Einführung in die Computerlinguistik
Kontextfreie Grammatiken

Hinrich Schütze

Center for Information and Language Processing

2019-01-14

Die Grundfassung dieses Foliensatzes wurde von Prof. Dr. Stefan Evert erstellt. Fehler und Mängel sind ausschließlich meine Verantwortung.

Outline

Outline

Kontextfreie Grammatiken / Sprachen

- Mächtiger als reguläre Sprachen / Automaten
- Rekursive Strukturen
- Konstituentengrammatik

Defintion kontextfreie Grammatik (CFG)

Eine kontextfreie Grammatik G über dem Alphabet Σ ist ein Quadrupel $G = (V, \Sigma, P, S)$. Die Elemente von V heißen Variablen oder Nichtterminalsymbole, entsprechend werden die Zeichen aus Σ auch als Terminalsymbole bezeichnet. Wir nehmen stets $V \cap \Sigma = \emptyset$ an. üblicherweise verwenden wir für Terminalsymbole Kleinbuchstaben $a, b, c, \ldots \in \Sigma$ und für Variablen Großbuchstaben $A, B, C, \ldots \in V$. Zur Unterscheidung von Wörtern $u, v, w, \ldots \in \Sigma^*$ bezeichnen wir Zeichenketten, die sowohl Variablen als auch Terminalsymbole enthalten können, als Terme und verwenden dafür griechische Kleinbuchstaben $\alpha, \beta, \gamma, \ldots \in (V \cup \Sigma)^*$. $S \in V$ ist eine spezielle Variable, die Startsymbol genannt wird. $P \subseteq V \times (V \cup \Sigma)^*$ schließlich ist die Menge der Produktionen: jede Produktion ist von der Form $A \to \alpha$, wobei A eine Variable und α ein beliebiger Term ist.

Produktionen

Eine Produktion $A \to \alpha \in P$ wird auch als Regel bezeichnet, A als linke Seite und α als rechte Seite der Regel.
Zur Vereinfachung der Notation dürfen Regeln $A \to \alpha_1$, $A \to \alpha_2$, $\ldots$, $A \to \alpha_n$ mit identischer linken Seite zusammengefasst werden: $A \to \alpha_1 | \alpha_2 | \ldots | \alpha_n$.

Ableitungsschritt

Ein Ableitungsschritt $\delta \Rightarrow_G \delta'$ überführt einen Term δ durch Ersetzung genau einer Variable in einen Term δ'. In der formalen Darstellung schreiben wir $\delta = \beta A \gamma$ und $\delta' = \beta \alpha \gamma$, wobei $A \in V$ die genannte Variable ist, die durch einen Term α ersetzt wird. Der Ableitungsschritt $\beta A \gamma \Rightarrow_G \beta \alpha \gamma$ ist zulässig, wenn es eine Produktion $A \to \alpha \in P$ gibt.

Ableitung

Eine Ableitung ist eine beliebige Folge von zulässigen Ableitungsschritten: $\alpha_1 \Rightarrow_G \alpha_2 \Rightarrow_G \ldots \Rightarrow_G \alpha_n$. Wir schreiben kurz $\alpha_1 \Rightarrow_G^* \alpha_n$ und sagen, dass α_n aus α_1 ableitbar ist. Der Index G kann dabei ausgelassen werden, sofern klar ist, bezüglich welcher Grammatik G die Ableitung durchgeführt wird.

Definition kontextfreie Sprache

Die von G beschriebene formale Sprache $\mathcal{L}\left[G\right]$ ist die Menge aller Wörter w, die aus dem Startsymbol ableitbar sind:

$$\mathcal{L}\left[G\right] := \{w \in \Sigma^* \mid S \Rightarrow^* w\}.$$

Eine Sprache $L \subseteq \Sigma^*$ heißt kontextfrei, wenn sie durch eine kontextfreie Grammatik G beschrieben werden kann, d.h. wenn $L = \mathcal{L}\left[G\right]$ gilt.

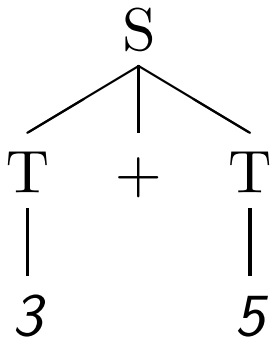
Beispiel

Als Beispiel betrachten wir eine kontextfreie Grammatik G_1 für einfache arithmetische Ausdrücke über dem Alphabet $\Sigma_1 = \{0, 1, \ldots, 9, +, *\}$. $G_1 = (V_1, \Sigma_1, P_1, S)$ ist folgendermaßen definiert:

$$
\begin{aligned}
V_1 &:= \{S, T\} \\
P_1 &:= \{S \to T, \\
&\qquad T \to T + T | T * T \\
&\qquad T \to 0|1|2|3|4|5|6|7|8|9\}.
\end{aligned}
$$

Die Variable T repräsentiert dabei jeweils einen arithmetischen Term. Das Wort $w = 3 + 5$ gehört zu $\mathcal{L}\left[G_1\right]$, da es aus S ableitbar ist: $S \Rightarrow T \Rightarrow T + T \Rightarrow 3 + T \Rightarrow 3 + 5$, also kurz $S \Rightarrow^* 3 + 5$.

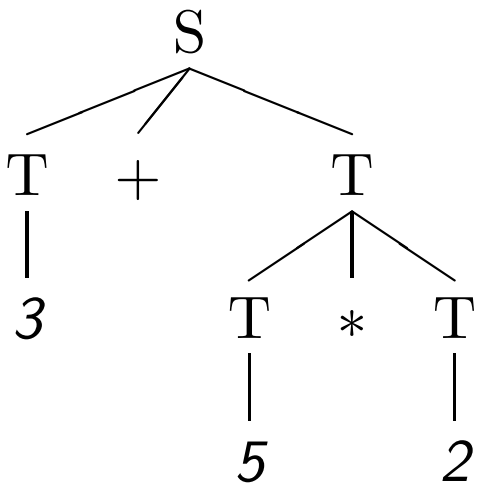
Ableitungsbaum für $S \Rightarrow T \Rightarrow T + T \Rightarrow 3 + T \Rightarrow 3 + 5$

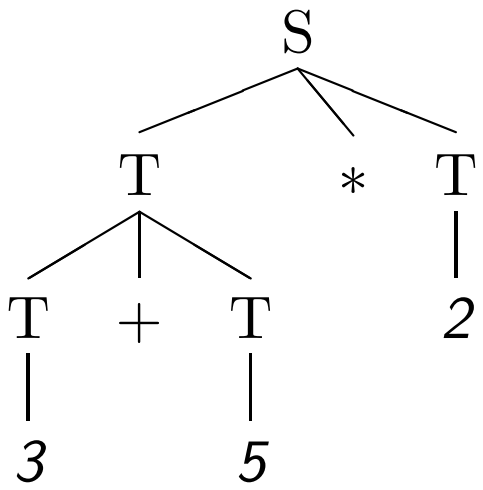


Die Ableitung von w bezüglich der Grammatik G_1 ist nicht eindeutig: eine andere mögliche Ableitung ist $S \Rightarrow T \Rightarrow T + T \Rightarrow T + 5 \Rightarrow 3 + 5$. Beide Varianten führen jedoch auf denselben Ableitungsbaum.

Ambiguität

Ein Wort w kann bezüglich einer CFG G auch mehrere verschiedene Ableitungsbäume besitzen. Z.B. hat $w = 3 + 5 * 2 \in \mathcal{L}\left[G_1\right]$ die folgenden beiden Ableitungsbäume:





Ableitung = Analyse

Die beiden Ableitungsbäume weisen w unterschiedliche Struktur zu, was für Anwendungen von großer Bedeutung ist. Man bezeichnet den Ableitungsbaum daher auch als Analyse von w durch die Grammatik. (Man denke z.B. an ein Taschenrechnerprogramm, das arithmetische Ausdrücke anhand ihres Ableitungsbaums auswertet. In diesem Fall wäre die linke Analyse die gewünschte, da $*$ stärker bindet als $+$). Eine Grammatik, in der es ein Wort w mit mehreren verschiedenen Ableitungsbäumen (bzw. Linksableitungen) gibt, heißt mehrdeutig / ambig.

Verbesserte Grammatik

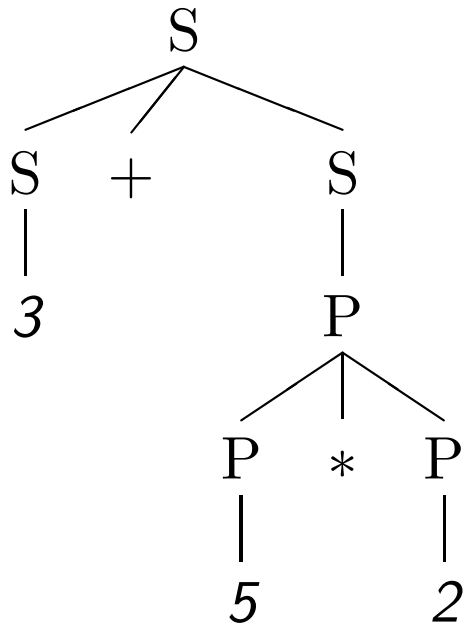
Im obigen Beispiel wäre es wünschenswert, G_1 so abzuändern, dass $w = 3 + 5 * 2$ nur noch eine Analyse besitzt (nämlich die durch den linken Baum dargestellte). Eine solche Grammatik ist $G_2 = (V_2, \Sigma_1, P_1, S)$ mit

$$
\begin{aligned}
V_2 &:= \{S, P\} \\
P_2 &:= \{S \to S + S | P, \\
&\qquad S \to 0|1|2|3|4|5|6|7|8|9, \\
&\qquad P \to P * P, \\
&\qquad P \to 0|1|2|3|4|5|6|7|8|9\}.
\end{aligned}
$$

Dabei stehen die Variablen S und P anschaulich für *Summe* und *Produkt*.

Verbesserte Grammatik: Beispiel $w = 3 + 5 * 2$

Bezüglich G_2 besitzt w nur noch die folgende eindeutige Analyse:



Outline

cfgtopdown.odp

Outline

cyk.odp

cyk,2009.odp

time,flies,cyk.pdf

Besonders klausurrelevant

- Formale Definition CFG:
  Terminale, Variablen, Startsymbol, Produktionen
- CYK
- Ableitungsbäume