Teopuя автоматов и формальных языков Parsing expression grammar

Автор: Екатерина Вербицкая

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

1 ноября 2016г.

В предыдущей серии

- Нисходящий анализ
- Множество FIRST
- Множество FOLLOW
- Удаление левой рекурсии
- Факторизация грамматики
- *LL*(1)-таблица
- LL(1) синтаксический анализ

В предыдущей серии

- Преимущества LL(1) синтаксического анализа
 - Прост и понятен
 - ▶ Разбор за O(n), где n длина входной строки
 - Применим для многих практически полезных задач (с оговорками)
- Недостатки LL(1) синтаксического анализа
 - Необходимость преобразовывать грамматику
 - Достаточно небольшой класс языков, которые действительно можно разобрать синтаксическим анализатором

- Грамматика для условного оператора if
 - ightharpoonup E
 ightarrow b
 - ▶ $S \rightarrow \underline{if} E \underline{then} S \underline{else} S | \underline{if} E \underline{then} S | a$
- if b then a if b then a else a

- Грамматика для условного оператора *if*
 - **▶** *E* → *b*
 - ▶ $S \rightarrow \underline{if} E \underline{then} S \underline{else} S | \underline{if} E \underline{then} S | a$
- if b then a if b then a else a
 - ightharpoonup if $b \underline{then} a (\underline{if} b \underline{then} a \underline{else} a)$
 - ightharpoonup if $b \underline{then} a (\underline{if} b \underline{then} a) \underline{else} a$

- Грамматика для условного оператора if
 - E → b
 - ▶ $S \rightarrow \underline{if} E \underline{then} S \underline{else} S | \underline{if} E \underline{then} S | a$
- Факторизуем
 - ► E → b
 - ▶ $S \rightarrow \underline{if} E \underline{then} S S' \mid a$
 - $S' \rightarrow \underline{\textit{else}} \, S \mid \varepsilon$
- Строим *LL*(1)-таблицу

N	FIRST	FOLLOW	а	b	if	then	else	\$
Ε	{b}	{ <u>then</u> }		b				

- Грамматика для условного оператора if
 - E → b
 - ▶ $S \rightarrow \underline{if} E \underline{then} S \underline{else} S | \underline{if} E \underline{then} S | a$
- Факторизуем
 - E → b
 - ▶ $S \rightarrow if E then S S' \mid a$
 - $S' \rightarrow \underline{\textit{else}} \, S \mid \varepsilon$
- Строим *LL*(1)-таблицу

Ν	FIRST	FOLLOW	a	b	if	then	else	\$
Ε	{b}	{ <u>then</u> }		b				
S	{ <u>if</u> , a}	{\$, <u>else</u> }	а		if E then S S'			

- Грамматика для условного оператора if
 - E → b
 - ▶ $S \rightarrow \underline{if} E \underline{then} S \underline{else} S | \underline{if} E \underline{then} S | a$
- Факторизуем
 - ightharpoonup E
 ightarrow b
 - ▶ $S \rightarrow if E then S S' \mid a$
 - ▶ $S' \rightarrow \underline{\textit{else}} \, S \, | \, \varepsilon$
- Строим *LL*(1)-таблицу

N	FIRST	FOLLOW	а	b	if	then	else	\$
Ε	{b}	{ <u>then</u> }		b				
S	{ <u>if</u> , a}	{\$, <u>else</u> }	а		if E then S S'			
5'	{ <u>else</u> }	{\$, <u>else</u> }					$\frac{\textit{else}}{arepsilon} \mathcal{S}$	ε

Две записи в одной ячейке!

Как бороться с dangling else problem

- ullet Приоритет $S' o \underline{\mathit{else}}\, S$ перед S' o arepsilon
- Изменить синтаксис
 - ▶ Добавить <u>endif</u> выражение
 - ▶ Запретить <u>if</u> сразу после <u>then</u>
 - ▶ Обязать использовать скобки, если есть *else*
 - ▶ Обязать использовать ветку <u>else</u>
 - Использовать разные ключевые слова для выражений:
 <u>if</u> E <u>then</u> S <u>else</u> и <u>if</u> E <u>do</u> S
- ...

Как бороться с dangling else problem

- ullet Приоритет $S' o \underline{\mathit{else}}\, S$ перед S' o arepsilon
- Изменить синтаксис
 - ▶ Добавить <u>endif</u> выражение
 - ► Запретить <u>if</u> сразу после <u>then</u>
 - ► Обязать использовать скобки, если есть *else*
 - ▶ Обязать использовать ветку <u>else</u>
 - Использовать разные ключевые слова для выражений:
 if E then S else и if E do S
- . . .
- Использовать другой формализм, в котором нет неоднозначностей!

Parsing Expression Grammars

- ullet РЕG G четверка (V, T, P, p_S) , где
 - ▶ V конечное множество нетерминалов
 - ► Т алфавит (конечное множество терминалов)
 - ▶ P функция из V в выражения (parsing expression)
 - ▶ p_S стартовое выражение

Parsing Expression Grammars

- \bullet PEG G четверка (V, T, P, p_S) , где
 - ▶ V конечное множество нетерминалов
 - ▶ Т алфавит (конечное множество терминалов)
 - ▶ P функция из V в выражения (parsing expression)
 - $ightharpoonup p_S$ стартовое выражение
- Parsing expression
 - ▶ Пустая строка ε
 - Терминал а
 - Нетерминал А
 - ▶ Последовательность p_1p_2 , где p_1, p_2 parser expression
 - ightharpoonup Упорядоченный выбор p_1/p_2 , где p_1,p_2 parser expression
 - ▶ 0-или-больше p^* , где p parser expression
 - ▶ Предикат He !p, где p parser expression

Пример Parsing Expression Grammars

- \bullet $E \rightarrow b$
- $S \rightarrow \underline{if} E \underline{then} S \underline{else} S / \underline{if} E \underline{then} S / a$

Интуиция

- С помощью фрагмента PEG разбираем ("матчим") *префикс* строки пока это возможно
- Неразобранный суффикс матчим следующим фрагментом PEG

Пример Parsing Expression Grammars

- $E \rightarrow b$
- $S \rightarrow \underline{if} E \underline{then} S \underline{else} S / \underline{if} E \underline{then} S / a$

```
• S ("if b then a else a") \rightarrow

(if E then S else S / if E then S / a) ("if b then a else a") \rightarrow

(if E then S else S) ("if b then a else a") \rightarrow

(E then S else S) ("E then E else E a") \rightarrow

(E then E else E ("E then E else E a") \rightarrow

(E else E ("E then E else E / if E then E / E ("E then E \text{ } E \text{ ("E then E else E ) ("E then E else E els
```

В чем разница между PEG и CFG?

- Выбор упорядочен!
 - Если получилось разобрать первой альтернативой, то вторая не рассматривается вообще
 - ▶ Как следствие, принципиально не бывает неоднозначностей

Формальнее: отношение PEG

$$G[p] \ xy \overset{\text{\tiny PEG}}{\leadsto} (y, x')$$

- Выражение p парсит строку xy, съедая x и оставляя y, возвращая x' как результат
- Если справа fail, значит, распарсить строку не удалось

Операционная семантика PEG: пустая строка

$$\overline{G[\varepsilon] \ x \overset{\text{PEG}}{\leadsto} (x, \, \varepsilon)}$$

Операционная семантика PEG: терминал

$$G[a] \ ax \stackrel{\text{peg}}{\leadsto} (x, \, a)$$

Операционная семантика PEG: терминал

$$\frac{\overline{G[a]\ ax} \overset{\text{PEG}}{\leadsto} (x,\, a)}{\overline{G[b]\ ax} \overset{\text{PEG}}{\leadsto} \text{fail}} \;, \, b \neq a$$

Операционная семантика PEG: терминал

$$\frac{\overline{G[a]} \ ax \overset{\text{PEG}}{\leadsto} (x, \, a)}{\overline{G[b]} \ ax \overset{\text{PEG}}{\leadsto} \text{fail}} \, , \, b \neq a$$

$$\frac{\overline{G[a]} \ \varepsilon \overset{\text{PEG}}{\leadsto} \text{fail}}{}$$

Операционная семантика PEG: переменная

$$\frac{G[P(A)] \ xy \overset{\text{\tiny PEG}}{\leadsto} (y,\,x')}{G[A] \ xy \overset{\text{\tiny PEG}}{\leadsto} (y,\,A[x'])}$$

Операционная семантика PEG: переменная

$$\frac{G[P(A)] \ xy \overset{\text{PEG}}{\leadsto} (y, x')}{G[A] \ xy \overset{\text{PEG}}{\leadsto} (y, A[x'])}$$

$$\frac{G[P(A)] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[A] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}$$

Операционная семантика PEG: последовательность

$$\frac{G[p_1] \ xyz \overset{\text{PEG}}{\leadsto} (yz, \, x') \quad G[p_2] \ yz \overset{\text{PEG}}{\leadsto} (z, \, y')}{G[p_1 \, p_2] \ xyz \overset{\text{PEG}}{\leadsto} (z, \, x'y')}$$

Операционная семантика PEG: последовательность

$$\frac{G[p_1] \ xyz \overset{\text{PEG}}{\leadsto} (yz, x') \quad G[p_2] \ yz \overset{\text{PEG}}{\leadsto} (z, y')}{G[p_1 \ p_2] \ xyz \overset{\text{PEG}}{\leadsto} (z, x'y')}$$

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} (y, x') \quad G[p_2] \ y \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p_1 \ p_2] \ xy \overset{\text{PEG}}{\leadsto} \text{fail}}$$

Операционная семантика PEG: последовательность

$$\frac{G[p_1] \ xyz \overset{\text{PEG}}{\leadsto} (yz, x') \quad G[p_2] \ yz \overset{\text{PEG}}{\leadsto} (z, y')}{G[p_1 \ p_2] \ xyz \overset{\text{PEG}}{\leadsto} (z, x'y')}$$

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} (y, x') \quad G[p_2] \ y \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p_1 \ p_2] \ xy \overset{\text{PEG}}{\leadsto} \text{fail}}$$

$$\frac{G[p_1] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p_1 \ p_2] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}$$

Операционная семантика PEG: выбор

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}{G[p_1 \, / \, p_2] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}$$

Операционная семантика PEG: выбор

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}{G[p_1 \, / \, p_2] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}$$

$$\frac{G[p_1] \ x \overset{\text{PEG}}{\leadsto} \text{fail} \quad G[p_2] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p_1 \, / \, p_2] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}$$

Операционная семантика PEG: выбор

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}{G[p_1 \, / \, p_2] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}$$

$$\frac{G[p_1] \ x \overset{\text{PEG}}{\leadsto} \text{fail} \quad G[p_2] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p_1 \, / \, p_2] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}$$

$$\frac{G[p_1] \ xy \overset{\text{PEG}}{\leadsto} \text{fail} \quad G[p_2] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}{G[p_1 \, / \, p_2] \ xy \overset{\text{PEG}}{\leadsto} (y, \, x')}$$

Операционная семантика PEG: предикат не

$$\frac{G[p] \ x \overset{\mathrm{PEG}}{\leadsto} \mathtt{fail}}{G[!p] \ x \overset{\mathrm{PEG}}{\leadsto} (x, \, \varepsilon)}$$

Операционная семантика PEG: предикат не

$$\frac{G[p] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[!p] \ x \overset{\text{PEG}}{\leadsto} (x, \varepsilon)}$$

$$\frac{G[p] \ xy \overset{\text{PEG}}{\leadsto} (y, x')}{G[!p] \ xy \overset{\text{PEG}}{\leadsto} \text{fail}}$$

Операционная семантика PEG: повторение

$$\frac{G[p] \ x \overset{\text{\tiny PEG}}{\leadsto} \texttt{fail}}{G[p^*] \ x \overset{\text{\tiny PEG}}{\leadsto} (x, \, \varepsilon)}$$

Операционная семантика PEG: повторение

$$\frac{G[p] \ x \overset{\text{PEG}}{\leadsto} \text{fail}}{G[p^*] \ x \overset{\text{PEG}}{\leadsto} (x, \varepsilon)}$$

$$\frac{G[p] \ xyz \overset{\text{PEG}}{\leadsto} (yz, x') \quad G[p^*] \ yz \overset{\text{PEG}}{\leadsto} (z, y')}{G[p^*] \ xyz \overset{\text{PEG}}{\leadsto} (z, x'y')}$$

Преимущества и недостатки PEG

- Преимущества
 - Никаких неоднозначностей
 - Наглядность
 - Синтаксический анализатор и спецификация языка по сути одно и то же
 - ightharpoonup Синтаксический анализ за O(n)
- Недостатки
 - ▶ Жадность: разбирается наиболее большой префикс
 - ★ PEG a*a не может разобрать ничего
 - ▶ Никто не знает, какой именно класс языков можно распознать
 - ★ Все регулярные языки распознаются
 - ★ Все LL(k) языки распознаются
 - ★ Не все КС языки распознаются
 - \star Язык палиндромов S o xSx/x не распознается
 - ★ Язык $\{a^nb^nc^n \mid n \geq 0\}$ не KC распознается
 - ▶ Проблемы с левой рекурсией

Еще примеры: арифметические выражения

```
\begin{array}{ccc} \textit{Expr} & \rightarrow & \textit{Sum} \\ \textit{Sum} & \rightarrow & \textit{Product} \left( (+ / -) \, \textit{Product} \right)^* \\ \textit{Product} & \rightarrow & \textit{Value} \left( (* / /) \, \textit{Value} \right)^* \\ \textit{Value} & \rightarrow & \left[ 0 - 9 \right] / \left( \, \textit{Expr} \, \right) \end{array}
```

Еще примеры: вложенные комментарии

```
\begin{array}{ccc} \textit{Begin} & \rightarrow & (* \\ \textit{End} & \rightarrow & *) \\ & \textit{C} & \rightarrow & \textit{Begin N* End} \\ & \textit{N} & \rightarrow & \textit{C} \ (!\textit{Begin !End Z}) \\ & \textit{Z} & \rightarrow & \textit{any single character} \end{array}
```

Еще примеры: не КС язык

$$S \rightarrow \&(Ac) a^+ B!.$$

 $A \rightarrow aA?b$
 $B \rightarrow bB?c$

Борьба с левой рекурсией: ограниченная левая рекурсия

• A^n имеет не более n леворекурсивных вызовов A, A^0 всегда завершается ошибкой

$$E^{0}$$
 ::= fail
 E^{1} ::= $E^{0} + n/n = \bot + n/n = n$
 E^{2} ::= $E^{1} + n/n = n + n/n$
 E^{3} ::= $E^{2} + n/n = (n + n/n) + n/n$
...
 E^{n} ::= $E^{n-1} + n/n$

Борьба с левой рекурсией

- Ищем значение п для каждого леворекурсивного нетерминала
- Подбирается такая граница, чтобы префикс, обработанный правилом, имел максимальную длину
- Промежуточные значения сохраняются в табличку L
 - ▶ $L[(A,x) \rightarrow X](B,y) = L(B,y)$, если $B \neq A$ или $y \neq x$
 - $L[(A,x) \to X](A,x) = X$

$$(A, xyz) \notin \mathcal{L} \qquad G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} (yz, x')$$

$$G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto (yz, x')] \stackrel{\mathrm{INC}}{\leadsto} (z, (xy)')$$

$$G[A] \quad xyz \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} (z, A[(xy)'])$$

$$(A, xyz) \notin \mathcal{L} \qquad G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} (yz, x')$$

$$G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto (yz, x')] \stackrel{\mathrm{INC}}{\leadsto} (z, (xy)')$$

$$G[A] \quad xyz \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} (z, A[(xy)'])$$

$$(A, x) \notin \mathcal{L} \quad G[P(A)] \quad x \quad \mathcal{L}[(A, x) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}$$

 $G[A] \ x \ \mathcal{L} \overset{\text{PEG}}{\leadsto} \text{fail}$

$$(A, xyz) \notin \mathcal{L}$$
 $G[P(A)]$ xyz $\mathcal{L}[(A, xyz) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} (yz, x')$
 $G[P(A)]$ xyz $\mathcal{L}[(A, xyz) \mapsto (yz, x')] \stackrel{\mathrm{INC}}{\leadsto} (z, (xy)')$

$$G[A]$$
 xyz $\mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} (z, A[(xy)'])$

$$\underbrace{(A, x) \notin \mathcal{L} \ G[P(A)] \ x \ \mathcal{L}[(A, x) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}}_{G[A] \ x \ \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}}$$

$$\underbrace{\mathcal{L}(A, xy) = \mathtt{fail}}_{G[A] \ xy} \ \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}$$

$$(A, xyz)
otin \mathcal{L} \qquad G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} (yz, x')$$
 $G[P(A)] \quad xyz \quad \mathcal{L}[(A, xyz) \mapsto (yz, x')] \stackrel{\mathrm{INC}}{\leadsto} (z, (xy)')$
 $G[A] \quad xyz \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} (z, A[(xy)'])$

$$\underbrace{(A, x) \notin \mathcal{L} \quad G[P(A)] \quad x \quad \mathcal{L}[(A, x) \mapsto \mathtt{fail}] \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}}_{G[A] \quad x \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}}$$

$$\underbrace{\mathcal{L}(A, xy) = \mathtt{fail}}_{G[A] \quad xy \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} \mathtt{fail}}$$

$$\underbrace{\mathcal{L}(A, xy) = (y, x')}_{G[A] \quad xy \quad \mathcal{L} \stackrel{\mathrm{PEG}}{\leadsto} (y, A[x'])}$$

Семантика отношения INC

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (yzw, x')] \stackrel{\text{PEG}}{\leadsto} (zw, (xy)')$$

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (zw, (xy)')] \stackrel{\text{INC}}{\leadsto} (w, (xyz)')$$

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (yzw, x')] \stackrel{\text{INC}}{\leadsto} (w, (xyz)')$$
, where $y \neq \varepsilon$

Семантика отношения INC

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (yzw, x')] \stackrel{\text{PEG}}{\leadsto} (zw, (xy)')$$

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (zw, (xy)')] \stackrel{\text{INC}}{\leadsto} (w, (xyz)')$$

$$G[P(A)] \ xyzw \ \mathcal{L}[(A, xyzw) \mapsto (yzw, x')] \stackrel{\text{INC}}{\leadsto} (w, (xyz)')$$
, where $y \neq \varepsilon$

Семантика отношения INC

$$\begin{split} G[P(A)] & \ xyzw \ \mathcal{L}[(A,xyzw) \mapsto (yzw,x')] \overset{\text{PEG}}{\leadsto} (zw,(xy)') \\ \frac{G[P(A)] & \ xyzw \ \mathcal{L}[(A,xyzw) \mapsto (zw,(xy)')] \overset{\text{INC}}{\leadsto} (w,(xyz)')}{G[P(A)] & \ xyzw \ \mathcal{L}[(A,xyzw) \mapsto (yzw,x')] \overset{\text{INC}}{\leadsto} (w,(xyz)')}, \text{where } y \neq \varepsilon \\ \frac{G[P(A)] & \ x \ \mathcal{L} \overset{\text{PEG}}{\leadsto} \text{fail}}{G[P(A)] & \ x \ \mathcal{L} \overset{\text{PEG}}{\leadsto} \mathcal{L}(A,x)} \\ \underline{G[P(A)] & \ xyz \ \mathcal{L}[(A,xyz) \mapsto (z,(xy)')] \overset{\text{PEG}}{\leadsto} (yz,x')} \end{split}$$

 $G[P(A)] \ xuz \ \mathcal{L}[(A, xyz) \mapsto (z, (xy)')] \stackrel{\text{INC}}{\sim} (z, (xy)')$