(702765)

Skript zur Vorlesung am 30.6.2000

Theoretische Informatik I

Aus der vorherigen Vorlesung:

Satz W:
Sei X ein Alphabet. Zu jeder regulären Sprache $R \subseteq X^*$ gibt es ein $n \in N$, so daß für alle Wörter $z \in R$ mit $|z| \geq n$ gilt:
- Es gibt eine Zerlegung $z = uvw$ mit $u,v,w \in X^*$, $v \neq \varepsilon$, $|uv| \leq n$, so daß für alle $i \geq 0$ gilt:
  - $uv^iw \in R$

Formalisiert ausgedrückt (in der Form „∀∃∀∃∀“):
$\forall R \in Reg\ \exists n \in N\ \forall z \in R, |z| \geq n\ \exists u,v,w \in X^*(z = uvw \wedge v \neq \varepsilon \wedge |uv| \leq n\ (\forall i \in N_0(uv^iw \in R)))$

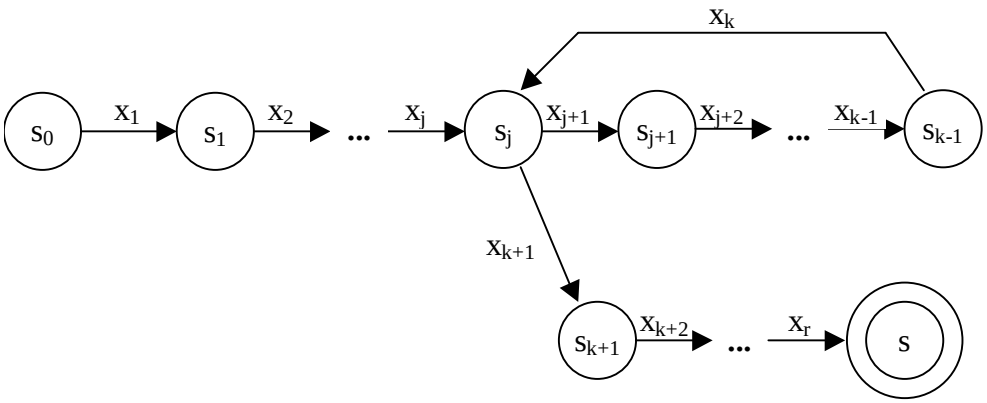
Im Klartext heißt das: In jeder regulären Sprache kann jedes Wort, welches eine gewisse Wortlänge n überschreitet, so in drei Teilwörter zerlegt werden, daß das mittlere Teilwort (welches nicht das leere Wort sein darf) beliebig vervielfacht oder auch weggelassen werden kann, ohne daß das gesamte Wort aus der Sprache herausfällt. Dieser Satz ist bekannt als das Pumping Lemma. Das Vervielfachen bezeichnet man auch als Pumpen.

Die Idee des Pumping Lemmas beruht darauf, daß
1. endliche Akzeptoren, um unbeschränkte Wörter erkennen zu können, in Zyklen laufen müssen, und daß
2. die einzige Möglichkeit, um unendliche Sprachen zu konstruieren, die Sternbildung ist, weshalb sehr lange Wörter viele Wiederholungen ein- und desselben Teilworts enthalten.

Vorlesung vom 30.06.2000:

Beweis des Pumping Lemmas (PL):
Sei R eine reguläre Sprache.
Zu R gibt es einen endlichen deterministischen Akzeptor $A = (X,S,\delta,s_0,F)$ mit $L(A) = R$.
Wir wählen nun $n := |S|$. Es sei $z \in R$ mit $|z| \geq n$.
Dann gibt es mindestens einen Zustand, der bei Eingabe von z zweimal durchlaufen wird.

$z = x_1...x_r$, $r \geq n$, $x_i \in X$ für $i = 1,...,r$



Für das **erste** Auftreten eines Zustands $s_j = s_k$ gilt:
$u := x_1...x_j$, $v := x_{j+1}...x_k \neq \varepsilon$
$|uv| = |x_1... x_k| \leq n$, weil es ja um das **erste** Auftreten geht
$w := x_{k+1}...x_r$

Wie man sieht, landet der Akzeptor im Zustand s_j, wenn er die Teilwörter u oder uv eingelesen hat (im Falle uv hat er den Zyklus bereits einmal durchlaufen). Er landet auch in diesem Zustand, wenn er die Teilwörter uv^i ($i \geq 0$) eingelesen hat, denn in diesem Fall durchläuft er den Zyklus i Mal.
Offenbar gilt also:
$\delta(s_0,u) = \delta(s_0,uv) = \delta(s_0,uv^i)$, $i \geq 0$

Das Teilwort w liest der Akzeptor also beginnend mit dem Zustand $s_j = s_k$ ein, egal ob v einmal, keinmal oder mehrmals im gesamten Wort enthalten ist. Damit gilt:
$\delta(s_0,uvw) = \delta(s_0,uv^iw) = s \in F$
$\Rightarrow uv^iw \in R$, $i \geq 0$

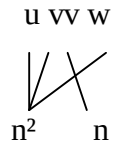
Zu beachten: Das PL ist keine Äquivalenz, sondern eine Implikation
L regulär $\Rightarrow$ L erfüllt das PL

Das Pumping Lemma wird oft benutzt, um zu beweisen, daß Sprachen **nicht** regulär sind, weil sie das PL **nicht** erfüllen. Das funktioniert aber nicht immer, denn es gibt auch nichtreguläre Sprachen, die das PL erfüllen.

Beispiel: $L = \{c^m a^n b^n | m,n \geq 0\} \cup \{a,b\}^*$
Es ist klar, daß Wörter aus $\{a,b\}^*$ das PL erfüllen. Deshalb betrachten wir dieWörter der anderen Menge. Dort existieren mehrere Pumpvarianten für die Wortzerlegung z = uvw:
$v = c^i \Rightarrow$ durch Pumpen von v landen wir in L *oder*
$v = a^i b^i \Rightarrow$ durch Pumpen von v landen wir in L

Beispiele für den Beweis der Nichtregularität durch das PL:
1. $L = \{a^n b^n | n \geq 0\}$ ist nicht regulär. Der Beweis erfolgte in der letzten Vorlesung.
2. $L = \{0^p | p$ ist Primzahl$\}$ist nicht regulär.
   Beweis durch Widerspruch: Wir nehmen an, L wäre regulär. Dann erfüllt L auch das PL. Wir wählen nun n wie im Satz W. Sei r eine Primzahl $r \geq 0$. Sei $z = 0^r \in L$, dann gibt es eine Zerlegung z = uvw mit $|uv| \leq n$, $u = 0^s$, $v = 0^t$, $t \neq 0$.
   Mit 0^r ist auch $0^{r+it} \in L$ für alle $i \geq 0$. Folglich sind alle Zahlen r+it Primzahlen. Nach spätestens t Zahlen kommt also stets eine Primzahl. Setzen wir i = r, dann ist r+rt eine Primzahl. Dann ist also auch r(1+t) eine Primzahl, andererseits sind aber r und t+1 Faktoren von r(1+t). Dies ist ein Widerspruch zur Primzahldefinition. L kann also nicht regulär sein, weil das Pumping Lemma nicht erfüllt wird.
3. $L = \{0^m | m$ ist Quadratzahl$\}$ist nicht regulär.
   Beweis durch Widerspruch: Wir nehmen an, L wäre regulär. Dann gibt es ein n el N, so daß sich jedes Wort z der Form 0^m,$m \geq n$, m Quadratzahl, zerlegen läßt in z = uvw mit den Eigenschaften des PL: $v \neq \varepsilon$, $|uv| \leq n$, $uv^iw \in L$, $i \geq 0$.
   Wir wählen jetzt speziell $z = 0^{n^2}$ und betrachten die zugehörige Zerlegung z = uvw.
   Dann ist offenbar aufgrund der Bedingungen des Pumping Lemma $1 \leq |v| \leq |uv| \leq n$.
   Ferner gilt für i = 2: $uv^2w \in L$.
   Andererseits: $n^2 = |z| = |uvw| < |uv^2w| \leq n^2 + n < n^2 + 2n + 1 = (n+1)^2$

|uv²w| anschaulich:



Wichtig ist hierbei: $n^2 < |uv^2w| < (n+1)^2$

Das bedeutet, daß $|uv^2w|$ zwischen zwei Quadratzahlen liegt. Somit ist $uv^2w \notin L$. Daraus folgt, daß das Pumping Lemma von L nicht erfüllt wird. Daher ist L nicht regulär.

4.7 Reguläre Grammatiken

Die bisher in den Vorlesungen benannten Konzepte im Bereich der Sprachentheorie sind für verschiedene Zwecke zu verwenden.

- Prozeß des Erkennens von Sprachen → Automaten
- Charakterisierungen von Sprachen → PL, Abschlußeigenschaften, algebraische Beschreibungen
- Beschreibungen von Sprachen → reguläre Ausdrücke
- *neu:* Erzeugungsprozesse von Sprachen → Grammatiken

Die Erzeugung von Sätzen/Wörtern einer Sprache erfolgt durch das Ersetzen von Symbolen. Welche Symbole wie ersetzt werden dürfen, bestimmen die Ableitungs- oder auch Produktionsregeln:

<Satz> → <Subjekt> <Prädikat> <Objekt>

<Subjekt> → Hund	<Prädikat> → beißt	<Objekt> → Katze
<Subjekt> → Katze	<Prädikat> → jagt	<Objekt> → Maus

<...>	Nichtterminalsymbole
<Satz>	Startsymbol
Hund, Katze, jagt,...	Terminalsymbole

Ableitungsprozeß:
<Satz> → <Subjekt> <Prädikat> <Objekt> → Katze <Prädikat> <Objekt>
→ Katze beißt <Objekt> → Katze beißt Maus

Dieses Beispiel führt uns zu einer formalen Definition von Grammatiken.

Definition X:

1) Eine Grammatik G beschreibt man durch ein 4-Tupel $G = (N, T, P, \sigma)$, wobei gilt:
   - N nichtleere, endliche Menge von Nichtterminalsymbolen
   - T nichtleere, endliche Menge von Terminalsymbolen
   - $N \cap T = \emptyset$
   - $\sigma \in N$ Startsymbol
   - $P \subseteq \{(\alpha,\beta) | \alpha,\beta \in (N \cup T)^*\}$ endliche Menge von Produktionsregeln

   Bemerkung: Statt (α,β) schreiben wir auch $\alpha \xrightarrow{\sigma} \beta$ oder nur $\alpha \longrightarrow \beta$, wenn der Bezug klar ist.

2) Seien $v,w \in (N \cup T)^*$. v ist ableitbar aus w (in Zeichen: $w \xrightarrow[G]{}{}^* v$ beziehungsweise $w \to^* v$ oder $w \to v$), wenn es Wörter
$v_1,...,v_k, u_1,...,u_k, z_1,...,z_k, \alpha_1,...,\alpha_{k-1}, \beta_1,...,\beta_{k-1} \in (N \cup T)^*$ gibt, so daß gilt:
$v_1 = w$, $v_k = v$, $v_i = u_i\alpha_i z_i$, $v_{i+1} = u_i\beta_i z_i$ für i = 1,...,k-1 und $(\alpha_i, \beta_i) \in P$ für i = 1,...,k-1

Die Ableitungsfolge sieht dann etwa so aus:
$w = u_1\alpha_1 z_1 \to u_1\beta_1 z_1 = u_2\alpha_2 z_2 \to u_2\beta_2 z_2 \to ... \to v$

Prinzipiell heißt das also, daß ein Wort v aus einem Wort w ableitbar ist, wenn es unter Ausnutzung der Ableitungsregeln möglich ist, die Symbole in w so zu ersetzen, daß am Ende v herauskommt.

3) Die von G erzeugte Sprache ist definiert durch $L(G) = \left\{ w \in T^* \mid \sigma \xrightarrow[G]{}{}^* w \right\}$

Beispiele von allgemeinen Grammatiken ohne Einschränkungen:

1. $L_1 = \{a^n b^n \mid n \in N_0\}$
$G_1 = (N_1, T_1, P_1, \sigma_1)$ mit
$T_1 = \{a,b\}$ $N_1 = \{\sigma_1\}$ $P_1 = \{\sigma_1 \to a\sigma_1 b, \sigma_1 \to \varepsilon\}$

Ableitungsfolge:
$\sigma_1 \to a\sigma_1 b \to aa\sigma_1 bb \to ... \to a^n\sigma_1 b^n \to a^n b^n$

2. korrekte Klammerung von `begin end`
$G_2 = (N_2, T_2, P_2, \sigma_2)$ mit
$T_2 = \{$`begin, end`$\}$ $N_2 = \{\sigma_2\}$ $P_2 = \{\sigma_2 \to$ `begin` σ_2 `end` $\sigma_2, \sigma_2 \to \varepsilon\}$

Ableitungsfolge:
$\sigma_2 \to$ `begin` σ_2 `end` $\sigma_2 \to$ `begin` σ_2 `end` $\to$ `begin begin` σ_2 `end` σ_2 `end`
$\to$ `begin begin end` σ_2 `end` $\to$ `begin begin end begin` σ_2 `end` σ_2 `end`
$\to$ `begin begin end begin` σ_2 `end end` $\to$ `begin begin end begin end end`

```
begin
  begin
  end
  begin
  end
end
```

$L(G_2)$ ist übrigens nicht regulär.

Die Typisierung von Grammatiken und der Bezug zu speziellen Sprachklassen erfolgt über die Einschränkung der Form von Produktionen. Das führt uns zur

Definition Y:
Eine Grammatik G = (N, T, P, σ) heißt rechtslineare Grammatik, wenn gilt:
Für alle $(\alpha,\beta) \in P$ gilt: $\alpha \in N$ und $\beta \in T^*$ oder $\beta = \beta' B$ mit $\beta' \in T^*$, $B \in N$
(Es gilt also, daß während der Ableitung in den einzelnen Schritten nur Nichtterminalsymbole auftreten oder die Nichtterminalsymbole rechts von Terminalsymbolen stehen. Neue Terminalsymbole können also nur rechts an die anderen Terminalsymbole angehängt werden.)

Analog definiert man die Linkslinearität einer Grammatik $G = (N, T, P, \sigma)$:

Für alle $(\alpha,\beta) \in P$ gilt: $\alpha \in N$ und $\beta \in T^*$ oder $\beta = B\beta'$ mit $\beta' \in T^*$, $B \in N$

Beispiel einer rechtslinearen Grammatik: $L = \{a\}^* \bullet \{b\}^*$
Sei $G = (N, T, P, \sigma)$ mit $N = \{\sigma,\sigma'\}$, $T = \{a,b\}$, $P = \{\sigma \to a\sigma, \sigma \to \sigma', \sigma' \to b\sigma', \sigma' \to \varepsilon\}$, $L(G) = L$

Satz Z:
Sei X ein Alphabet. Zu jeder regulären Sprache $L \subseteq X^*$ gibt es eine rechtslineare Grammatik G mit $L = L(G)$ und umgekehrt.
Beweis:

„$\Rightarrow$“ Sei $L \subseteq X^*$ regulär. Dann gibt es einen endlichen deterministischen Akzeptor $A = (X,S,\delta,s_0,F)$ mit $L = L(A)$.

Idee: Konstruiere eine Grammatik aus der Zustandsübergangsfunktion δ

Sei $N := S$, $T := X$, $\sigma := s_0$

$P := \{s \to xs' | \forall s,s' \in S, x \in X: \delta(s,x) = s'\} \cup \{s \to \varepsilon | s \in F\}$

Zu zeigen ist jetzt: $L(G) = L$

Sei $w \in L(G)$, $w = x_1...x_k$, $x_i \in X$, $i = 1...k$

$\Rightarrow s_0 \xrightarrow[G]{}{}^* w \Rightarrow s_0 \to x_1s_1 \to x_1x_2s_2 \to ... \to x_1...x_ks_k \to x_1...x_k$

$\Rightarrow \forall i \in \{0,...,k-1\}: \delta(s_i,x_{i+1}) = s_{i+1}$ und $s_k \in F$

$\Rightarrow \delta^*(s_o,w) = s_k \in F$

$\Rightarrow w \in L(A) = L$

$\Rightarrow L(G) = L$

„$\Leftarrow$“ Diese Beweisrichtung wird in der nächsten Vorlesung verfolgt