上下文无关文法

- 上下文无关文法
- 语法分析树
- 文法和语言的歧义性
- 文法的化简与范式
  - 消除无用符号
  - 消除 ε-产生式
  - 消除单元产生式
  - 乔姆斯基范式
  - 格雷巴赫范式

为什么要化简

- 典型问题：给定 CFG G 和串 w，判断 $w \in \mathbf{L}(G)$?
- 编译器设计和自然语言处理的基本问题之一
- 但文法的形式非常自由，过于复杂不易于自动处理
- 以不改变语言为前提，化简文法和限制文法的格式

例 7. 如下文法中，有无意义的变元和产生式

$$
\begin{aligned}
S &\to 0DS1D \mid B \mid \varepsilon \\
B &\to BC1 \mid 0CBC \\
A &\to A0 \mid A1 \mid \varepsilon \\
C &\to D \\
D &\to \varepsilon
\end{aligned}
$$

变元A和B

对文法定义的语言无贡献

$S \nRightarrow A$ $B \nRightarrow V$

文法的化简

1. 消除无用符号: 对文法定义语言没有贡献的符号
2. 消除 ε 产生式: $A \to \varepsilon$ (得到语言 $L - \{\varepsilon\}$)
3. 消除单元产生式: $A \to B$

无用符号

定义

CFG $G = (V, T, P, S)$, 符号 $X \in (V \cup T)$:

❶ 如果 $S \overset{*}{\Rightarrow} \alpha X \beta$, 称 X 是可达的;

❷ 如果 $\alpha X \beta \overset{*}{\Rightarrow} w$ $(w \in T^*)$, 称 X 是产生的;

❸ 如果 X 同时是产生的和可达的, 即 既·又

$$S \overset{*}{\Rightarrow} \alpha X \beta \overset{*}{\Rightarrow} w \quad (w \in T^*),$$

则称 X 是有用的, 否则称 X 为无用符号.

消除无用符号

S 也是吗　　可作终点　　变元有穷

迭代会停止

计算“产生的”符号集

1. 每个 T 中的符号都是产生的;
2. $A \to \alpha \in P$ 且 α 中符号都是产生的, 则 A 是产生的.

计算“可达的”符号集

可作起点

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

删除全部含有“非产生的”和“非可达的”符号的产生式

定理 18

每个非空的 CFL 都能被一个不带无用符号的 CFG 定义.

略证

注意

- 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

顺序!

例 8. 消除如下文法无用符号

$$S \to \cancel{A}B \mid a$$
$$A \to b$$

B是非产生的

$S \to a$

~~$A \to b$~~

A是非可达的

$S \to a$

消除 ε-产生式

定义

文法中形如 $A \to \varepsilon$ 的产生式称为ε-产生式.
如果变元 $A \overset{*}{\Rightarrow} \varepsilon$, 称 A 是可空的.

- ε-产生式在文法定义语言时, 除产生空串外没有其他帮助
- 对于 CFL L, 消除其文法中全部的 ε-产生式后, 得到语言 $L - \{\varepsilon\}$

1 确定“可空变元”

❶ 如果 $A \to \varepsilon$, 则 A 是可空的;

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

2 替换产生式

将含有可空变元的一条产生式 $A \to X_1X_2\cdots X_n$, 用一组产生式 $A \to Y_1Y_2\cdots Y_n$ 代替, 其中:

❶ 若 X_i 不是可空的, Y_i 为 X_i; 保留

❷ 若 X_i 是可空的, Y_i 为 X_i 或 ε;

❸ 但 Y_i 不能全为 ε.

全都写出来

定理 19

任何 CFG G, 都存在一个不带无用符号和 ε-产生式的 CFG G', 使 $\mathbf{L}(G') = \mathbf{L}(G) - \{\varepsilon\}$.

例 9. 消除 CFG $G = (\{S, A, B\}, \{a, b\}, P, S)$ 的 ε-产生式.

①

$$\begin{aligned} \underline{S} &\to AB \\ \underline{A} &\to AaA \mid \cancel{\varepsilon} \\ \underline{B} &\to BbB \mid \cancel{\varepsilon} \end{aligned}$$

②

$$\begin{aligned} S &\to AB \mid A \mid B \\ A &\to AaA \mid aA \mid Aa \mid a \\ B &\to BbB \mid Bb \mid bB \mid b \end{aligned}$$

解: CFG G' 为

$$\begin{aligned} S &\to AB \mid A \mid B \\ A &\to AaA \mid Aa \mid aA \mid a \\ B &\to BbB \mid Bb \mid bB \mid b \end{aligned}$$

消除单元产生式

派

确定“单元对”

如果有 $A \overset{*}{\Rightarrow} B$，则称 $[A, B]$ 为单元对.

传

1. $A \to B \in P$，则 $[A, B]$ 是单元对;
2. 若 $[A, B]$ 和 $[B, C]$ 都是单元对，则 $[A, C]$ 是单元对.

右 → 右

消除单元产生式

1. 删除全部形为 $A \to B$ 的单元产生式;
2. 对每个单元对 $[A, B]$，将 B 的产生式复制给 A.

定理 20

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来定义.

例 10. 消除文法的单元产生式

$$
\begin{aligned}
S &\to A \mid B \mid 0S1 \\
A &\to 0A \mid 0 \\
B &\to 1B \mid 1
\end{aligned}
$$

(S, A)

(S, B)

解: 单位对为 $[S, A]$ 和 $[S, B]$, 带入得:

$$
\begin{aligned}
S &\to 0S1 \\
S &\to 0A \mid 0 \\
S &\to 1B \mid 1
\end{aligned}
\qquad
\begin{aligned}
A &\to 0A \mid 0 \\
B &\to 1B \mid 1
\end{aligned}
$$

剔去掉

文法化简的可靠顺序

1. 消除 ε-产生式;
2. 消除单元产生式;
3. 消除非产生的无用符号;
4. 消除非可达的无用符号.

限制文法格式

将任意形式的文法转换为:

1. 乔姆斯基范式 (CNF, Chomsky Normal Form)
2. 格雷巴赫范式 (GNF, Greibach Normal Form)

乔姆斯基范式

定理 21 (乔姆斯基范式 CNF)

每个不带 ε 的 CFL 都可以由这样的 CFG G 定义，G 中每个产生式的形式都为

$$A \to BC \quad \text{或} \quad A \to a$$

这里的 A, B 和 C 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
- 因此存在算法判断任意字符串 w 是否在给定的 CFL 中
- 利用 CNF 的多项式时间解析算法 — CYK 算法

CFG 转为 CNF 的方法

❶ 将产生式

$$A \to X_1 X_2 \cdots X_m \quad (m \geq 2)$$

中每个终结符 a 替换为新变元 C_a,

❷ 增加新产生式

$$C_a \to a,$$

❸ 引入新变元 $D_1, D_2, \cdots, D_{m-2}$, 将产生式

$$A \to B_1 B_2 \cdots B_m \quad (m \geq 2)$$

替换为一组级联产生式

$$\begin{aligned} A &\to B_1 D_1 \\ D_1 &\to B_2 D_2 \\ &\cdots \\ D_{m-2} &\to B_{m-1} B_m. \end{aligned}$$

例 11. CFG $G = (\{S, A, B\}, \{a, b\}, P, S)$, 产生式集合 P 为:

$$
\begin{aligned}
S &\to bA \mid aB \\
A &\to bAA \mid aS \mid a \\
B &\to aBB \mid bS \mid b
\end{aligned}
$$

请设计等价的 CNF 文法.

解: CNF 为

$$
\begin{aligned}
&S \to C_bA \mid C_aB \\
&A \to C_aS \mid C_bD_1 \mid a \qquad D_1 \to AA \qquad C_a \to a \\
&B \to C_bS \mid C_aD_2 \mid b \qquad D_2 \to BB \qquad C_b \to b
\end{aligned}
$$

格雷巴赫范式

定理 22 (格雷巴赫范式 GNF)

每个不带 ε 的 CFL 都可以由这样的 CFG G 定义, G 中每个产生式的形式都为

$$A \to a\alpha$$

其中 A 是变元, a 是终结符, α 是零或多个变元的串.

- GNF 每个产生式都会引入一个终结符
- 长度为 n 的串的派生恰好是 n 步

例 12. 将以下文法转换为 GNF.

$$
\begin{aligned}
&S \to AB \\
&A \to aA \mid bB \mid b \\
&B \to b
\end{aligned}
$$

解: GNF 为

$$
\begin{aligned}
&S \to aAB \mid bBB \mid bB \\
&A \to aA \mid bB \mid b \\
&B \to b
\end{aligned}
$$

直接左递归

定义

文法中形式为 $A \to A\alpha$ 的产生式，称为直接左递归.

消除直接左递归

❶ 若 A 产生式

$$A \to A\alpha_1 \mid A\alpha_2 \mid \cdots \mid A\alpha_n \mid \beta_1 \mid \beta_2 \mid \cdots \mid \beta_m$$

其中 $\alpha_i \neq \varepsilon$, β_j 不以 A 开始;

❷ 引入新变元 B, 并用如下产生式替换

$$A \to \beta_1 \mid \beta_2 \mid \cdots \mid \beta_m \mid \beta_1 B \mid \beta_2 B \mid \cdots \mid \beta_m B$$

$$B \to \alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_n \mid \alpha_1 B \mid \alpha_2 B \mid \cdots \mid \alpha_n B$$

间接左递归

定义

文法中如果有形式为

$$
\begin{aligned}
A &\to B\alpha \mid \ldots \\
B &\to A\beta \mid \ldots
\end{aligned}
$$

的产生式，称为间接左递归.

- 会有 $A \Rightarrow B\alpha \Rightarrow A\beta\alpha$, 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 $A_1, A_2, \cdots, A_n$;
2. 通过代入, 使产生式都形如

$$\begin{aligned} A_i &\to A_j\alpha \\ A_i &\to a\alpha \end{aligned}$$

   但要求 $i \le j$;
3. 消除直接左递归 $A_i \to A_i\beta$, 再代入其他产生式.

例 13. Convert the following grammar to GNF.

$$\begin{aligned} S &\to AB \\ A &\to BS \mid b \\ B &\to SA \mid a \end{aligned}$$

解: 1. 重命名变元, 代换 $i > j$ 的 A_j

$$
\begin{aligned}
A_1 &\to A_2A_3 \\
A_2 &\to A_3A_1 \mid b \\
A_3 &\to a \mid \cancel{A_1A_2} \mid \cancel{A_2A_3A_2} \mid \\
&\quad A_3A_1A_3A_2 \mid bA_3A_2
\end{aligned}
$$

2. 消除直接左递归

$$
\begin{aligned}
A_1 &\to A_2A_3 \\
A_2 &\to A_3A_1 \mid b \\
A_3 &\to bA_3A_2 \mid a \mid bA_3A_2B_1 \mid aB_1 \\
B_1 &\to A_1A_3A_2 \mid A_1A_3A_2B_1
\end{aligned}
$$

3. A_3 代入到 A_2, A_2 代入到 A_1, A_1 代入 B_1

$$
\begin{aligned}
A_3 &\to bA_3A_2 \mid a \mid bA_3A_2B_1 \mid aB_1 \\
A_2 &\to bA_3A_2A_1 \mid aA_1 \mid bA_3A_2B_1A_1 \mid aB_1A_1 \mid b \\
A_1 &\to bA_3A_2A_1A_3 \mid aA_1A_3 \mid bA_3A_2B_1A_1A_3 \mid aB_1A_1A_3 \mid bA_3 \\
B_1 &\to bA_3A_2A_1A_3A_3A_2 \mid aA_1A_3A_3A_2 \mid bA_3A_2B_1A_1A_3A_3A_2 \mid \\
&\quad aB_1A_1A_3A_3A_2 \mid bA_3A_3A_2 \mid bA_3A_2A_1A_3A_3A_2B_1 \mid aA_1A_3A_3A_2B_1 \mid \\
&\quad bA_3A_2B_1A_1A_3A_3A_2B_1 \mid aB_1A_1A_3A_3A_2B_1 \mid bA_3A_3A_2B_1
\end{aligned}
$$