

2024

КЫК, || и их друзья

29 августа 2024 г.

Содержание

I	Чистка грамматики	1
1	Исключение непорождающих переменных	1
2	Исключение недостижимых символов	2
3	Полезные и бесполезные символы и продукции	2
3.1	Избавляемся от эпсилон-продукций	3
3.2	Избавляемся от единичных продукций	3
4	Алгоритм очистки грамматики!!!	3
II	Нормальная форма Хомского	4
5	Алгоритм построения CNF	4
III	КЯК парсер	5
IV	LL1 парсер	7
V	LR(0): то, ради чего мы сюда пришли	10

Часть I

Чистка грамматики

Граматики могут быть "плохими": там могут быть лишние символы, лишние правила
Такие грамматики хочется упростить, и убрать из них всяческий мусор

1 Исключение непорождающих переменных

Примем следующие обозначения:

T - алфавит (набор терминалов языка)
 N - набор переменных (нетерминалов, символов)
 P - множество продукций (порождающих правил)
 S - стартовый символ
 $G = \langle T, N, P, S \rangle$ - грамматика

T^* - замыкание алфавита
 L - язык

w - терминал

$L(T^*)$ - множество всех возможных строк терминалов

N^G - множество всех возможных строк нетерминалов

$L(T \cup N^G)$ - множество всех возможных строк из терминалов и нетерминалов

Переменная порождает строчку нетерминалов $A \in N^G$, если есть продукции вида:

1. $A \rightarrow w$, где $w \in L(T^*)$
2. $A \rightarrow \alpha$, где $\alpha \in L(T \cup N^G)^+$

И на примере это выглядит так:

$T = \{a, b, c\}$

$N = \{A, B, C, S\}$

$P = \{S \rightarrow AB, S \rightarrow C, A \rightarrow aA, A \rightarrow a, B \rightarrow bB, C \rightarrow c\}$

Действуем итерационно:

1. Ищем правила первого типа: $A \rightarrow a, C \rightarrow c$
То есть C, A - порождающие
2. Теперь, когда базис собран - ищем правила второго типа, содержащие нетерминалы и порождающие символы:
 $S \rightarrow C, A \rightarrow aA$, - то есть S - порождающий
3. Итого, в наши порождающие символы в итоге попали $\{C, A, S\}$

A вот B ничего в итоге не порождает, а значит возможно он нам и не нужен?

2 Исключение недостижимых символов

Попробуем пройти в другую сторону - от стартового символа

1. S - достигим
2. $A \rightarrow \alpha$, если A достигим, то все символы из α достигимы

В общем-то достаточно логичный набор правил

Рассмотрим что получится на примере:

$S \rightarrow AB, A \rightarrow C, C \rightarrow c, B \rightarrow bB, D \rightarrow aC, D \rightarrow bc$

S достигим - значит A, B достигимы

A достигим - значит C достигим

B достигим - значит b достигим

C достигим - значит c достигим

И остались недостижимыми D, a - может они нам тоже не нужны?

3 Полезные и бесполезные символы и продукции

Собственно символ - **полезный** - если он используется в порождении какой-либо строки терминалов из стартового символа грамматики (ну в общем достижимый и порождающий)

A ещё бывают бесполезные продукции, и с ними надо подробнее разобраться

$A \rightarrow \varepsilon$ - это **эпсилон-продукция**

$A \rightarrow B$ - это **единичная продукция**

И есть одна интересная теоремка - если L - контекстно-свободный язык, то $L - \{\varepsilon\}$ можно представить в виде контекстно-свободной грамматики без эпсилон-продукций

И есть из этой теоремки забавное следствие: Любой контекстно-свободный язык, содержащий пустую строку, можно представить контекстно-свободной грамматикой вида: $G = \langle T, N, P \cup \{S \rightarrow \varepsilon\}, S \rangle$, где P не содержит эпсилон-продукций

3.1 Избавляемся от эпсилон-продукций

Идём дальше - теперь разберёмся с такой штукой как **зануляемый символ**, оно нам пригодится
Логично - что это символ A такой что

1. $A \rightarrow \varepsilon$
2. $A \rightarrow \alpha$, где α - строка из зануляемых символов

A теперь... АЛГОРИТМ (хотя как завещала Бабалова - алгоритм был сначала)

1. Находим зануляемые символы
2. (могу сюда вставить честное определение, но....) Во всех продукциях добавляем всевозможные комбинации, где зануляемые символы явно занулены (убраны к чертовой матери)
3. Искключаем все эпсилон-продукции, кроме той, что со стартовым символом слева
4. Искключаем все бесполезные символы

ПРИМЕР: $S \rightarrow ABC$, $A \rightarrow aA \mid \varepsilon$, $C \rightarrow c \mid \varepsilon$, $B \rightarrow bB \mid A$

1. Видно что зануляемыми у нас в итоге являются: A, C, B, S
2. Вот тут начинается веселье

$S \rightarrow ABC$ - заменяем на $S \rightarrow ABC|BC|AC|AB|C|B|A|\varepsilon$
 $A \rightarrow aA$ - заменяем на $A \rightarrow aA|a|\varepsilon$
 $C \rightarrow c|\varepsilon$ - не не изменилось
 $B \rightarrow bB|A$ - заменяем на $B \rightarrow bB|A|b|\varepsilon$

3. В итоге остаются только:

$S \rightarrow ABC|BC|AC|AB|C|B|A|\varepsilon$
 $A \rightarrow aA|a$
 $C \rightarrow c$
 $B \rightarrow bB|A|b$

3.2 Избавляемся от единичных продукций

Теперь нам бы ещё надо избавиться от единичных продукций

1. Продукции вида $A \rightarrow^* B$ и $B \rightarrow \alpha$ заменяем на $A \rightarrow \alpha$ - по сути просто сокращаем множество транзитивных переходов до одного
2. (Искключаем все бесполезные символы)

ПРИМЕРЧИК:

$S \rightarrow ABE$, $A \rightarrow aA|a$, $C \rightarrow c$, $B \rightarrow A|D$, $D \rightarrow C$, $E \rightarrow B|C|e$

И теперь заменяем все эти единичные переходы...

$S \rightarrow ABE$, $A \rightarrow aA|a$, $C \rightarrow c$, $B \rightarrow aA|a|c$, $D \rightarrow c$, $E \rightarrow aA|a|c|e$

Ну и теперь, в принципе можно убрать лишние символы: C, D

$S \rightarrow ABE$, $A \rightarrow aA|a$, $B \rightarrow aA|a|c$, $E \rightarrow aA|a|c|e$

4 Алгоритм очистки грамматики!!!

1. Искключаем эпсилон-продукции
2. Искключаем единичные продукции
3. Искключаем все бесполезные символы
4. Ну и если стартовый символ зануляемый - оставляем $S \rightarrow \varepsilon$

Часть II

Нормальная форма Хомского

Грамматика является грамматикой в нормальной форме Хомского, если включает только продукции вида:

1. $A \rightarrow BC$, где $A, B, C \in N$
2. $A \rightarrow a$, где $A \in N, a \in T$

И это очень строгие ограничения - и зачем оно нам вообще надо?
Ну в общем-то чтобы применить наш простой парсер КЯК

5 Алгоритм построения CNF

1. Очистка грамматики - и теперь грамматике только продукции из 1 терминала или длина тела не меньше 2х
2. Для каждого терминала создаём продукцию $A_a \rightarrow a$ и заменяем a на A_a во всех продукциях с длиной ≥ 2 - и теперь у нас есть только продукции из 1 терминала или только из нетерминалов, длины ≥ 2
3. Ну и теперь заменяем все продукции с длиной ≥ 2 на грамматики по такому алгоритму
 $A \rightarrow B\alpha$, где $|\alpha| \geq 2$ превращаем в $A \rightarrow BC$ и $C \rightarrow \alpha$

А теперь - пример всего этого дела.

$$\begin{aligned} E &\rightarrow E + T | T \\ T &\rightarrow T * F | F \\ F &\rightarrow id | (E) \end{aligned}$$

Как тут видно, это ни разу не нормальная форма Хомского. Так что идём по нашему алгоритму.

Для начала - очистим грамматику

Получаем уже вот так

$$\begin{aligned} E &\rightarrow E + T | T * F | id | (E) \\ T &\rightarrow T * F | id | (E) \\ F &\rightarrow id | (E) \end{aligned}$$

А теперь начинается колдовство!

Для наших терминалов $(,), +, *, id$ создаём нетерминалы, и заменяем на них в продукциях, с длиной больше 2

$$\begin{aligned} E &\rightarrow EPT | TMF | id | LER \\ T &\rightarrow TMF | id | LER \\ F &\rightarrow id | LER \\ L &\rightarrow (\\ R &\rightarrow) \\ P &\rightarrow + \\ M &\rightarrow * \\ I &\rightarrow id \end{aligned}$$

Выглядит это конечно жутковато (но можем сразу выкинуть I , потому что он нигде не нужен). Но вот теперь следуем второму шагу - заменяем всё так, чтобы получились нормальные продукции.

$$\begin{aligned} E &\rightarrow EA | TB | id | LC \\ T &\rightarrow TB | id | LC \\ F &\rightarrow id | LC \\ L &\rightarrow (\\ R &\rightarrow) \\ P &\rightarrow + \\ M &\rightarrow * \\ A &\rightarrow PT \\ B &\rightarrow MF \\ C &\rightarrow ER \end{aligned}$$

Вуаля, наша грамматика в нормальной форме Хомского.

Часть III

КЯК парсер

Рассмотрим пример грамматики ниже. Возьмём для разбора строку *baaba*.

$$S \rightarrow AB \mid BC$$

$$A \rightarrow BA \mid a$$

$$B \rightarrow CC \mid b$$

$$C \rightarrow AB \mid a$$

Суть КЯК в рассмотрении подстрок от самых маленьких к целой строке. По итогу нам нужно узнать, является ли строка каким-то из нетерминалов (т.е. словом языка, порождённого грамматикой). Если совсем кратко - принадлежит ли строка языку.

Приступим к разбору. Ещё раз напомним, что мы его ведём снизу вверх - от терминалов к нетерминалам и так пока не охватим всю строку.

$$B \rightarrow b$$

$$A \rightarrow a$$

$$C \rightarrow a$$

Идём дальше. Это были подстроки длины 1. *baaba* также имеет подстроки длины 2, которые формируются из подстрок длины 1 (\times - декартово произведение):

$$ba = b \times a = B \times A = BA = A$$

$$ba = b \times a = B \times C = BC = S$$

Пока у нас есть $A \rightarrow ba, S \rightarrow ba$. Которые получаются, как мы выяснили, вот так:

$$A \rightarrow BA \rightarrow bA \rightarrow ba$$

$$S \rightarrow BC \rightarrow bC \rightarrow ba$$

Тупо последовательно заменяем нетерминалы, пока не доберёмся до терминалов. К сожалению, дальше мы упоремся и запутаемся, если будем так продолжать. Это была всего лишь первая 2-подстрока, а я уже устала.

Видимо, кто-то из авторов тоже устал, и придумал лестницу КЯК или как она там. Выглядит она следующим образом (цифры - длины подстрок, НТ - нетерминалы):

5	НТ				
4	НТ	НТ			
3	НТ	НТ	НТ		
2	НТ	НТ	НТ	НТ	
1	НТ	НТ	НТ	НТ	НТ
	b	a	a	b	a

Эта таблица показывает, из каких нетерминалов какую строку мы можем построить. Пока что для нас она выглядит так:

5	НТ				
4	НТ	НТ			
3	НТ	НТ	НТ		
2	НТ	НТ	НТ	НТ	
1	В	А,С	А,С	В	А,С
	b	a	a	b	a

И, учитывая что мы уже успели проанализировать первое *ba*:

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	S,A	HT	HT	HT	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Теперь давайте заполнять эту таблицу. Для следующих подстрок мы получим:

$$aa = a \times a = A, C \times A, C = AA, CA, AC, CC = B$$

$$ab = a \times b = A, C \times B = AB, CB = S, C$$

$$bb = b \times b = B \times B = \emptyset$$

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 3.

$$baa = b \times aa = B \times B = BB = \emptyset$$

$$baa = ba \times a = S, A \times A, C = SA, AA, SC, AC = \emptyset$$

$$aab = a \times ab = A, C \times S, C = AS, CS, AC, CC = B$$

$$aab = aa \times b = B \times B = \emptyset$$

$$aba = a \times ba = A, C \times S, A = AS, CS, AA, CA = \emptyset$$

$$aba = ab \times a = S, C \times A, C = SA, CA, SC, CC = B$$

5	HT				
4	HT	HT			
3	\emptyset	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 4.

$$baab = b \times aab = B \times B = BB = \emptyset$$

$$baab = ba \times ab = S, A \times S, C = SS, AS, SC, AC = \emptyset$$

$$baab = baa \times b = \emptyset \times B = \emptyset$$

$$aaba = a \times aba = A, C \times B = AB, CB = S, C$$

$$aaba = aa \times ba = B \times S, A = BS, BA = A$$

$$aaba = aab \times a = B \times A, C = BA, BC = A, S$$

5	HT				
4	\emptyset	S,A,C			
3	\emptyset	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 5.

$$baaba = b \times aaba = B \times S, A, C = BS, BA, BC = A, S$$

$$baaba = ba \times aba = S, A \times B = SA, AB = S, C$$

$$baaba = baa \times ba = \emptyset \times S, A = \emptyset$$

$$baaba = baab \times a = \emptyset \times A, C = \emptyset$$

5	S,A,C				
4	∅	S,A,C			
3	∅	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Таким образом, строка может быть получена путем последовательного раскрытия (развертки) нетерминалов S, A, C.

УПРАЖНЕНИЕ. Убедиться, что все указанные нетерминалы преобразуются в искомую строку.

Часть IV

LL1 парсер

Всё начинается с таблицы парсинга. Для составления таблицы необходимо вначале посчитать множества символов FIRST и FOLLOW для каждого нетерминала NT.

Заметим, что первое правило грамматики для стартового нетерминала, если не указано иное. После стартового нетерминала может идти конец строки, для остальных в общем случае это неверно. Рассмотрим следующую грамматику:

$$A \rightarrow CB$$

$$B \rightarrow +CB \mid \varepsilon$$

$$C \rightarrow ED$$

$$D \rightarrow *ED \mid \varepsilon$$

$$E \rightarrow id \mid (A)$$

FIRST(NT) - множество символов на первой позиции строки NT (aka begin()):

$$FIRST(A) = FIRST(C)$$

$$FIRST(B) = '+' \cup \varepsilon$$

$$FIRST(C) = FIRST(E)$$

$$FIRST(D) = '*' \cup \varepsilon$$

$$FIRST(E) = 'id' \cup '('$$

FOLLOW(NT) - множество символов на первой после последней позиции строки NT (aka end()) (т.е.

1) если видим $P \rightarrow \dots NT Q$, то добавляем $FIRST(Q)$ (если Q непусто) или $FOLLOW(P)$ (если Q пусто)

2) если Q может быть ε , то добавляем $FOLLOW(Q)$

3) ε не может быть в FOLLOW, если появляется, то его надо убрать)

$$FOLLOW(A) = ')' \cup '$'$$

$$FOLLOW(B) = FOLLOW(A)$$

$$FOLLOW(C) = FIRST(B) \cup FOLLOW(A) \cup FOLLOW(B) \setminus \varepsilon$$

$$FOLLOW(D) = FOLLOW(C) \setminus \varepsilon$$

$$FOLLOW(E) = FIRST(D) \cup FOLLOW(C) \cup FOLLOW(D) \setminus \varepsilon$$

Получим:

NT	FIRST	FOLLOW
A	id, (), \$
B	+, ε), \$
C	id, (+,), \$
D	*, ε	+,), \$
E	id, (*, +,), \$

Теперь, таблица парсинга. Она содержит производящие правила на пересечении терминалов и нетерминалов. Заполняется она так для каждого нетерминала NT:

1. Если $\varepsilon \in FIRST(NT)$, то пишем правило, которым получился ε , во всех строках из FOLLOW(NT).
2. Для всех прочих символов (не $\$$) из FIRST(NT) в соответствующих столбцах пишется правило, которым этот символ был получен.

NT/T	id	(+)	*	\$
A	$A \rightarrow CB$	$A \rightarrow CB$				
B			$B \rightarrow +CB$	$B \rightarrow \varepsilon$		$B \rightarrow \varepsilon$
C	$C \rightarrow ED$	$C \rightarrow ED$				
D			$D \rightarrow \varepsilon$	$D \rightarrow \varepsilon$	$D \rightarrow *ED$	$D \rightarrow \varepsilon$
E	$E \rightarrow id$	$E \rightarrow (A)$				

Удивительно, но алгоритм предельно прост. Мы всего лишь инициализируем стек стартовым нетерминалом, а в конец рассматриваемой строки добавляем конечный символ $\$$.

Далее происходит следующее. Мы достаём символ из строки и нечто со стека. Если нечто - терминал и совпало с символом, то *por* и двигаем текущий символ. Если не совпало - ошибка. (При любой ошибке выходим из цикла.)

Если же нечто - нетерминал, то мы ищем в таблице запись на пересечении нечто и текущего символа строки. Если запись не найдена, то ошибка. Иначе *por*. Правая часть записи добавляется в стек, так что первый её элемент наверху.

Так повторяется, пока текущий символ не станет концом строки или не будет получена ошибка. Если ошибок не было, то строка принадлежит языку.

Рассмотрим пример. Грамматика

$$\begin{aligned} S' &\rightarrow S\$ \\ S &\rightarrow xYzS \mid a \\ Y &\rightarrow xYz \mid y \end{aligned}$$

и строки $xyyza$ и $xyyzz$.

Таблица 1:

NT	FIRST	FOLLOW
S'	x,a	\$
S	x,a	\$
Y	x,y	z,\$

Таблица 2:

NT/T	a	x	y	z	\$
S'	$S' \rightarrow S\$$	$S' \rightarrow S\$$			
S	$S \rightarrow a$	$S \rightarrow xYzS$			
Y		$Y \rightarrow xYz$	$Y \rightarrow y$		

Начинаем разбор строки $xyyza$:

Стек: S'

Строка: $xyyza\$$

Символ: x

Нечто - S'

Нечто - нетерминал, и потому ищем пересечение S' и x. Находим $S' \rightarrow S\$$.

Стек: $\$S$

Строка: $xyyza\$$

Символ: x

Нечто - S

Нечто - нетерминал, и потому ищем пересечение S и x. Находим $S \rightarrow xYzS$.

Стек: $\$SzYx$

Строка: $xyyza\$$

Символ: x

Нечто - x

Нечто - терминал, и он совпал с текущим символом.

Стек: $SSzY$

Строка: $xyzza\$$

Символ: x

Нечто - Y

Нечто - нетерминал, и потому ищем пересечение Y и x . Находим $Y \rightarrow xYz$.

Стек: $SSzzYx$

Строка: $xyzza\$$

Символ: x

Нечто - x

Нечто - терминал, и он совпал с текущим символом.

Стек: $SSzzY$

Строка: $yzza\$$

Символ: y

Нечто - Y

Нечто - нетерминал, и потому ищем пересечение Y и y . Находим $Y \rightarrow y$.

Стек: $SSzzzy$

Строка: $yzza\$$

Символ: y

Нечто - y

Нечто - терминал, и он совпал с текущим символом.

Стек: $SSzz$

Строка: $zza\$$

Символ: z

Нечто - z

Нечто - терминал, и он совпал с текущим символом.

Стек: SSz

Строка: $za\$$

Символ: z

Нечто - z

Нечто - терминал, и он совпал с текущим символом.

Стек: SS

Строка: $a\$$

Символ: a

Нечто - S

Нечто - нетерминал, и потому ищем пересечение S и a . Находим $S \rightarrow a$.

Стек: Sa

Строка: $a\$$

Символ: a

Нечто - a

Нечто - терминал, и он совпал с текущим символом.

Стек: S

Строка: S

Символ: S

Нечто - S

Нечто - S , и он совпал с текущим символом.

Стек:

Строка:

Стек опустел, значит цикл закончился. Ошибки не было, посему строка $xyzza$ принадлежит языку.

Теперь строка $xyzzz$. Здесь всё то же самое до момента:

Стек: $\$Sz$
Строка: $zz\$$

Символ: z
Нечто - z
Нечто - терминал, и он совпал с текущим символом.
Стек: $\$S$
Строка: $z\$$

Символ: z
Нечто - S
Нечто - нетерминал, и потому ищем пересечение S и z . Не нашли. ОШИБКА. Выходим.
Таким образом, строка $xyzzz$ не принадлежит языку.

Часть V

LR(0): то, ради чего мы сюда пришли

Мы преодолели долгий путь, но тут на нас напали конечные автоматы. Присаживайтесь поудобнее, сейчас начнётся реальная мясорубка.

LR-челики строятся на двух основных операциях: сдвиг и свертка:

Сдвиг:
Стек: aBC , Строка: def
Стек: $aBCd$, Строка: ef

Свертка (предполагая правило $A \rightarrow BC$):
Стек: aBC
Стек: aA

Вся сложность парсинга состоит в том, что произвольную строку не спарсить, тыкая наугад либо сдвиг либо свертку. Последовательностей действий может быть разной, и какая-то будет успешной, а какая-то нет (это как игра в карты, непонятно когда какую карту выложить, а когда придержать). Вообще, ТА - то ещё казино. Особенно когда грамматика некорректная. Ну да ладно.

Короче, какой-то больной на голову человек придумал анализировать сдвиги через ввод дополнительного символа. В разных источниках его обозначают по-разному, мы будем использовать $_$. Производящее правило с $_$ в правой части правила называется LR(0)-ситуацией.

Рассмотрим грамматику

$$\begin{aligned} S &\rightarrow TF \\ F &\rightarrow +T \\ T &\rightarrow a \mid (F) \end{aligned}$$

Поехали строить LR(0) состояния. Изначально вспомогательный символ всегда находится в начале правой части производящего правила.

$$S \rightarrow _TF$$

Далее рекурсивно повторяем это для нетерминала перед $_$, в данном случае T .

$$T \rightarrow _a, T \rightarrow _(F)$$

Итак, 1. $S \rightarrow _TF, T \rightarrow _a, T \rightarrow _(F)$.

Теперь двигаем $_$ во всех LR(0)-ситуациях. Сдвиг относительно каждого отдельного терминала или нетерминала порождает новое состояние.

$$\begin{aligned} 2. S &\rightarrow T_F \\ 3. T &\rightarrow a_ \\ 4. T &\rightarrow (_F) \end{aligned}$$

2е и 4е состояния порождает ситуацию $_NT$, так что опять плодят в себе другие состояния:

$$F \rightarrow _ + T$$

В итоге

$$2. S \rightarrow T_F, F \rightarrow _ + T$$

$$3. T \rightarrow a_$$

$$4. T \rightarrow (_F), F \rightarrow _ + T$$

Далее из 2

$$5. S \rightarrow TF_$$

$$6. F \rightarrow +_T, T \rightarrow _a, T \rightarrow _(F)$$

Из 4

$$7. T \rightarrow (F_)$$

Из 6

$$8. F \rightarrow +T_$$

Из 7

$$9. T \rightarrow (F)_$$

Соберём состояния воедино:

$$1. T \rightarrow _a, T \rightarrow _(F), S \rightarrow _TF$$

$$2. F \rightarrow _ + T, S \rightarrow T_F$$

$$3. T \rightarrow a_$$

$$4. T \rightarrow (_F), F \rightarrow _ + T$$

$$5. S \rightarrow TF_$$

$$6. F \rightarrow +_T, T \rightarrow _a, T \rightarrow _(F)$$

$$7. T \rightarrow (F_)$$

$$8. F \rightarrow +T_$$

$$9. T \rightarrow (F)_$$

Сост/Элм	а	+	()	конец	F	T
1	сдвиг 3		сдвиг 4				сдвиг 2
2		сдвиг 6				сдвиг 5	
3 свертка $T \rightarrow a$							
4						сдвиг 7	
5					свертка $S \rightarrow TF$, успех		
6							сдвиг 8
7						сдвиг 9	
8 свертка $F \rightarrow +T$							
9 свертка $T \rightarrow (F)$							

Таким образом, таблица парсинга LR(0) построена.