y \Leftrightarrow x = x_1 u x_2$ und $y = x_1 v x_2$ mit $u \longrightarrow v \in P$ und $x_1, x_2 \in (N \cup \Sigma)^*$
2) *y aus x in n Schritten ableitbar:*
   $x \vdash_G^n y \Leftrightarrow x \vdash_G x_1 \vdash_G \cdots \vdash_G x_{n-1} \vdash_G y$ für $x_1, \ldots, x_{n-1} \in (N \cup \Sigma)^*$
3) *y aus x ableitbar:*
   $x \vdash_G^* y \Leftrightarrow x \vdash_G^n y$ für ein $n \geq 0$
4) Die durch G *erzeugte Sprache* ist
   $L(G) := \{w \in \Sigma^* \mid S \vdash_G^* w\}$.

Die Chomsky-Hierarchy

Definition Chomsky-Hierarchy

Es sei $G = (N, \Sigma, P, S)$ eine Grammatik.

Typ 0		Jede Grammatik G
Typ 1	monoton	Falls alle Regeln in G *nicht verkürzend* sind, also die Form $w \longrightarrow u$ haben wobei $w, u \in (\Sigma \cup N)^+$ und $\lvert u \rvert \geq \lvert w \rvert$
Typ 2	kontextfrei	Falls alle Regeln in G die Form $A \longrightarrow w$ haben mit $A \in N, w \in (\Sigma \cup N)^*$
Typ 3	rechtslinear	Falls alle Regeln in G die Form $A \longrightarrow uB$ oder $A \longrightarrow u$ haben mit $A, B \in N,\ u \in \Sigma^*$.

Definition Sprachklassen

Für $i = 0, 1, 2, 3$ ist die *Klasse der Typ-i-Sprachen definiert* als

$$\mathcal{L}_i := \{L(G) \mid G \text{ ist Grammatik vom Typ } i\}.$$

Sprachklassen der Chomsky-Hierarchie

Die Sprachtypen bilden eine *strikte Hierarchie*.

Satz: Chomsky-Hierarchie

Es gilt $\mathcal{L}_3 \subsetneqq \mathcal{L}_2 \subsetneqq \mathcal{L}_1 \subsetneqq \mathcal{L}_0$.

Satz: Typ 3 = regulär

Die Typ-3-Sprachen sind genau die regulären Sprachen, d. h.:

$$\mathcal{L}_3 = \{L \mid L \text{ ist regulär}\}$$

Wir haben bereits gezeigt:

Satz

Es gilt $\mathcal{L}_3 \subsetneqq \mathcal{L}_2$.

Vereinfachung kontextfreier Grammatiken

Terminierende und erreichbare Symbole; reduzierte Grammatik

Es sei $G = (N, \Sigma, P, S)$ eine kontextfreie Grammatik.

1) $A \in N$ heißt *terminierend*, falls es ein $w \in \Sigma^*$ gibt mit $A \vdash_G^* w$.
2) $A \in N$ heißt *erreichbar*, falls es $u, v \in (\Sigma \cup N)^*$ gibt mit $S \vdash_G^* uAv$.
3) G heißt *reduziert*, falls alle Elemente von N *erreichbar* und *terminierend* sind.

Lemma

Für jede kontextfreie Grammatik $G = (N, \Sigma, P, S)$ ist die Menge der terminierenden Symbole (in polynomieller Zeit) berechenbar.

Lemma

Für jede kontextfreie Grammatik $G = (N, \Sigma, P, S)$ ist die Menge der erreichbaren Nichtterminalsymbole (polynomieller Zeit) berechenbar.

Leerheitsproblem für kontextfreie Grammatiken

Korollar

Zu jeder kontextfreien Grammatik G mit $L(G) \neq \emptyset$ kann in (polynomieller Zeit) eine äquivalente reduzierte kontextfreie Grammatik konstruiert werden.

Definition: Leerheitsproblem für kontextfreie Grammatiken

- Gegeben: kontextfreie Grammatik G.
- Frage: gilt $L(G) = \emptyset$?

Satz

Das Leerheitsproblem ist für kontextfreie Grammatiken in polynomieller Zeit entscheidbar.

Chomsky-Normalform

Definition

Eine kontextfreie Grammatik (N, Σ, P, S) ist in *Chomsky-Normalform*, wenn alle Ableitungsregeln die folgende Form haben:

- $A \longrightarrow a$ oder $A \longrightarrow BC$ mit $A, B, C \in N,\ a \in \Sigma$
- $S \longrightarrow \varepsilon$ ist erlaubt, wenn es keine Regeln $A \to BC$ mit $S \in \{B, C\}$ gibt.

Vorgehen zum erstellen der CNF:

1. Aufheben der Mischung von Terminalen und Nichtterminalen auf den rechten Seiten von Produktionen. [→ separierte Grammatiken]
2. Aufbrechen langer Wörter auf den rechten Seiten von Produktionen.
3. Eliminieren von Regeln der Form $A \longrightarrow \varepsilon$ (ε-*Regeln*).
4. Eliminieren von Regeln der Form $A \longrightarrow B$ (*Kettenregeln*).

Das Wortproblem für kontextfreie Grammatiken

Definition: Wortproblem für (feste) kontextfreie Grammatik G

- Gegeben: $w \in \Sigma^*$
- Frage: gilt $w \in L(G)$?

- Für Grammatiken in *Chomsky-Normalform* gilt:
  Ableitung eines Wortes $w \in L(G)$ hat *höchstens die Länge* $2|w| + 1$:
  - Produktionen der Form $A \longrightarrow BC$ verlängern um 1, sie werden genau $(|w| - 1)$ mal angewendet.
  - Produktionen der Form $A \longrightarrow a$ erzeugen genau ein Terminalsymbol von w, sie werden genau $|w|$ mal angewendet.
  - Das leere Wort wird durch $S \longrightarrow \varepsilon$ abgeleitet:
    Ableitung der Länge $2|w| + 1$.

Das Wortproblem für kontextfreie Grammatiken

- Es gibt nur endlich viele Ableitungen der Länge höchstens $2|w| + 1$.
- Um das Wortproblem zu entscheiden können wir
  - alle diese Ableitungen aufzählen und
  - prüfen, ob sie das Eingabewort w erzeugen.
- Aber: *exponentielle Laufzeit*, denn es kann exponentiell viele Ableitungen der Länge $2|w| + 1$ geben.
- Unser Ziel: polynomielle Laufzeit.

Der CYK-Algorithmus (Cocke, Younger, Kasami)

Definition

Sei $G = (N, \Sigma, P, S)$ kf. Grammatik in Ch. NF und $w = a_1 \cdots a_n \in \Sigma^*$. Sei

- $w_{ij} := a_i \cdots a_j \quad$ (für $i \leq j$)
- $N_{ij} := \{A \in N \mid A \vdash_G^* w_{ij}\}$.

1) $S \in N_{1n} \quad \Leftrightarrow w \in L(G)$

2) $A \in N_{ii} \quad \Leftrightarrow A \vdash_G^* a_i \quad \Leftrightarrow A \longrightarrow a_i \in P$

3) $A \in N_{ij}$ für $i < j \quad \Leftrightarrow A \vdash_G^* a_i \cdots a_j$

$\Leftrightarrow \exists A \longrightarrow BC \in P$ und ein k mit $i \leq k < j$ mit

$B \vdash_G^* a_i \cdots a_k$ und $C \vdash_G^* a_{k+1} \cdots a_j$

$\Leftrightarrow \exists A \longrightarrow BC \in P$ und k mit $i \leq k < j$ mit

$B \in N_{ik}$ und $C \in N_{(k+1)j}$

Beispiel

- Grammatik $G = (N, \Sigma, P, S)$ mit

$$
\begin{aligned}
P = \quad & \{S \longrightarrow SA \mid a, \\
& A \longrightarrow BS, \\
& B \longrightarrow BB \mid BS \mid b \mid c\}
\end{aligned}
$$

- $w = abacba$.

	$N_{1,6} = \{S\}$					
	$N_{1,5} = \emptyset$	$N_{2,6}=\{A,B\}$				
	$N_{1,4} = \emptyset$	$N_{2,5} = \{B\}$	$N_{3,6} = \{S\}$			
	$N_{1,3} = \{S\}$	$N_{2,4} = \{B\}$	$N_{3,5} = \emptyset$	$N_{4,6}=\{A,B\}$		
	$N_{1,2} = \emptyset$	$N_{2,3}=\{A,B\}$	$N_{3,4} = \emptyset$	$N_{4,5} = \{B\}$	$N_{2,3}=\{A,B\}$	
	$N_{1,1} = \{S\}$	$N_{2,2} = \{B\}$	$N_{3,3} = \{S\}$	$N_{4,4} = \{B\}$	$N_{5,5} = \{B\}$	$N_{6,6} = \{S\}$
$w =$	a	b	a	c	b	a

Der CYK-Algorithmus (Cocke, Younger, Kasami)

```
for i := 1 to n do
    N_ii := {A | A ⟶ a_i ∈ P}

for ℓ := 1 to n − 1 do   // wachsende Teilwortlänge ℓ = j − i
    for i := 1 to n − ℓ do   // Startposition i von Teilwort
        j := i + ℓ   // Endposition j von Teilwort
        N_ij := ∅
        for k := i to j − 1 do   // Mögliche Trennpositionen k
            N_ij := N_ij ∪ {A | ∃A ⟶ BC ∈ P mit B ∈ N_ik und C ∈ N_(k+1)j}

if S ∈ N_1n then output w ∈ L(G)
```

Der CYK-Algorithmus (Cocke, Younger, Kasami)

Satz

Für jede Grammatik G in Chomsky-Normalform entscheidet der CYK-Algorithmus das Wortproblem für $w \in \Sigma^*$ in Zeit $\mathcal{O}(|w|^3)$.

Beweis.

- Erste for-Schleife: $|w|$ Schritte
- Dann drei geschachtelte Schleifen, jeweils $\leq |w|$ Schritte,
  - also $\leq |w|^3$ Schritte in der innersten Schleife.
  - Suche nach den Produktionen $A \longrightarrow a_i$ und $A \longrightarrow BC$ nur konstante Zeit, denn G wird als konstant angenommen.

Das Äquivalenzproblem

Definition Äquivalenzproblem

- Gegeben: kontextfreie Grammatiken G_1, G_2.
- Frage: Gilt $L(G_1) = L(G_2)$?

- Wir werden in *Theoretische Informatik 2* beweisen:

Satz

Das Äquivalenzproblem für kontextfreie Sprachen ist *unentscheidbar*.

Zusammenfassung Entscheidungsprobleme

	Wortproblem	**Leerheitsprob.**	**Äquivalenzprob.**
Typ 0			Unentscheidbar
Typ 1			Unentscheidbar
Typ 2	Polyzeit	Polyzeit	Unentscheidbar
Typ 3 NEA/reg. Ausdr.	Linearzeit	Linearzeit	Exponentialzeit
Typ 3 DEA	Linearzeit	Linearzeit	Polynomialzeit