stochastische kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel,

$$(V, \Sigma, R, S)$$

mit Nicht-Terminalsymbolen V,
Terminalsymbolen Σ,
Regeln R der Form $X \rightarrow \beta$ (mit $X \in V, \beta \in (V \cup \Sigma)^*$)
und dem Startsymbol $S \in V$.

Warum Wahrscheinlichkeiten dazu?

- Wertung von Hypothesen
  (z.B. bei Spracherkennung),
- Auswahl des wahrscheinlichsten Ergebnisses (Parsing),
- frühzeitiges Ausschalten unwahrscheinlicher Analyseversuche (Parsing, Effizienzsteigerung)

stochastische kontextfreie Grammatik

Eine *stochastische* oder *probabilistische* kontextfreie Grammatik G ist ein 5-Tupel, (V, Σ, R, S, P),

mit V, Σ, R, S wie vorher und

P eine Funktion von R nach [0,1] mit

$$\forall X \in V : \sum_{\beta \in (V \cup \Sigma)^*} P(X \rightarrow \beta) = 1$$

Die Summe der Wahrscheinlichkeiten für die Regeln mit einem bestimmten Nichtterminal auf der linken Seite muss 1 sein.

Beispiel:

$$\begin{array}{lcll} S & \rightarrow & NP\,VP & (1.0) \\ VP & \rightarrow & V & (0.5) \\ VP & \rightarrow & V\;Adv & (0.5) \\ NP & \rightarrow & \textit{Hans} & (1.0) \\ V & \rightarrow & \textit{schläft} & (1.0) \\ Adv & \rightarrow & \textit{ständig} & (1.0) \end{array}$$

Wahrscheinlichkeit einer Zeichenfolge: Summe der W'keiten der Parsebäume dazu.

W'keit eines Parsebaums: W'keit der erzeugenden Linksableitung

Wahrscheinlichkeit einer Ableitung

Wahrscheinlichkeit einer Linksableitung:

Seien die Regeln durchnummeriert und die Ableitung repräsentiert durch die Folge der verwendeten Regeln. Sei X_m eine Zufallsvariable, die von der Regel bestimmt wird, die in Schritt m verwendet wurde.

Dann ist eine Linksableitung ein stochastischer Prozess $X_1, .., X_M$ mit der Regelmenge der Grammatik als Zustandsmenge.

Die W'keit einer Linksableitung ist:

$$\begin{aligned} & P(X_1 = r_{i_1}, .., X_M = r_{i_M}) \\ & \quad = P(X_M = r_{i_M} \mid X_1 = r_{i_1}, .., X_{M-1} = r_{i_{M-1}}) \\ & \qquad \cdot P(X_1 = r_{i_1}, .., X_{M-1} = r_{i_{M-1}}) \\ & \quad = \Pi_{m=1}^{M} P(X_m = r_{i_m} \mid X_1 = r_{i_1}, .., X_{m-1} = r_{i_{m-1}}) \end{aligned}$$

nötig: alle Faktoren, d.h. alle bedingten W'keiten für jede Regel.

Annahme, Regeln unabhängig von vorangegangenen Ableitungsschritten:

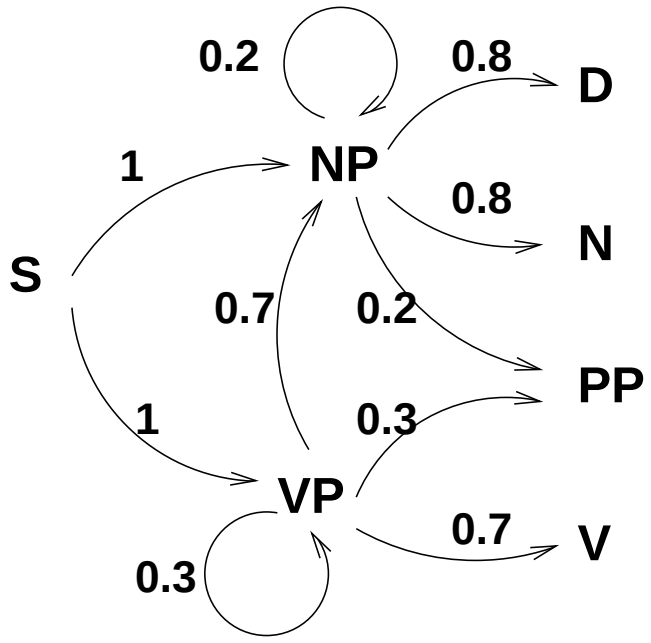
$$P(X_m = r_{i_m} \mid X_1 = r_{i_1}, .., X_{m-1} = r_{i_{m-1}}) = P(r_{i_m})$$

also $P(X_1 = r_{i_1}, .., X_M = r_{i_M}) = \prod_{m=1}^{M} P(r_{i_m})$

Parsingbeispiel

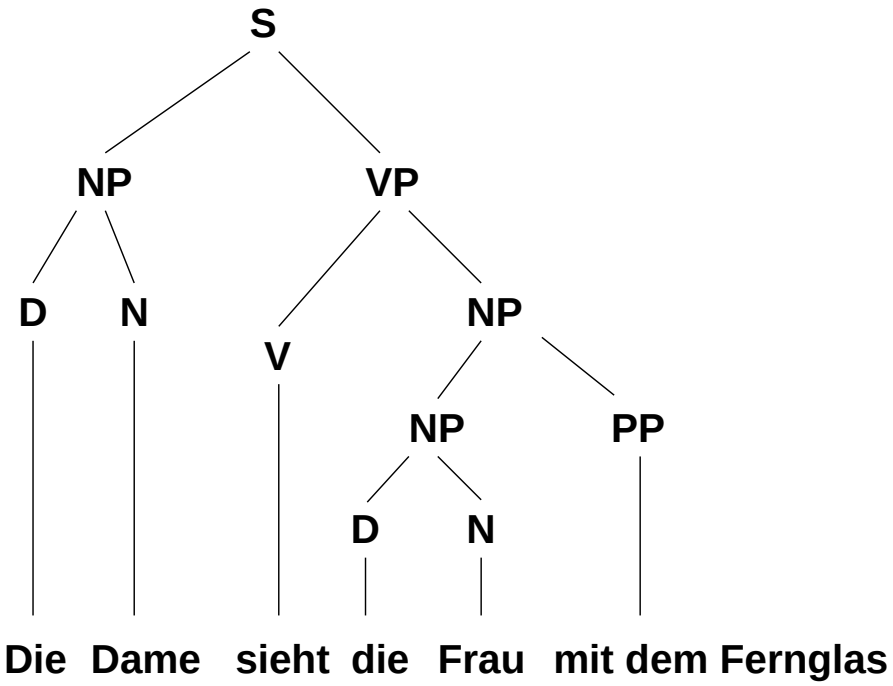
Grammatik:

1	S	$\rightarrow$	$NP\,VP$	(1.0)
2	VP	$\rightarrow$	$V\,NP$	(0.7)
3	VP	$\rightarrow$	$VP\,PP$	(0.3)
4	NP	$\rightarrow$	$D\,N$	(0.8)
5	NP	$\rightarrow$	$NP\,PP$	(0.2)
6	PP	$\rightarrow$	'mit dem Fernglas'	(1.0)
7	V	$\rightarrow$	'sieht'	(1.0)
8	N	$\rightarrow$	'Dame'	(0.4)
9	N	$\rightarrow$	'Frau'	(0.6)
10	D	$\rightarrow$	'die'	(1.0)

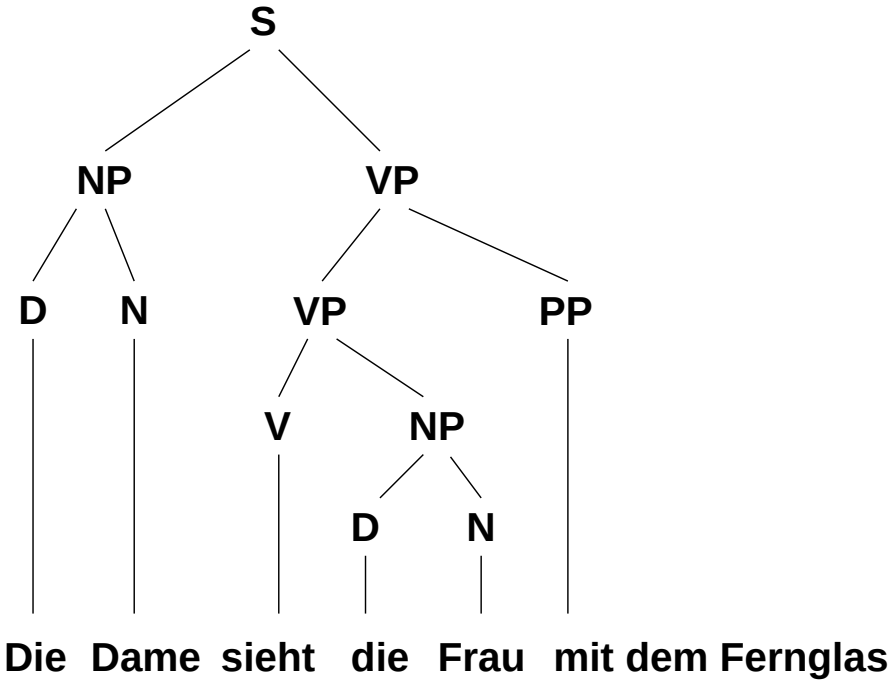


Lesarten von 'Die Dame sieht die Frau mit dem Fernglas'

I:



II:



Die Ableitungen

$S \Rightarrow^* w$ über (ohne lexikalische Ersetzungen):

I:

(S, NP VP, D N VP,

D N V NP,

D N V NP PP, D N V D N PP)

= (1, 4, 2, 5, 4)

II:

(S, NP VP, D N VP,

D N VP PP,

D N V NP PP, D N V D N PP)

= (1, 4, 3, 2, 4)

Welche davon ist wahrscheinlicher?

$$P(I) = (1 \cdot 0.8 \cdot 0.7 \cdot 0.2 \cdot 0.8) = 0.09$$

$$P(I) = (1 \cdot 0.8 \cdot 0.3 \cdot 0.7 \cdot 0.8) = 0.13$$

Notationen zu Parsebäumen

$\tau_1 \circ \tau_2$:
der am weitesten links stehende Nichtterminal-Knoten von τ_1 wird durch τ_2 ersetzt, linker Knoten und Wurzel müssen dabei das gleiche Nichtterminal als Label tragen.

$l(n)$: Label des Knotens n.

$R(\tau)$: Wurzel von τ.

$Y(\tau)$:
“yield“ des Baumes τ, Ergebnis der Regelanwendung, generierter String.

$L(\tau)$:
Label des am weitesten links stehenden Knoten in τ.

Korrespondenz zu partiellen Parsebäumen

Regel $X \rightarrow X_1, ..X_K$ entspricht Teilbaum mit Wurzel x und Knoten $X1, ..X_k$.

Sei τ_m der Teilbaum zu r_m, $\tau_1 \circ .. \circ \tau_M = \mathbf{t}_m$ die Folge der Teilbäume der Ableitung $r_{i_1}, .., r_{i_M}$. Dabei ist τ_M der vollständige Parsebaum.

Dann ist die Wahrscheinlichkeit eines Parsebaums:

$$P(\tau) = P(\mathbf{t}_M) = \prod_{m=1}^{M} P(\tau_m \mid \mathbf{t}_{m-1})$$

$$P(\tau_m \mid \tau_1 \circ .. \circ \tau_{m-1}) = P(r_m \mid r_1, .., r_{m-1})$$

Gesucht: Extraktor-Funktion g, die die relevanten Eigenschaften von $\mathbf{t}_m$ zur W'keitbestimmung extrahiert, mit:

$$P(\tau_{k+1} \mid \mathbf{t}_k) \approx P(\tau_{k+1} \mid g(\mathbf{t}_k))$$

und das ist der Label des am weitesten links stehenden Knotens im Ergebnis von τ_k, $g(\tau) = L(Y(\tau))$.

Ein Parser für SKFGs

Adaption des Viterbi-Algorithmus, Variante des CYK-Parsers, Zeitkomplexität $O(N^3T^3)$,
Speicherkomplexität $O(NT^2)$.

$V_N = X_1, ..X_N, S = X_1, \omega = w_1, ..w_T$, G in CNF (Analyse-W'keiten bleiben erhalten).

Akkumulator $\delta_n(X_i)$, Knoten n, Nichtterminale X_i.

Jeder Knoten n bestimmt einen Teilstring mit den Stringpositionen (s,t), sei $\mathbf{w}_{s,t} = w_{s+1}, .., w_t$ dieser Teilstring.

Die dazu inverse Funktion (nur partiell definiert!) bestimmt aus einem Teilstring denjenigen Parsebaumknoten, der am nächsten zur Wurzel liegt.

$$\delta_{s,t}(X_i) = \max_{\tau: Y(\tau) = \mathbf{w}_{s,t}} P(\tau \mid l(R(\tau)) = X_i),$$
$$1 \leq i \leq N, 0 \leq s < t \leq T$$

W'keit des wahrscheinlichsten Baums mit $\omega = \mathbf{w}_{0T}$:

$$\delta_{0,t}(s) = \max_{\tau: Y(\tau) = \omega} P(\tau \mid l(R(\tau)) = S)$$

Der Baum selbst ist:

$$\arg\max_{\tau: Y(\tau) = \omega} P(\tau \mid l(R(\tau)) = S)$$

weiter: Parser

Sei $p_{i \to jk} = P(X_i \to X_j X_k \mid X_i), p_{i \to w} = P(X_i \to w \mid X_i)$.

Konstruktion des Parsebaums, bottom-up:

1. Initialisierung

   $\forall i, t : 1 \leq i, \leq N, 1 \leq t \leq T$

   $$\delta_{t-1,t}(X_i) = p_{i \to w_t}$$

2. Rekursion

   $\forall i, r, t : 1 \leq i, \leq N, 1 \leq r + 1 < t \leq T$

   $$\delta_{r,t}(X_i) = \max_{\substack{1 \leq j \leq N \\ 1 \leq k \leq N \\ r < s < t}} p_{i \to jk} \delta_{r,s}(X_j) \delta_{s,t}(X_k)$$

   $$\begin{bmatrix} \iota_{r,t}(X_i) \\ \kappa_{r,t}(X_i) \\ \sigma_{r,t}(X_i) \end{bmatrix} = \arg\max_{\substack{1 \leq j \leq N \\ 1 \leq k \leq N \\ r < s < t}} p_{i \to jk} \delta_{r,s}(X_j) \delta_{s,t}(X_k)$$

3. Rekonstruktion, $n = (s, t)$

   $$\begin{aligned} \text{left}(n) &= \begin{cases} nil & \text{wenn } t - s \leq 2 \\ (s, \sigma_{s,t}(l(n))) & \text{sonst} \end{cases} \\ \text{right}(n) &= \begin{cases} nil & \text{wenn } t - s \leq 2 \\ (\sigma_{s,t}(l(n)), t) & \text{sonst} \end{cases} \\ l(\text{left}(n)) &= X_{\iota_{s,t}(l(n))} \\ l(\text{right}(n)) &= X_{\kappa_{s,t}(l(n))} \end{aligned}$$

weiter Parser

Begründungen

Initialisierung:

Die Nichtterminale, die die Terminale erzeugen, erhalten die W'keiten der lexikalischen Regeln.

Rekursion:

ι, κ sammeln die Indizes der rechten Seiten der verwendeten Regeln auf, σ speichert die String-Position zwischen X_ι und X_κ.

$\delta_{r,t}(X_i)$ wird aus den W'keiten für X_j und X_k berechnet:

$$\begin{aligned}
&\max_{\tau:Y(\tau)=\mathbf{w}_{r,t}} P(\tau \mid l(R(\tau)) = X_i) \\
&\quad = \max_{j,k,s}\big[P(X_i \rightarrow X_j\ X_k \mid X_i) \\
&\qquad \cdot \max_{\tau':Y(\tau')=\mathbf{w}_{r,s}} P(\tau' \mid l(R(\tau')) = X_j) \\
&\qquad \cdot \max_{\tau'':Y(\tau'')=\mathbf{w}_{s,t}} P(\tau'' \mid l(R(\tau')) = X_k)\big]
\end{aligned}$$

Parameterschätzungen für prob. Grammatiken

- wenn ein annotierter Korpus vorliegt:

  relative Häufigkeiten verwenden, Achtung bei spärlichen Daten

- wenn nicht:
  - Zuerst alle gültigen Parsebäume erzeugen, dann eine Gleichverteilung für alle Bäume pro Satz annehmen.

    Häufigkeitszählung der verwendeten Regeln (nach linkem Symbol), gewichtet mit der W'keit des Baumes: ergibt neue Verteilung für die Regeln, daraus neue Verteilung für die Bäume; das kann beliebig wiederholt werden.

    Problem: Komplexität, Anzahl der Parsebäume exponentiell zur Stringlänge.
  - *Inside-Outside*-Algorithmus

Parameterschätzungen für prob. Grammatiken

Inside-Outside-Algorithmus

Idee: Verwende die aktuellen W'keiten der Regeln, um davon abhängige andere Masse einzuschätzen.

Ziel: Finde die Menge von Regel-W'keiten, mit denen es am ehesten möglich ist, den Trainings-Korpus zu generieren.

Die Variablen:

Die *Inside*-W'keit $I_i^\omega(s,t)$ schätzt die W'keit

$$P(X_i \Rightarrow^* \mathbf{w}_{s,t} \mid X_i),$$

dass $\mathbf{w}_{s,t}$ abgeleitet wird, wenn X_i vorliegt.

Die *Outside*-W'keit $O_i^\omega(s,t)$ schätzt die W'keit

$$P(S \Rightarrow^* \mathbf{w}_{0,s} X_i \mathbf{w}_{t,T} \mid S),$$

dass die o.a. Zeichenkette von S abgeleitet wird, wenn S vorliegt.

- Inside-Variablen-Initialisierung:

  $\forall i, t : 1 \leq i \leq N, 1 \leq t \leq T$

  $$I_i^{\omega}(t-1, t) = p_{i \to w_t}$$

- Inside-Variablen-Rekursion

  $\forall i, r, t : 1 \leq i \leq N, 1 \leq r < t \leq T$

  $$I_i^{\omega}(r, t) = \sum_{j,k=1}^{N} \sum_{r<s<t} P_{i \to jk} \cdot I_j^{\omega}(r, s) \cdot I_k^{\omega}(s, t)$$

- Outside-Variablen-Rekursion

  $\forall i, t : 1 \leq i \leq N, 1 \leq t \leq T$

  O^{ω}

Herleitung

Inside:

$$P(X_i \Rightarrow^* \mathbf{w}_{r,t} \mid X_i) =$$
$$\Sigma_{j,k=1}^{N} P(X_i \Rightarrow X_j X_k \Rightarrow^* \mathbf{w}_{r,t} \mid X_i) =$$
$$\Sigma_{j,k=1}^{N} \Sigma_{r<s<t} P(X_i \Rightarrow X_j X_k \mid X_i) \cdot P(X_j \Rightarrow^* \mathbf{w}_{r,s} \mid X_j) \cdot P(X_k \Rightarrow^* \mathbf{w}_{s,t} \mid X_k)$$

Outside:

$$P(S \Rightarrow^* \mathbf{w}_{0,s} X_i \mathbf{w}_{t,T} \mid S) =$$
$$\Sigma_{j,k=1}^{N} \Sigma_{s=0}^{r-1} P(S \Rightarrow^* \mathbf{w}_{0,s} X_j \mathbf{w}_{t,T} \Rightarrow^* \mathbf{w}_{0,s} X_k X_i \mathbf{w}_{t,T} \Rightarrow^* \mathbf{w}_{0,s} \mathbf{w}_{s,r} X_i \mathbf{w}_{t,T}) +$$
$$\Sigma_{s=t+1}^{T} P(S \Rightarrow^* \mathbf{w}_{0,r} X_j \mathbf{w}_{s,T} \Rightarrow^* \mathbf{w}_{0,r} X_i X_k \mathbf{w}_{s,T} \Rightarrow^* \mathbf{w}_{0,r} X_i \mathbf{w}_{t,s} \mathbf{w}_{s,T}) =$$
$$\Sigma_{j,k=1}^{N} \Sigma_{s=0}^{r-1} P(S \Rightarrow^* \mathbf{w}_{0,s} X_j \mathbf{w}_{t,T}) P(X_j \Rightarrow^* X_k X_i \mid X_j) P(X_k \Rightarrow^* \mathbf{w}_{s,r} \mid X_k) +$$
$$\Sigma_{s=t+1}^{T} P(S \Rightarrow^* \mathbf{w}_{0,r} X_j \mathbf{w}_{s,T} P(X_j \Rightarrow^* X_i X_k \mid X_j) P(X_k \Rightarrow^* \mathbf{w}_{t,s} \mid X_k) =$$

Komplexität für beide Variablen $O(N^3 T^3)$.

Wiedereinschätzungsgleichung

Idee:

$$p_{i \to w} = P(X_i \to w \mid X_i) = \frac{P(X_i \to w)}{P(X_i)}$$

$$p_{i \to jk} = P(X_i \to X_j X_k \mid X_i) = \frac{P(X_i \to X_j X_k)}{P(X_i)}$$

mit:

$$P(X_i) = \frac{1}{\mid W \mid} \sum_{\omega \in W} \frac{P_i^\omega}{P^\omega}$$

$$P(X_i \to w) = \frac{1}{\mid W \mid} \sum_{\omega \in W} \frac{\sum_{1 \le t \le T, w_t = w} O_i^\omega(t-1, t) \cdot p_{i \to w}}{P^\omega}$$

$$P(X_i \to X_j X_k) = \frac{1}{\mid W \mid} \sum_{\omega \in W} \frac{\sum_{0 \le r < s < t \le T} O_i^\omega(r, t) \cdot p_{i \to jk} \cdot I_j^\omega(r, s) \cdot I_k^\omega(s, t)}{P^\omega}$$