Tabellenparser für übersetzte Grammatiken

Aus der Grammatik seien die Hilfsrelationen

(i) die *löschbaren Kategorien* $\{A \mid A \Rightarrow^+ \epsilon\} \subseteq Cat$,

(ii) die *Kettenrelation* $\{\,(A, B) \mid\ A \Rightarrow^+ B\,\} \subseteq Cat^2$,

(iii) die *Linke-Ecken-Relation*

$$<_{lc} \quad := \quad \{(B, A) \mid (A \to \alpha B \gamma) \in P,\ \alpha \Rightarrow^* \epsilon\}$$

und deren reflexive, transitive Hülle

$$<^*_{lc} \quad := \quad \{(B, A) \mid \exists \gamma \in Cat^*\ A \Rightarrow^* B\gamma\}$$

(iv) und folgende Teilmengen von $\dot{P}$, für $E, D \in Cat, c \in \Sigma$:

$$\begin{aligned}
&\mathit{predict}_{ep}(E, c) := \\
&\qquad \{(A \to \alpha \cdot C\beta) \in \dot{P} \mid A <^*_{lc} E,\ \alpha \Rightarrow^* \epsilon,\ c <^*_{lc} C\} \\
&\mathit{predict}_{lc}(E, D, c) := \\
&\qquad \{(A \to \alpha B \beta \cdot C\gamma) \in \dot{P} \mid \\
&\qquad\qquad A <^*_{lc} E,\ \alpha\beta \Rightarrow^* \epsilon,\ B \Rightarrow^* D,\ c <^*_{lc} C\}
\end{aligned}$$

vorberechnet. Wir benutzen sie nun, um

(i) die Suche nach passenden Grammatikregeln während der der Leftcorner- bzw. Earley-*Erkennung* einzuschränken,

(ii) dafür zu sorgen, daß nur kanonische Lösch- und Kettenableitungen gesucht werden.

Aus (ii) erhalten wir ein für jede Eingabe terminierendes *Analyse*verfahren, wobei in der Tabelle die Kanten $i \stackrel{A \to \alpha \cdot \beta}{\longrightarrow} j$ durch α-Bäume zu partiellen Analysen erweitert werden.

Verallgemeinerter Earley-Erkenner

Speichern von *Erwartungen* $j \cdot B$, „bei Position j beginne eine B-Phrase“, ersparen das Suchen nach Kanten $i \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} j$.

Vor.: Kontextfreie Grammatik $G = (N, \Sigma, S, P)$,
Hilfsrelationen $Cat \Rightarrow^* \epsilon$, $Cat \Rightarrow^* Cat$, $<^*_{lc}$, $predict_{ep}$

Eingabe: $w = a_1 \cdots a_n \in \Sigma^*$,

Methode: Konstruiere für die Positionen $0 \leq i \leq j \leq n$ alle Kanten $i \stackrel{A \to \alpha \cdot \beta}{\longrightarrow} j$, wo $(A \to \alpha\beta) \in P$, nach den Regeln:

$$\frac{start(S)}{0 \cdot S} \quad (Start) \qquad \frac{i \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} j}{j \cdot B} \quad (Expect)$$

$$\frac{j \cdot C, \quad (A \to \alpha \cdot B\beta) \in predict_{ep}(C, a_{j+1})}{j \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} j} \quad (Predict)$$

$$\frac{j = i + 1}{i \stackrel{a_j \to a_j \cdot}{\longrightarrow} j} \quad (Scan)$$

$$\frac{i \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} k, \quad B \Rightarrow^* D, \quad k \stackrel{D \to \delta \cdot}{\longrightarrow} j, \quad \beta \Rightarrow^* \epsilon}{i \stackrel{A \to \alpha B\beta \cdot}{\longrightarrow} j} \quad (Complete_1)$$

$$\frac{\begin{array}{c} i \stackrel{A \to \alpha \cdot B\beta C\gamma}{\longrightarrow} k, \quad B \Rightarrow^* D, \quad k \stackrel{D \to \delta \cdot}{\longrightarrow} j, \\ \beta \Rightarrow^* \epsilon, \quad j < n, \quad a_{j+1} <^*_{lc} C \end{array}}{i \stackrel{A \to \alpha B\beta \cdot C\gamma}{\longrightarrow} j} \quad (Complete_2)$$

Ausgabe: Eine Kante $0 \stackrel{S \to \alpha \cdot}{\longrightarrow} n$ wird erzeugt $\iff$ $w \in L(G)$.

Verallgemeinerter Leftcorner-Erkenner

Der Aufbau der Kanten zu $w = a_1 \cdots a_n$ erfolgt durch

(i) Festlegen von Erwartungskategorien für die Resteingabe:

$$\frac{start(S)}{0 \cdot S} \quad (Start) \qquad \frac{i \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} j}{j \cdot B} \quad (Expect)$$

(ii) Lesen der Eingabewörter:

$$\frac{j = i + 1}{i \stackrel{a_j \to a_j \cdot}{\longrightarrow} j} \quad (Scan)$$

(iii) Voraussagen brauchbarer Regeln:

$$\frac{i \cdot E, \qquad i \stackrel{D \to \delta \cdot}{\longrightarrow} j, \qquad (A \to \alpha B\beta \cdot C\gamma) \in predict_{lc}(E, D, a_{j+1})}{i \stackrel{A \to \alpha B\beta \cdot C\gamma}{\longrightarrow} j} \quad (Predict)$$

(iv) Abschluß unter Verkettung von Kanten:

$$\frac{i \stackrel{A \to \alpha \cdot B\beta}{\longrightarrow} k, \quad B \Rightarrow^* D, \quad k \stackrel{D \to \delta \cdot}{\longrightarrow} j, \quad \beta \Rightarrow^* \epsilon}{i \stackrel{A \to \alpha B\beta \cdot}{\longrightarrow} j} \quad (Complete_1)$$

$$\frac{i \stackrel{A \to \alpha \cdot B\beta C\gamma}{\longrightarrow} k, \quad B \Rightarrow^* D, \quad k \stackrel{D \to \delta \cdot}{\longrightarrow} j, \quad \beta \Rightarrow^* \epsilon, \quad j < n, \quad a_{j+1} <^*_{lc} C}{i \stackrel{A \to \alpha B\beta \cdot C\gamma}{\longrightarrow} j} \quad (Complete_2)$$

Neben den Erwartungskategorien $j \cdot B$ wird das nächste Eingabewort und $c <^*_{lc} C$ zum Filtern der Regeln benutzt wird.

Beobachtungen

(i) Die Menge P der Grammatikregeln wird zur Parsezeit nicht mehr durchsucht; stattdessen nur die Teilmengen $predict_{ep}$ bzw. $predict_{lc}$ und die Tabelle der Kanten.

(ii) Die verallgemeinerten Earley- und Leftcorner-Erkenner unterscheiden sich *nur* in (*Predict*).

Der Leftcorner-Erkenner erzeugt nur $i \stackrel{A \to \alpha \cdot \beta}{\longrightarrow} j$ mit $i < j$, d.h. wo α einen nicht-leeren Teil der Eingabe erkennt.

(iii) In (*Scan*) werden *ad-hoc*-Regeln $(a \to a)$ verwendet, die in $(D \to \delta)$ in (*Complete*) eingesetzt werden.

(iv) Zum *Erkennen* kann man Kanten $i \stackrel{A \to \alpha \cdot \beta}{\longrightarrow} j$ zu $i \stackrel{A \to \cdot \beta}{\longrightarrow} j$ vereinfachen, da der α-Teil nicht benutzt wird.

(v) Eine Kante $k \stackrel{D \to \delta \cdot}{\longrightarrow} j$ wird nur für $k < j$ erzeugt.

Erweiterung der verallgemeinerten Erkenner zu Earley- und Leftcorner-Tabellenparsern

(i) Die Kanten $i \xrightarrow{A \to \alpha \cdot \beta} j$ stehen für partielle Analysen von $a_{i+1} \cdots a_j$, d.h. sind Adressen für Baumfolgen mit Wurzeln α und Blättern $a_{i+1} \cdots a_j$.

(ii) In (*Complete*)-Schritten werden diese Baumfolgen um einen Baum zu $B \Rightarrow^* D \to \delta \Rightarrow^* a_{k+1} \cdots a_j$ mit $k < j$ erweitert.

(iii) In (*Complete*$_1$) wird für $i < k$ ein echt verzweigender Knoten A erzeugt, für $i = k$ muß man eine kanonische Ableitung $A \Rightarrow^* D$ aufbauen, also $B \Rightarrow^* D$ ignorieren.

(iv) (*Complete*$_2$) erzeugt nur echt verzweigende Knoten A, mit $k < j < l$ in $k \xrightarrow{B} j$ und $j \xrightarrow{C} l$ für ein l.

(v) Jede vollständige Analyse $A \Rightarrow^* a_i \ldots a_j \epsilon$ beginnt daher mit einer Kettenableitung $A \Rightarrow^* D$ zu einem Blatt D (und $i = j$) oder einem Verzweigungsknoten D.

Daraus folgt, daß nur kanonische Bäume aufgebaut werden, und daß der Parser terminiert.

Bem. Kanten $i \xrightarrow{D \to \delta \cdot} j$ kann man zu $i \xrightarrow{D} j$ vereinfachen, d.h. man braucht die *vollständigen* Analysen nur nach der Wurzel zu unterscheiden.