第五章 上下文无关文法

2025年6月18日 21:03

上下文无关文法

定义：CFG(Context-Free Grammer)是一个四元组$G=(V,T,P,S)$，其中：
V为变元(Variable),T为终结符(Terminal),P为产生式(Production),
S为初始符号(Start Symbol),S∈V
每个产生式包含：一个变元(head),一个产生式符号→,一个$(V\cup T)^*$
中的符号串(body).
产生式$A\rightarrow\alpha_1,\cdots,A\rightarrow\alpha_n\Leftrightarrow A\rightarrow\alpha_1|\cdots|\alpha_n$

归约和派生：
归约：从字符串到文法变元的分析过程
派生：从文法变元到字符串的分析过程
最左/右派生：只替换符号串中最左/右边变元的派生过程，记为$\underset{lm}{\Rightarrow}$和$\underset{rm}{\Rightarrow}$
任何派生都有等价的最左派生和最右派生

文法的语言：
定义：$CFG\ G=(V,T,P,S)$的语言定义为$L(G)=\{w|w\in T^*,S\overset{*}{\underset{G}{\Rightarrow}}w\}$，且
w仅有终结符组成，初始符号S能派生出w

上下文无关语言：
定义：L=L(G),则称L为CFL(Context-Free-Language)
上下文无关是指在文法派生的每一步$\alpha A\beta\Rightarrow\alpha\gamma\beta$，符号串γ仅根据A产生式派生，无需依赖A的上下文α和β

文法的等价性：若L(G1)=L(G2)，则G1和G2是等价的

句型：初始符号S派生出的符号串称为G的句型(sentential form)
左/右句型：S $\overset{*}{\underset{lm}{\Rightarrow}}/\overset{*}{\underset{rm}{\Rightarrow}}\alpha$
只含有终结符的句型称为G的句子(sentence)
L(G)为文法G全部的句子

语法分析树：(parse tree)，将派生或归约的过程表示成树形结构

语法树和派生的等价性：
定理：A∈V，那么文法G中A⇒α当且仅当G中存在以A为根节点为α的语法树
以下命题等价：①串w在变元A的语言中 ②存在以A为根节点，产物为w的语法分析树 ③A $\overset{*}{\Rightarrow}$ w ④A $\overset{*}{\underset{lm}{\Rightarrow}}$ w ⑤A $\overset{*}{\underset{rm}{\Rightarrow}}$ w

文法和语言的歧义性

定义：若CFG G使某符号串有不同的语法分析树，则称文法G是歧义的(ambiguity)
固有歧义性：CFL L的所有文法都为歧义的，L是固有歧义的(Inherent Ambiguity)
判断任何给定CFG G是否歧义 或 CFL L是否固有歧义 均为不可判定的

文法的化简与范式

文法的化简：

消除无用符号：从文法开始符号派生到终结符串的过程中用不到

可达的(reachable)：$S \overset{*}{\Rightarrow} \alpha X\beta$，则X是可达的

产生的(generating)：$X \overset{*}{\Rightarrow} w(w \in T^*)$，则X是产生的

有用的：同时为可达的和产生的，否则称X为无用符号

消除无用符号：删除全部含有无用符号的产生式

$A \to \alpha \in P$ 且 α 中符号都是产生的, 则 A 是产生的. (也包括 $\alpha = \varepsilon$ 时)

$A \to \alpha \in P$ 且 A 是可达的, 则 α 中符号都是可达的.

※先寻找并消除全部“非产生的”符号，再寻找并消除全部“非可达的”符号，否则可能消除不完整

消除ε产生式：形如 $A \to \varepsilon$ 的产生式，除了 ε 自身,没有贡献语言中其他的串

如果 $B \to \alpha$ 且 α 中的每个符号都是可空的, 则 B 是可空的.

替换带有可空变元的产生式：

将含有可空变元的一条产生式 $A \to X1X2 \cdots Xn$, 用一组产生式 $A \to Y1Y2 \cdots Yn$ 代替, 其中:① 若 Xi 不是可空的, Yi 为 Xi ② 若 Xi 是可空的, Yi 为 Xi 或 ε，但 Yi 不能全为 ε.

消除单元产生式(unit productions)：仅仅增加了推导 (或归约) 的步骤

单元对：若$A \Rightarrow B$, 则称 [A,B] 为单元对. $A \to B \in P$, 则 [A,B] 是单元对;

单元对具有传递性。

对每个单元对[A,B]，删除形为A→B的单元产生式，并将B产生式复制给A

定理：每个非空的 CFL 都能被一个不带无用符号的 CFG 定义.

定理：任何 CFG G, 都存在一个不带无用符号和 ε-产生式的 CFG G' ,

使 $L(G') = L(G) - \{\varepsilon\}$.

定理：每个不带 ε 的 CFL 都可由一个不带无用符号, ε-产生式和单元产生式的文法定义.

文法的范式：

乔姆斯基范式：CNF,(Chomsky Normal Form)

每个不带 ε 的 CFL 都可由这样的 CFG G 定义, G 中每个产生式都形为

$A \to BC$ 或 $A \to a$,其中 A, B 和 C 都是变元, a 是终结符

利用 CNF 派生长度为 n 的串, 刚好需要 2n−1 步

CFG转为CNF的方法：

①将产生式$A \to X_1X_2 \cdots X_m (m \geq 2)$中每个终结符a替换为新变元$C_a$，并增加新产生式$C_a \to a$

②引入新变元$D_1, D_2, \cdots, D_{m-2}$，将产生式$A \to B_1B_2 \cdots B_m (m > 2)$替换为一组产生式:$A \to B_1D_1, D_1 \to B_2D_2, \cdots, D_{m-2} \to B_{m-1}B_m$

格雷巴赫范式：GNF,(Greibach Normal Form)

每个不带 ε 的 CFL 都可由这样的 CFG G 定义, G 中每个产生式都形为

$A \to a\alpha$, 其中A 是变元, a 是终结符, α 是零或多个变元的串

长度为 n 的串的派生恰好是 n 步

GNF 每个产生式都会引入一个终结符

直接左递归：(Immediate left-recursion)：形式为 A → Aα 的产生式

消除直接左递归：

①若A产生式$A \rightarrow A\alpha_1 \left| A\alpha_2 \right| \cdots \left| A\alpha_n \right| \beta_1 \left| \beta_2 \right|_{\ldots} \left| \beta_m \right.$，其中 αi≠ε, βj 不以 A 开始

②引入新变元B，并用如下产生式替换：

$A \rightarrow \beta_1 | \beta_2 | \cdots | \beta_m | \beta_1 B | \beta_2 B | \cdots | \beta_m B$

$B \rightarrow \alpha_1 | \alpha_2 | \cdots | \alpha_n | \alpha_1 B | \alpha_2 B | \cdots | \alpha_n B$

间接左递归：(Indirect left-recursion)：

形如A → Bα | ... ； B → Cβ | ... ; C → Aγ | ...无法通过代换消除递归

消除间接左递归：

① 将文法中变元重命名为 A1, A2, ··· , An;

② 通过代入, 使产生式都形如$A_i \rightarrow A_j\alpha, A_i \rightarrow a\alpha$，且 i ≤ j;

③消除直接左递归$A_i \rightarrow A_j\alpha$, 再代入其他产生式

引理：若$A \rightarrow \alpha_1 B \alpha_2$为P中的一个产生式，且$B \rightarrow \beta_1 | \beta_2 | \cdots | \beta_n$是P中的全部B产生式，将产生式$A \rightarrow \alpha_1 B \alpha_2$从P中删除，并增加: $\mathrm{A} \rightarrow \alpha_1\beta_1\alpha_2 | \alpha_1\beta_2\alpha_2 | \cdots | \alpha_1\beta_n\alpha_2$，得到文法G',则G'=G

触发指数时间：解析某些字符串，如果选择的产生式不正确, 会发生回溯, 甚至要回溯到初始状态, 复杂度是指数时间