Formale Methoden 1

Gerhard Jäger
Gerhard.Jaeger@uni-bielefeld.de

Uni Bielefeld, WS 2007/2008

28. November 2007

Theorie formaler Sprachen

Formale Sprache:

- Menge von Symbolketten
- Formale Sprachen modellieren zunächst nur den Form-Aspekt natürlicher Sprachen.
- Annahme, dass jede Zeichenkette eindeutig einer Sprache entweder zugehört oder nicht zugehört $\Rightarrow$ Idealisierung
- alle interessanten formalen Sprachen sind unendlich (also unendliche Mengen von endlichen Zeichenketten)
- formale Grammatik: endliche Beschreibung einer formalen Sprache
- (Sprach-)Automaten: abstrakte Maschinen (Computerprogramme), die entscheiden können, ob eine Kette zu einer bestimmten formalen Sprache gehört oder nicht

Grundlagen

- Gegeben sei eine **endliche** Menge A von Symbolen (das *Alphabet* oder *Vokabular*)
- (Zeichen-)*Kette* über A: endliche Sequenz von Elementen von A
- Beispiel:
  - $A = \{a, b, c\}$
  - Ketten über A:
    - abc
    - $acbbcab$
    - $bacbbca$
- *Länge* einer Kette: Anzahl der Symbol-Vorkommen in der Kette (wenn das selbe Symbol mehrfach vorkommt, wird es mehrfach gezählt)
  - $l(abc) = 3$
  - $l(acbbcab) = 7$
  - $l(bacbbca) = 7$

Grundlagen

- Eine Kette der Länge n über das Vokabular A kann mengentheoretisch modelliert werden zum Beispiel:
  - als Funktion von $\{m \in \mathbb{N} | 1 \leq m \leq n\}$ in A, oder
  - als n-Tupel, also Element von A^n.

  Die Art der Reduktion von Ketten auf Mengen spielt im Weiteren keine Rolle und kann implizit gelassen werden.
- Es ist zu unterscheiden zwischen einem Element $a \in A$ und der Kette a von Länge 1, die nur aus a besteht. Wenn Verwechslungsgefahr besteht, schreiben wir $\langle a \rangle$ für die Kette, die nur das Symbol a enthält.
- Es gibt genau eine Kette der Länge 0, die **leere Kette**. Sie wird notiert als ϵ (manchmal auch als e oder als $\langle\rangle$, da es sich gewissermaßen um ein 0-Tupel handelt).
- Die Menge aller endlichen Ketten über A (einschließlich der leeren Kette) wird mit A^* bezeichnet.

Verkettung

- wichtigste Operation über Ketten: *Verkettung* (auch *Konkatenation* genannt)
- Aneinanderreihung zweier Ketten:
  - $abc \frown abc = abcabc$
  - $daaac \frown \epsilon = daaac$
  - $\epsilon \frown cabbba = cabba$
- assoziativ: für beliebige Ketten $u, v, w \in A^*$:

$$(u \frown v) \frown w = u \frown (v \frown w)$$

- ϵ ist **neutrales Element** für Verkettung:

$$\epsilon \frown u = u = u \frown \epsilon$$

Umkehrung einer Kette

- Notation: Wenn u eine Kette ist, ist u^R die Umkehrung dieser Kette.
- z.B.: $(acbab)^R = babca$
- für leere Kette gilt: $\epsilon^R = \epsilon$
- rekursive Definition:

Definition

Sei A ein Alphabet.

1. Wenn v eine Kette von Länge 0 ist (also $v = \epsilon$), dann $v^R = v$.
2. Wenn v eine Kette von Länge $n + 1$ ist, dann hat sie die Form wa (mit $w \in A^*$ und $a \in A$). Es gilt: $(wa)^R = aw^R$.

Grundlagen

- Zusammenhang zwischen Verkettung und Umkehrung:

$$(u \frown v)^R = v^R \frown u^R$$

- **Teilkette**: v ist eine *Teilkette* von $u \in A^*$ gdw. es $z, w \in A^*$ gibt und $u = z \frown v \frown w$.
- Wenn v eine Teilkette von u ist und $l(v) < l(u)$, dann ist v eine **echte Teilkette** von u.
- **Präfix:** v ist ein *Präfix* von $u \in A^*$ gdw. es ein $w \in A^*$ gibt so dass $u = v \frown w$.
- **Suffix:** v ist ein *Suffix* von $u \in A^*$ gdw. es ein $w \in A^*$ gibt so dass $u = w \frown v$.

Sprachen

Formale Sprache

Eine (formale) **Sprache** über ein Alphabet A ist eine Teilmenge von A^*, also eine Menge von Ketten über A.

- Sprachen können endlich oder unendlich sein.
- Linguistisch interessant sind v.a. unendliche Sprachen.
- Nicht für alle Sprachen gibt es endliche Beschreibungen. Wissenschaftlich untersuchbar sind nur solche Sprachen, für die es eine endliche Beschreibung gibt.
- Humboldt: (Natürliche) Sprachen machen „von endlichen Mitteln einen unendlichen Gebrauch“ $\Rightarrow$ natürliche Sprachen sind unendlich, aber sie sind endlich beschreibbar.

Beispiele für formale Sprachen

- $L = \{x \in \{a, b\}^* | x$ enhält die selbe Anzahl von a und b (in beliebiger Reihenfolge)$\}$
- $L_1 = \{x \in \{a, b\}^* | x = a^n b^b (n \geq 0)$, d.h. eine Kette von n mal a, gefolgt von der gleichen Anzahl von $b)\}$
- $L_2 = \{x \in \{a, b\}^* | x$ enthält n-mal b und n^2-mal a, für $n \in \mathbb{N}\}$

Grammatiken

(Formale) Grammatiken sind präzise Beschreibungen von formalen Sprachen. Eine Grammatik besteht aus

- zwei Alphabeten, dem **Terminalalphabet** V_T und dem **Nicht-Terminalalphabet** V_N,
- einem **Startsymbol** S, sowie
- einer Menge von **(Ersetzungs-)Regeln**. Eine Ersetzungsregel besteht aus zwei Teilen, der **linken Seite** und der **rechten Seite**.

Eine **Ableitung** für eine Grammatik erhält man, indem man mit der Kette S beginnt und sukzessive Teilketten, die der linken Seite einer Regel entsprechen, durch die rechte Seite der selben Regel ersetzt.

Beispiel

$$
\begin{aligned}
V_T \text{ (Terminalalphabet)} &= \{a, b\} \\
V_N \text{ (Nicht-Terminalalphabet)} &= \{S, A, B\} \\
S \text{ (Startsymbol)} & \\
R \text{ (Regeln)} &= \left\{ \begin{array}{lcl} S & \rightarrow & ABS \\ S & \rightarrow & \epsilon \\ AB & \rightarrow & BA \\ BA & \rightarrow & AB \\ A & \rightarrow & a \\ B & \rightarrow & b \end{array} \right\}
\end{aligned}
$$

Grammatiken

- Konvention: Terminal-Symbole werden als Kleinbuchstaben geschrieben und Nicht-Terminalsymbole als Großbuchstaben.
- **Ableitung** für die o.g. Grammatik:

  $S \Rightarrow ABS \Rightarrow ABABS \Rightarrow ABAB \Rightarrow ABBA \Rightarrow ABbA \Rightarrow aBbA \Rightarrow abbA \Rightarrow abba$
- Auf $abba$ kann keine Ersetzungsregel mehr angewandt werden, weil sie ausschließlich aus Terminalsymbolen besteht. Eine solche Kette heißt **Terminalkette**.
- Die Sprache, die von einer Grammatik **generiert** wird, ist die Menge aller Terminalketten, die durch die Regeln der Grammatik aus dem Startsymbol abgeleitet werden können.

Grammatiken

Definition ((Formale) Grammatik)

Eine (formale) **Grammatik** ist ein 4-Tupel $\langle V_T, V_N, S, R \rangle$, wobei V_T und V_N endliche disjunkte Mengen sind (also $V_T \cap V_N = \emptyset$), $S \in V_N$, und $R \in (V_T \cup V_N)^* \times (V_T \cup V_N)^*$ ist. Dabei gilt, dass die linke Seite jeder Regel mindestens ein Element von V_N enthält.

Üblicherweise werden Regeln geschrieben als $L \longrightarrow R$ statt $\langle L, R \rangle$.

Definition (Ableitung)

Sei $G = \langle V_T, V_N, S, R \rangle$ eine Grammatik. Eine **Ableitung** für G ist eine Folge von Ketten $x_1, x_2, \ldots, x_n (n \geq 1)$, so dass $x_1 = S$, und für jedes x_i mit $2 \leq i \leq n$ gilt:

- $x_i = uvw$,
- es gibt eine Regel $v \rightarrow z \in R$, und
- $x_{i+1} = uzw$.

Grammatiken

Definition (Generierung)

Eine Grammatik G **generiert** eine Kette $x \in V_T^*$ genau dann wenn es eine Ableitung $x_1, \ldots, x_n$ für G gibt, so dass $x_n = x$.

Definition (generierte Sprache)

Die von einer Grammatik G **generierte Sprache** (geschrieben als $L(G)$) ist die Menge aller Ketten, die von G generiert werden.