Reguläre Ausdrücke

Definition (Reguläre Ausdrücke)

Sei Σ ein Alphabet, dann gilt:

(i) $\emptyset$ ist ein regulärer Ausdruck über Σ.

(ii) ε ist ein regulärer Ausdruck über Σ.

(iii) Für jedes $a \in \Sigma$ ist a ein regulärer Ausdruck über Σ.

(iv) Wenn α und β reguläre Ausdrücke über Σ sind, so sind auch $\alpha\beta$, $(\alpha \mid \beta)$ und $(\alpha)^*$ reguläre Ausdrücke über Σ.

Sprache eines regulären Ausdrucks

Definition (Sprache eines regulären Ausdrucks)
Sei Σ ein Alphabet und γ ein regulärer Ausdruck über Σ, dann wird die von γ beschriebene Sprache $L(\gamma) \subseteq \Sigma^*$ wie folgt definiert.

(i) Für $\gamma = \emptyset$ gilt $L(\gamma) = \emptyset$.

(ii) Für $\gamma = \varepsilon$ gilt $L(\gamma) = \{\varepsilon\}$.

(iii) Für $\gamma = a$ mit $a \in \Sigma$ gilt $L(\gamma) = \{a\}$.

(iv) Für $\gamma = \alpha\beta$ gilt $L(\gamma) = L(\alpha) \cdot L(\beta)$.

(v) Für $\gamma = (\alpha \mid \beta)$ gilt $L(\gamma) = L(\alpha) \cup L(\beta)$.

(vi) Für $\gamma = (\alpha)^*$ gilt $L(\gamma) = (L(\alpha))^*$.

Beispiel eines regulären Ausdrucks

$$
\begin{aligned}
L((0 \mid (0 \mid 1)^*00)) &= L(0) \cup L((0 \mid 1)^*00)) \\
&= L(0) \cup (L((0 \mid 1)^*0)) \cdot L(0)) \\
&= L(0) \cup ((L((0 \mid 1)^*)) \cdot L(0)) \cdot L(0)) \\
&= L(0) \cup (((L((0 \mid 1)))^* \cdot L(0)) \cdot L(0)) \\
&= L(0) \cup (((L(0) \cup L(1))^* \cdot L(0)) \cdot L(0)) \\
&= \{0\} \cup (((\{0\} \cup \{1\})^* \cdot \{0\}) \cdot \{0\}) \\
&= \{0\} \cup ((\{0,1\}^* \cdot \{0\}) \cdot \{0\}) \\
&= \{0\} \cup (\{0,1\}^* \cdot \{00\}),
\end{aligned}
$$

Das heißt, die vom Ausdruck $(0 \mid (0 \mid 1)^*00)$ beschriebene Sprache ist die Menge aller Wörter über dem Alphabet $\{0,1\}$, die gleich 0 sind oder auf 00 enden.

Äquivalenz regulärer Ausdrücke

Definition

Zwei reguläre Ausdrücke β und γ heißen äquivalent, in Zeichen $\beta \equiv \gamma$, wenn $L(\beta) = L(\gamma)$ gilt.

Beispiel:

$$((a \mid b))^* \equiv ((a \mid b)(a \mid b))^*((a \mid b) \mid \varepsilon)$$

Rechenregeln für reguläre Ausdrücke

$$(A \mid (B \mid C)) \equiv ((A \mid B) \mid C)$$

$$A(B \mid C) \equiv (AB \mid AC)$$

$$(B \mid C)A \equiv (BA \mid CA)$$

$$(A \mid B) \equiv (B \mid A)$$

$$(A \mid A) \equiv A$$

$$(A \mid \emptyset) \equiv (\emptyset \mid A) \equiv A$$

$$A\emptyset \equiv \emptyset A \equiv \emptyset$$

$$A\varepsilon \equiv \varepsilon A \equiv A$$

$$((A)^*)^* \equiv (A)^*$$

$$(\emptyset)^* \equiv \varepsilon$$

$$((A \mid \varepsilon))^* \equiv (A)^*$$

$$(A)^* \equiv (\varepsilon \mid A)A^* \equiv ((\varepsilon \mid A))^*A$$

Bemerkung zu regulären Ausdrücken

1. Wir vereinbaren, dass wir Klammern, die nicht notwendigerweise gebraucht werden, weglassen können. Zum Beispiel können wir statt $(\alpha \mid (\beta \mid \gamma))$ auch $(\alpha \mid \beta \mid \gamma)$ schreiben. Wir schreiben auch $L(\alpha \mid \beta)$ statt $L((\alpha \mid \beta))$ sowie a^* statt $(a)^*$ für $a \in \Sigma$.

2. Wir benutzen die abkürzende Schreibweise α^n für $\underbrace{\alpha\alpha \ldots \alpha}_{n\text{-mal}}$.

3. Wir benutzen die abkürzende Schreibweise α^+ für $\alpha^*\alpha$.

4. In der Literatur findet man oft auch abweichende Schreibweisen der regulären Ausdrücke. Zum Beispiel findet man für $(\alpha \mid \beta)$ auch $(\alpha + \beta)$ oder auch $(\alpha \cup \beta)$. Auch wird natürlich oft $\alpha \cdot \beta$ für $\alpha\beta$ zugelassen.

5. Oft wird in der Literatur zwischen regulärem Ausdruck und beschriebener Sprache nicht unterschieden, das heißt, man identifiziert einen regulären Ausdruck mit der beschriebenen Sprache.

Weitere Beispiele regulärer Ausdrücke

$(a \mid b)^*$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$.

$(a \mid b)^+$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$, die nicht dem leeren Wort entsprechen.

$(a \mid b)^* aba (a \mid b)^*$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$, die das Teilwort aba haben.

$(a \mid b)^* a (a \mid b)^2$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$, deren drittletztes Symbol ein a ist.

$((a \mid b)(a \mid b))^*$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$, deren Länge gerade ist.

$(b \mid \varepsilon)(ab)^*(a \mid \varepsilon)$ beschreibt die Menge aller Wörter über dem Alphabet $\{a, b\}$, die nicht das Teilwort aa und nicht das Teilwort bb enthalten.

Satz von Kleene

Satz (Kleene)
Die Menge der durch reguläre Ausdrücke beschreibbaren Sprachen ist genau die Menge der regulären Sprachen.

Beweis (Konstruktionen)

- regulärer Ausdruck $\longrightarrow$ NEA: Induktion über Aufbau der Ausdrücke
- NEA $\longrightarrow$ regulärer Ausdruck: Lösung von Gleichungssystemen

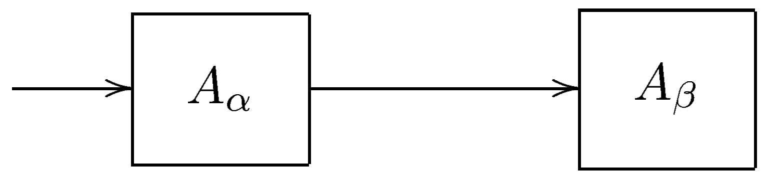
Regulärer Ausdruck $\rightarrow$ NEA

1. Automaten für die Ausdrücke $\emptyset$, ε und a:

2. Seien A_α bzw. A_β NEA mit $T(A_\alpha) = L(\alpha)$ bzw. $T(A_\beta) = L(\beta)$
Konstruiere NEA für $\alpha\beta$, $(\alpha \mid \beta)$ und α^*
(formale Konstruktion: siehe Skript)

Regulärer Ausdruck → NEA – Fortsetzung

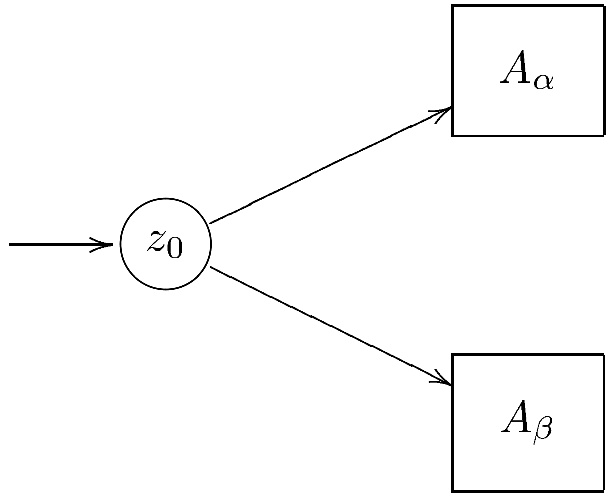
(a) Automat für $\alpha\beta$:



- Verschmelze Endzustände von A_α mit Startzustand von A_β. ("Reihenschaltung")
- neuer Startzustand: Startzustand von A_α
- neue Endzustände: Endzustände von A_β

Regulärer Ausdruck $\to$ NEA – Fortsetzung

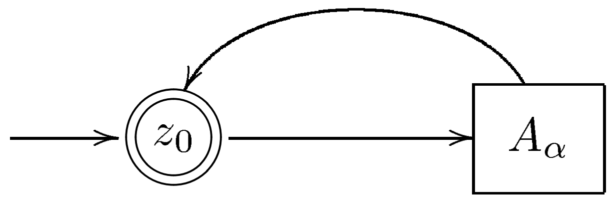
(b) Automat für $(\alpha \mid \beta)$:



- Neuer Startzustand mit Kanten zu den Nachfolgern der Startzustände von A_α und A_β ("Parallelschaltung").
- neue Endzustände: Endzustände von A_α und A_β

Regulärer Ausdruck $\rightarrow$ NEA – Fortsetzung

(c) Automat für $(\alpha)^*$:



- Füge für jede Kante zu einem Endzustand eine gleiche Kante zum Startzustand ein.
- neuer Endzustand: Startzustand von A_α

NEA $\rightarrow$ Regulärer Ausdruck

- Aufstellen eines “linearen Gleichungssystems” für einen NEA
- Auflösen des linearen Gleichungssystems $\rightarrow$ regulärer Ausdruck
- (Wegen der besseren Lesbarkeit geben wir nicht die Ausdrücke, sondern ihre Sprachen an.)

Gleichungssystem zu einem NEA

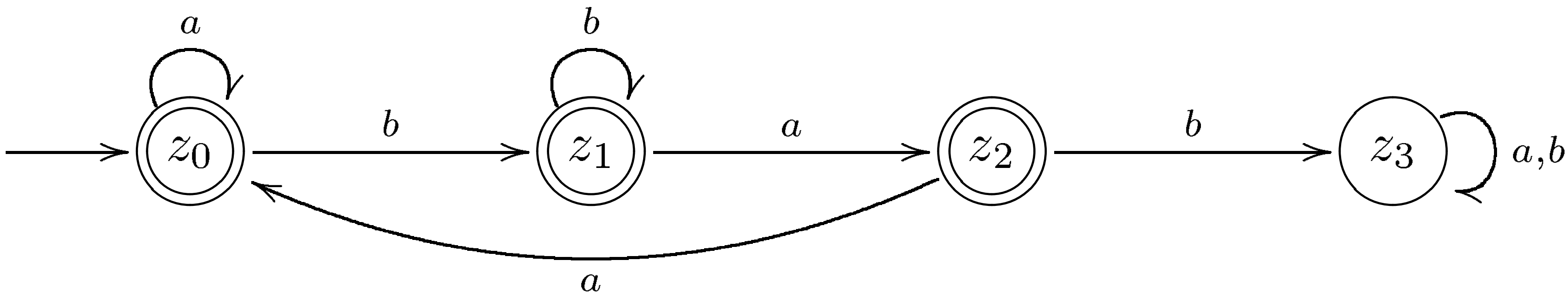
NEA $A = (Z, \Sigma, \delta, z_0, E)$

Für $z \in Z$ sei $A_z = (Z, \Sigma, \delta, z, E)$ und $L_z = T(A_z)$.
Für alle $z \in Z$ gilt:

$$L_z = \bigcup_{z' \in Z} \; \bigcup_{a : z' \in \delta(z,a)} a \cdot L_{z'} \cup E_z \qquad \text{mit } E_z = \begin{cases} \emptyset, & \text{falls } z \notin E \\ \{\varepsilon\}, & \text{sonst} \end{cases}$$

Lineares Gleichungssystem mit den Variablen $L_z, z \in Z$.
Falls eine eindeutige Lösung existiert, so ist durch L_{z_0} der gesuchte reguläre Ausdruck bekannt.

Gleichungssystem zu einem NEA-Beispiel



$$
\begin{aligned}
L_{z_0} &= \{a\} \cdot L_{z_0} \cup \{b\} \cdot L_{z_1} \cup \{\varepsilon\} \\
L_{z_1} &= \{b\} \cdot L_{z_1} \cup \{a\} \cdot L_{z_2} \cup \{\varepsilon\} \\
L_{z_2} &= \{a\} \cdot L_{z_0} \cup \{b\} \cdot L_{z_3} \cup \{\varepsilon\} \\
L_{z_3} &= \{a, b\} \cdot L_{z_3}
\end{aligned}
$$

Auflösen linearer Gleichungssysteme

Lemma. Für $B, C \subseteq \Sigma^*$ mit $\varepsilon \notin B$ gilt:
Die Gleichung $L = B \cdot L \cup C$ besitzt die einzige Lösung $L = B^* \cdot C$.

Auflösung von Gleichungssytemen durch sukzessive Eliminierung der Variablen L_z unter Nutzung des Lemmas

Auflösen linearer Gleichungssysteme - Beispiel

Für das Gleichungssystem auf Folie 169 ergibt sich

$$L_{z_3} = \{a, b\} \cdot L_{z_3} \rightarrow L_{z_3} = \{a, b\}^* \cdot \emptyset = \emptyset$$

$$L_{z_2} = \{a\} \cdot L_{z_0} \cup \{b\} \cdot L_{z_3} \cup \{\varepsilon\} \rightarrow L_{z_2} = \{a\} \cdot L_{z_0} \cup \{\varepsilon\}$$

$$L_{z_1} = \{b\} \cdot L_{z_1} \cup \{a\} \cdot L_{z_2} \cup \{\varepsilon\} = \{b\} \cdot L_{z_1} \cup \{aa\} \cdot L_{z_0} \cup \{a, \varepsilon\}$$
$$\rightarrow L_{z_1} = \{b\}^*\{aa\} \cdot L_{z_0} \cup \{b\}^*\{a, \varepsilon\}$$

$$L_{z_0} = \{a\} \cdot L_{z_0} \cup \{b\} \cdot L_{z_1} \cup \{\varepsilon\} = \{a\} \cdot L_{z_0} \cup \{b\}^+\{aa\} \cdot L_{z_0} \cup \{b\}^+\{a, \varepsilon\} \cup \{\varepsilon\}$$
$$\rightarrow L_{z_0} = (\{a\} \cup \{b\}^+\{aa\})^* \cdot (\{b\}^+\{a, \varepsilon\} \cup \{\varepsilon\})$$

Regulärer Ausdruck: $(a \mid b^+aa)^*(b^+(a \mid \varepsilon) \mid \varepsilon)$

Das Pumping Lemma

Satz (Pumping Lemma)

Sei L eine reguläre Sprache.
Dann gibt es eine Konstante $k \in \mathbb{N}$, so dass
für alle Wörter $z \in L$ mit $|z| \geq k$
eine Zerlegung $z = uvw$ existiert, so dass gilt:

(i) $|uv| \leq k$ und $|v| \geq 1$,

(ii) für alle $i \in \mathbb{N}$ gilt $uv^i w \in L$.

Bedeutung:
Wichtiges Hilfsmittel um zu beweisen, dass eine Sprache nicht regulär ist

Pumping Lemma – Beweis

- Reguläre Sprache L werde durch DEA $A = (Z, \Sigma, \delta, z_0, E)$ mit k Zuständen akzeptiert.
- Betrachte Wort $x \in L$ mit $|x| \geq k$. Es gilt $\hat{\delta}(z_0, x) = q$ für ein $q \in E$.
- Nach Einlesen der ersten k Buchstaben von x wurden $k+1$ Zustände angenommen.
  Nach dem Schubfachprinzip müssen 2 dieser Zustände gleich sein, d.h. es gibt eine Zerlegung $x = uvw$ mit $|uv| \leq k$, $|v| > 0$, $\hat{\delta}(z_0, u) = \hat{\delta}(z_0, uv) = z_u$, $\hat{\delta}(z_0, uvw) = q$.
- Nach Definition von $\hat{\delta}$ gilt:
  $\hat{\delta}(z_u, v) = z_u$, d.h. $\hat{\delta}(z_u, v^i) = z_u$ für alle $i \geq 0$
- Es folgt $\hat{\delta}(z_0, uv^iw) = \hat{\delta}(\hat{\delta}(z_0, uv^i), w) = \hat{\delta}(z_u, w) = q$, d.h. $uv^iw \in L$.

Anwendung des Pumping Lemmas

Erfüllt eine Sprache nicht die Folgerung des Pumping-Lemmas, so kann sie auch nicht regulär sein.

Um zu beweisen, dass eine Sprache L nicht die Folgerung des Pumping-Lemmas erfüllt, muss man Folgendes zeigen:

Für jede natürliche Zahl k
existiert ein Wort z mit $|z| \geq k$,
so dass für jede Zerlegung $z = uvw$ mit $|uv| \leq k$ und $|v| \geq 1$
ein $i \in \mathbb{N}$ existiert, so dass $uv^iw \notin L$ gilt.

Anwendungen des Pumping Lemmas - Beispiel

Satz: Die Menge PAL der Palindrome über $\{a, b\}$ ist nicht regulär.

Beweis

Wir zeigen, dass PAL nicht die Behauptung des Pumping-Lemmas erfüllt.

- Sei k eine beliebige natürliche Zahl.
- Wähle $z = a^k b a^k$. (Offensichtlich gilt $|z| \geq k$ und $z \in PAL$).
- Für jede Zerlegung $z = uvw$ mit $|uv| \leq k$ und $|v| \geq 1$ gilt:
  $u = a^r$, $v = a^s$, $w = a^{k-r-s} b a^k$ mit $r + s \leq k$ und $s \geq 1$.
- Wähle $i = 2$. Es gilt $uv^2w = a^r a^{2s} a^{k-r-s} b a^k = a^{k+s} b a^k \notin PAL$.

Anwendungen des Pumping Lemmas

Mit Hilfe des Pumping Lemmas kann man auch die Nichtregularität der Sprachen

$$L = \{a^{n^2} \mid n \geq 1\},$$
$$L = \{a^{2^n} \mid n \geq 1\},$$
$$L = \{a^p \mid p \text{ Primzahl}\},$$
$$L = \{a^n b^n \mid n \geq 1\},$$
$$L = \{ww \mid w \in \{a,b\}^*\},$$
$$L = \text{Menge der regulären Ausdrücke über } \Sigma$$

und vieler anderer zeigen.

Abschlusseigenschaften regulärer Sprachen

Satz Sind A und B reguläre Sprachen über Σ, dann sind auch

(i) $A \cup B$,

(ii) $A \cap B$,

(iii) $A \cdot B$,

(iv) $\Sigma^* \setminus A$ und

(v) A^*

reguläre Sprachen.

Wortproblem und andere Entscheidungsprobleme

Satz Das Wortproblem für ein Wort der Länge n und einen DEA ist mit einem Zeitaufwand von $O(n)$ entscheidbar.

Satz Das Wortproblem für ein Wort der Länge n und einen NEA mit k Transitionen ist mit einem Zeitaufwand von $O(n \cdot k)$ entscheidbar.

Satz Das Leerheitsproblem, das Endlichkeitsproblem, das Schnittproblem für NEA sind in linearer Zeit (Leerheit, Endlichkeit) bzw. in quadratischer Zeit (Schnitt) entscheidbar.

Satz Das Äquivalenzproblem für DEA ist in quadratischer Zeit entscheidbar.

Satz Das Äquivalenzproblem für NEA ist in exponentieller Zeit entscheidbar (und $\mathbb{NP}$-hart, genauer: **PSPACE**-vollständig).