Глава 4. Восходящий синтаксический анализ

4.2. Грамматики простого предшествования

Одним из наиболее легких подходов к решению проблемы поиска и своевременной свертки основы является реализация восходящего синтаксического анализа для небольшого класса КС-грамматик, называемых *грамматиками простого предшествования*. Технология синтаксического анализа для таких грамматик предполагает введение специальных бинарных отношений между каждой парой символов грамматики (как терминалов, так и нетерминалов). Эти отношения управляют выбором основ сентенциальных форм для последующей их свертки.

Основным недостатком этого метода является применимость его лишь в узком классе грамматик простого предшествования.

4.2.1. Отношения предшествования

Пусть $G = (V_T, V_N, P, S)$ – контекстно-свободная грамматика, а строка $\alpha XY\beta$ – правосторонняя сентенциальная форма, где $\alpha, \beta \in (V_T \cup V_N)^*$, $X, Y \in V_T \cup V_N$. В некоторый момент (на одном из этапов процесса последовательных сверток сентенциальной формы) возникает одна из следующих возможных ситуаций:

1. Y – самый левый символ (*заголовок*) основы сентенциальной формы, а X не входит в основу. В этом случае говорят, что символ Y *предшествует* символу X (поскольку символ Y должен быть свернут раньше символа X), и записывают в виде $X \lessdot Y$.

2. X и Y входят в одну и ту же основу. В этом случае говорят, что X и Y имеют *равное предшествование* (поскольку сворачиваются одновременно), и записывают в виде $X \doteq Y$.

3. X – последний символ (*окончание*) основы, а Y не входит в основу. В этом случае говорят, что символ X *предшествует* символу Y (поскольку символ X должен быть свернут раньше символа Y), и записывают в виде $X \gtrdot Y$.

Отношения $\lessdot$, $\doteq$, $\gtrdot$ называются *отношениями предшествования*. Следует заметить, что, хотя эти отношения похожи на арифметические отношения $<$, $=$, $>$, они имеют совершенно иные свойства. В частности, они не обладают свойствами коммутативности и ассоциативности. Из отношения $X \gtrdot Y$, не следует, что существует отношение $Y \lessdot X$. Эти отношения не являются симметричными, из отношения $X \doteq Y$ не следует $Y \doteq X$. Для одной и той же грамматики может быть так, что $X \lessdot Y$ и $X \gtrdot Y$, или для некоторых пар символов не выполняется ни одно из отношений предшествования.

Формально эти отношения для символов $X, Y \in V_T \cup V_N$ грамматики определяются следующим образом:

1. $X \doteq Y$, если существует некоторая продукция $A \to \alpha XY\beta$, $A \in V_N$, $\alpha, \beta \in (V_T \cup V_N)^*$. Это значит, что в правосторонней сентенциальной форме X и Y входят в одну и ту же основу.

2. $X \lessdot Y$, если существует некоторая продукция $A \to \alpha XB\beta$, $A, B \in V_N$, $\alpha, \beta \in (V_T \cup V_N)^*$, такая, что $B \overset{+}{\Rightarrow} Y\delta$, $\delta \in (V_T \cup V_N)^*$. Это значит, что в правосторонней сентенциальной форме основа начинается с символа Y (Y является заголовком основы).

3. $X \gtrdot Y$, если существует некоторая продукция $A \to \alpha BZ\beta$, $A, B \in V_N$, $Z \in V_T \cup V_N$, $\alpha, \beta \in (V_T \cup V_N)^*$, такая, что $B \overset{+}{\Rightarrow} \gamma X$ и $Z \overset{*}{\Rightarrow} Y\delta$, $\gamma, \delta \in (V_T \cup V_N)^*$. Это значит, что в правосторонней сентенциальной форме основа завершается символом X (X является окончанием основы). Следует заметить, что в правосторонней сентенциальной форме справа от основы может быть только терминальная строка. Поэтому в данном случае символ Y может быть только терминалом, т. е. $Y \in V_T$, и отношение $\gtrdot$ определяется на множестве $(V_T \cup V_N) \times V_T$. Заметим также, что если $Y\delta$ выводится из Z за нуль шагов, то $Z = Y$.

Контекстно-свободная грамматика $G = (V_T, V_N, P, S)$ называется *грамматикой простого предшествования*, если:

1) не содержит ε-продукций;

2) никакие две продукции грамматики не имеют совпадающих правых частей (грамматики, в которых нет двух продукций с одинаковыми правыми частями, называются *обратимыми*);

3) любые два символа, составляющие элемент множества $(V_T \cup V_N) \times (V_T \cup V_N)$, связаны одним и тем же отношением предшествования.

Отношения предшествования обычно записывают в виде *матрицы предшествования*, строки и столбцы которой соответствуют символам грамматики. На пересечении i-й строки и j-го столбца записывается отношение предшествования между соответствующими символами грамматики. Элементами матрицы являются знаки $\lessdot$, $\doteq$, $\gtrdot$ или «пусто». Последний случай означает, что соответствующие символы не могут стоять рядом ни в одной сентенциальной форме.

4.2.2. Вычисление отношений предшествования

Формальный процесс вычисления отношений предшествования для символов $X, Y \in V_T \cup V_N$ заданной КС-грамматики можно представить такой последовательностью действий (в описании действий строки $\alpha, \beta \in (V_T \cup V_N)^*$, $A \in V_N$):

1. Определить для каждого нетерминала X грамматики множество $L(X) = \{Y \mid X \overset{+}{\Rightarrow} Y\alpha\}$, т. е. множество символов грамматики (как терминалов, так и нетерминалов), с которых могут начинаться строки, выводимые из нетерминала X. Для этого необходимо построить отношение <LEFT>, определяемое следующим образом: X <LEFT> Y, если в грамматике существует продукция вида $X \to Y\beta$. Затем вычислить отношение <LEFT>$^+$ как транзитивное замыкание отношения <LEFT>. Тогда $L(X)$ есть множество символов Y, для которых выполняется отношение X <LEFT>$^+$ Y.

2. Вычислить отношение <LEFT>* как рефлексивно-транзитивное замыкание отношения <LEFT>. Очевидно, что отношение <LEFT>* легко вычисляется по отношению <LEFT>$^+$, поскольку имеет место соотношение <LEFT>* = <LEFT>$^+ \cup I$, где I – отношение тождественности. Отношение <LEFT>* понадобится для вычисления отношения $\gtrdot$.

3. Определить для каждого нетерминала X грамматики множество $R(X) = \{Y \mid X \overset{+}{\Rightarrow} \alpha Y\}$, т. е. множество символов грамматики, являющихся крайними справа в строках, выводимых из нетерминала X. Для этого необходимо построить отношение <RIGHT>, определяемое следующим образом: Y <RIGHT> X, если в грамматике существует продукция вида $X \rightarrow \beta Y$. Затем вычислить отношение <RIGHT>$^{+}$ как транзитивное замыкание отношения <RIGHT>. Тогда $R(X)$ есть множество символов Y, для которых выполняется отношение Y <RIGHT>$^{+}$ X.

4. Построить для всех символов грамматики отношение $\doteq$ по его определению, т. е. $X \doteq Y$, если в грамматике существует продукция вида $A \rightarrow \alpha XY\beta$.

5. Вычислить отношение $\lessdot$. Из его формального определения следует, что $X \lessdot Y$, если в грамматике имеется продукция вида $A \rightarrow \alpha XB\beta$, где $B \in V_N$, и $Y \in L(B)$. Таким образом, отношение $\lessdot$ можно вычислить как произведение отношений $\doteq$ и <LEFT>$^{+}$, т. е. ($\lessdot$) = ($\doteq$) (<LEFT>$^{+}$).

6. Вычислить отношение $\gtrdot$. Из его формального определения следует, что $X \gtrdot Y$ (напомним, что отношение определено только для $Y \in V_T$), если существует продукция

а) вида $A \rightarrow \alpha BY\beta$, где $B \in V_N$, $Y \in V_T$, и $X \in R(B)$;

б) вида $A \rightarrow \alpha BZ\beta$, где $B, Z \in V_N$, и $X \in R(B)$, $Y \in L_T(Z)$, где $L_T(Z) \subseteq L(Z)$ – подмножество терминалов множества $L(Z)$.

Таким образом, ($\gtrdot$) = (<RIGHT>$^{+}$) ($\doteq$) (<LEFT>*). Поскольку Y может быть только терминалом, при вычислении произведения отношений следует рассматривать только те отношения X <LEFT>* Y, где $Y \in V_T$.

7. Построить матрицу предшествования, объединив матрицы отношений $\doteq$, $\lessdot$ и $\gtrdot$ в одну и заменив единицы на соответствующие обозначения ($\doteq$, $\lessdot$, $\gtrdot$) отношений.

Пример процесса вычисления отношений предшествования для грамматики с продукциями

$S \to AB$

$A \to aA \mid a$

$B \to bB \mid b$

представлен на рис. 4.3.

$$S \to AB$$
$$A \to aA \mid a$$
$$B \to bB \mid b$$

$<\mathrm{LEFT}>$

	S	A	B	a	b
S		1			
A				1	
B					1

$<\mathrm{LEFT}>^{+}$

	S	A	B	a	b
S		1		1	
A				1	
B					1

$<\mathrm{LEFT}>^{*}$

	S	A	B	a	b
S	1	1		1	
A		1		1	
B			1		1
a				1	
b					1

$<\mathrm{RIGHT}>$

	S	A	B
S			
A		1	
B	1		1
a		1	
b			1

$<\mathrm{RIGHT}>^{+}$

	S	A	B
S			
A		1	
B	1		1
a		1	
b	1		1

$\doteq$

	S	A	B	a	b
S					
A			1		
B					
a		1			
b			1		

$\lessdot$

	S	A	B	a	b
S					
A					1
B					
a				1	
b					1

$\gtrdot$

	S	A	B	a	b
S					
A					1
B					
a					1
b					

Матрица предшествования

	S	A	B	a	b
S					
A			$\doteq$		$\lessdot$ $\gtrdot$
B					
a		$\doteq$		$\lessdot$	$\gtrdot$
b			$\doteq$		$\lessdot$

Рис. 4.3. Процесс вычисления отношений предшествования

Из отношений <LEFT>$^{+}$ и <RIGHT>$^{+}$ следует, что

$L(S) = \{A, a\}$, $L(A) = \{a\}$, $L(B) = \{b\}$,

$R(S) = \{B, b\}$, $R(A) = \{A, a\}$, $R(B) = \{B, b\}$.

Грамматика не является грамматикой простого предшествования, поскольку она не удовлетворяет третьему условию, требующему, чтобы любые два символа грамматики были связаны одним и тем же отношением предшествования. В нашем случае одновременно выполняются $A \lessdot b$ и $A \gtrdot b$. Действительно $A \lessdot b$, так как имеется продукция $S \to AB$ и $b \in L(B)$, т. е. $B \overset{+}{\Rightarrow} b$, а $A \gtrdot b$, так как существует продукция $S \to AB$, такая, что $A \in R(A)$ и $b \in L(B)$, т. е. $A \overset{+}{\Rightarrow} aA$ и $B \overset{*}{\Rightarrow} b$.

Более кратко для вычисления вручную (пусть имеется продукция вида $A \to \alpha XY\beta$):

1) $X \doteq Y$ для любых символов (терминалов и нетерминалов),

2) если $Y \in V_N$, то X связано отношением $\lessdot$ со всеми элементами из $L(Y)$,

3) если $X \in V_N$, то

   а) если $Y \in V_T$, то все элементы из $R(X)$ связаны отношением $\gtrdot$ с элементом Y,

   б) если $Y \in V_N$, то все элементы из $R(X)$ связаны отношением $\gtrdot$ со всеми терминалами из $L(Y)$.