Введение в Теоретическую Лингвистику

Филипп Дудчук
Саша Подобряев

E-mail: seminars@linguistics.msx.ru
http://seminars.narod.ru/spring2005

Лекция 10
29.04.2005

Грамматики и автоматы

0 Чтение

Пентус, А. Е. и М. Р. Пентус. 2004. *Теория формальных языков.* М.: Изд-во ЦПИ при механико-математическом факультете МГУ. С. 3–12.

Книга доступна в формате PDF (532 Kb) на http://lpcs.math.msu.su/~pentus/ftp/papers/tfyav11.pdf (ссылка есть на домашней странице семинара).

1 Исходные понятия

Натуральные числа: $\mathbb{N} = \{0, 1, 2, \ldots\}$.

Алфавит — конечное непустое множество, элементы которого называются **символы** (**буквы**) алфавита: $\Sigma = \{a, b, c, \ldots\}$. **Слово** в алфавите Σ — конечная последовательность элементов из Σ. Например, *aabba* является словом в алфавитах $\Sigma_1 = \{a, b\}$, $\Sigma_2 = \{a, b, c, e\}$, но не является словом в алфавите $\Sigma_3 = \{a, c, d, f\}$. **Пустое слово** — слово, не содержащее ни одного символа. Пустое слово будем обозначать ε. Множество всех слов в алфавите Σ обозначается Σ^*. Условимся также, что $\Sigma^+ = \Sigma^* - \{\varepsilon\}$. **Языком** в алфавите Σ будем называть множество слов $L \subseteq \Sigma^*$.

(1) $aabba \in \{a, b\}^*$

(2) $aabba \notin \{c, d\}^*$

Конкатенация слов x и y ($x \cdot y$) — слово xy, получающееся приписыванием y справа от x. Нетрудно заметить факт (3).

(3) $\forall x \forall y \forall \Sigma\, [x \in \Sigma^* \wedge y \in \Sigma^* \leftrightarrow xy \in \Sigma^*]$

Если $x \in \Sigma^*$, $n \in \mathbb{N}$, то $x^n \in \Sigma^*$ и x^n — слово, состоящее из x, записанного n раз подряд.

(4) $(ab)^3 = ababab$

(5) $ab^3 = abbb$

(6) $x^0 = \varepsilon$

2 Порождающие грамматики

Кроме алфавита $\Sigma = \{a, b, c, \ldots\}$, введем в рассмотрение еще один алфавит $N = \{A, B, C, \ldots\}$. Символы из Σ будем называть **основными (терминальными)**, а символы из N — **вспомогательными (нетерминальными)**. Соответственно, Σ — основной алфавит, а N — вспомогательный. Кроме того, зафиксируем некоторый символ S из алфавита N и назовем его **начальным**.

Тогда кортеж $<N, \Sigma, P, S>$ будем называть порождающей грамматикой G при выполнении следующих условий.

(i) $N \cap \Sigma = \varnothing$

(ii) P конечно и $P \subset (N \cup \Sigma)^+ \times (N \cup \Sigma)^*$

Элементы множества P — пары $<\alpha, \beta>$, каждая из которых является правилом подстановки: $\alpha \to \beta$.

Язык, порождаемый грамматикой G над алфавитом Σ, есть множество слов алфавита Σ, каждое из которых можно получить за конечное число шагов из начального символа грамматики G, т.е. применив правила подстановки грамматики G конечное число раз.

2.1 Конкретный пример 1

$G_1 = <\{S, A\}, \{a\}, \{S \to aAa, A \to \varepsilon\}, S>$. Более упрощенная запись:

G_1: $S \to aAa$
$A \to \varepsilon$

Язык, порождаемый G, состоит ровно из одного слова: $L(G_1) = \{aa\}$. Для этого языка можно предложить более простую грамматику G_2:

G_2: $S \to aa$

Грамматики G_1 и G_2 порождают один и тот же язык. Такие грамматики будем называть **эквивалентными**. «Обогатим» грамматику G_1.

$G_3 = <\{S, A\}, \{a\}, \{S \to aAa, A \to \varepsilon, A \to aAa\}, S>$

G_3: $S \to aAa$
$A \to \varepsilon$
$A \to aAa$

Нетрудно заметить, что $L(G_3) = \{aa, aaaa, aaaaaa, \ldots\} = \{(aa)^n \mid n \geq 1\} = \{a^{2n} \mid n \geq 1\}$. Если из G_3 удалить правило $A \to \varepsilon$, окажется, что $L(G_3) = \varnothing$.

2.2 Конкретный пример 2

Существует грамматика, порождающая язык, эквивалентный множеству всех слов в алфавите. Для примера зафиксируем алфавит $\Sigma = \{a, b, c\}$

G_4: $S \to \varepsilon$
$S \to aS$
$S \to bS$
$S \to cS$

В общем случае язык $L = \Sigma = \{a_1, \ldots a_n\}^*$ порождается грамматикой
$G_g = <\{S\}, \{a_1, \ldots a_n\}, \{S \to a_1S, S \to a_2S, S \to a_3S, ..., S \to a_nS\}, S>$.

2.3 Классы грамматик. Иерархия Хомского

Контекстная (контекстно-зависимая) грамматика (= грамматика типа 1) — порождающая грамматика, каждое правило которой имеет вид $\eta A \theta \to \eta \alpha \theta$, где $A \in N$, $\eta \in (N \cup \Sigma)^*$, $\theta \in (N \cup \Sigma)^*$ и $\alpha \in (N \cup \Sigma)^+$.

Например, грамматика G_5 не является контекстной (последние три правила не имеют требуемый вид).

G_5: $S \to TS$
$S \to US$
$S \to b$
$Tb \to Ab$
$A \to a$
$TA \to AAT$
$UAb \to b$
$UAAA \to AAU$

Говоря неформально, последние три правила требуют стирание контекста, что в контекстных грамматиках запрещено.

Контекстно-свободная (КС-)грамматика (= грамматика типа 2) — порождающая грамматика, в которой все правила имеют вид $A \to \alpha$, где $A \in N$, $\alpha \in (N \cup \Sigma)^*$.

Линейная грамматика — порождающая грамматика, каждое правило которой имеет вид $A \to u$ или $A \to uBv$, где $A \in N$, $B \in N$, $u \in \Sigma^*$, $v \in \Sigma^*$.

Например, грамматика G_6 является КС-грамматикой, но не является линейной грамматикой (первые два правила не имеют требуемого вида).

G_6: $S \to TT$
$T \to cTT$
$T \to bT$
$T \to a$

Праволинейная грамматика (= грамматика типа 3) — порождающая грамматика, каждое правило которой имеет вид $A \to u$ или $A \to uB$, где $A \in N$, $B \in N$, $u \in \Sigma^*$.

Например, грамматика G_7 является линейной, но не праволинейной (первое правило не имеет требуемого вида).

G_7: $S \to aSa$
$S \to T$
$T \to bT$
$T \to \varepsilon$

По приведенным примерам нетрудно догадаться, что, например, всякая праволинейная грамматика есть частный случай линейных, КС-, контекстных и порождающих грамматик. Действительно, класс праволинейный грамматик входит в класс линейных грамматик, который входит в класс КС-грамматик, который входит в класс контекстных грамматик, который, в свою очередь, входит в класс порождающих грамматик.

Языки, порождаемые праволинейными грамматиками, являются **автоматными**, т. е. такими, для которых можно построить конечный автомат.

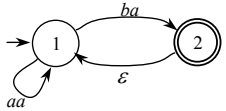
3 (Недетерминированные) конечные автоматы

Построим конечный автомат M. Зафиксируем алфавит Σ и четыре конечных множества Q, Δ , I и F, такие что:

(i) Q — множество **состояний** автомата M
(ii) $I \subseteq Q$ — множество **начальных состояний** автомата M
(iii) $F \subseteq Q$ — множество **допускающих состояний** автомата M
(iv) $\Delta = Q \times \Sigma^* \times Q$ — множество (допустимых) **переходов** автомата M

Тогда $M = \langle Q, \Sigma, \Delta, I, F \rangle$ есть искомый конечный автомат.

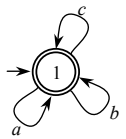
Например, $M = \langle \{1, 2\}, \{a, b\}, \{\langle 1, aa, 1\rangle, \langle 1, ba, 2\rangle, \langle 2, \varepsilon, 1\rangle\}, \{1\}, \{2\}\rangle$ является конечным автоматом и может быть изображен графически так:



Нетрудно заметить, какой праволинейной грамматикой порождается язык, допускаемый данным автоматом.

G_M: $S \to aaS$
$S \to baT$
$T \to S$
$T \to \varepsilon$

Обратная задача: построим автомат, например, для грамматики G_4.

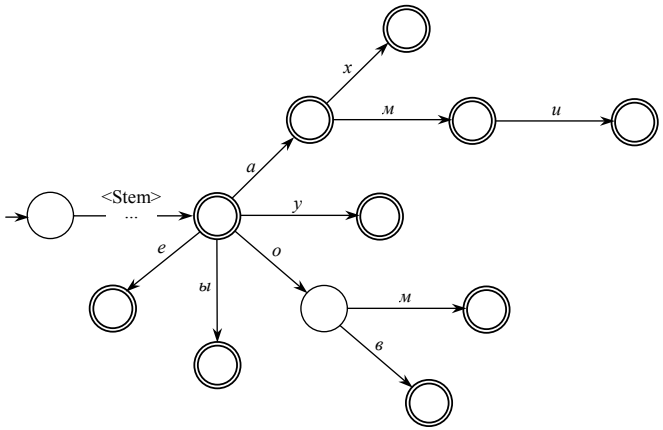


Конкретный пример: русская именная морфология

Построим автомат, распознающий фрагмент русского языка — склонение существительных типа *стол*.

стол	стол *ы*
стол *а*	стол *о в*
стол *у*	стол *а м*
стол *о м*	стол *а м и*
стол *е*	стол *а х*

Обозначим основу склоняемого существительного переменной <Stem>. Ее область значений — основы морфологического класса, в который входит существительное *стол* (например, *мат*, *класс*, *семинар*, ...). Зафиксируем алфавит $\Sigma = \{ а, в, е, и, м, о, у, х, ы \}$.



Вопрос для домашнего обдумывания: можно ли «упростить» полученный автомат? (Под упрощением автомата здесь понимается уменьшение количества состояний).