

Задание 8

Контекстно-свободные языки и магазинные автоматы II

Ключевые слова¹: язык, контекстно-свободный язык, магазинный автомат, грамматика, метод математической индукции.

1 КС-грамматики

1.1 Определение

Вспомним сначала что такое грамматика.

Определение 1. Грамматикой Γ называется четвёрка (N, T, P, S) , где

- N – множество нетерминальных символов
- T – множество терминальных символов
- P – множество правил вывода, $P \subseteq (N \cup T)^* \times (N \cup T)^*$.
- S – аксиома, $S \in N$.

При этом, $N \cap T = \emptyset$, нетерминалы обычно обозначаются заглавными буквами A, B, C, \dots , терминалы обычно обозначаются строчными буквами и/или цифрами, смешанные цепочки из $(N \cup T)^*$ обозначают греческими буквами α, β, γ . Слово $w \in T^*$ порождается грамматикой Γ , если существует последовательность правил вывода, начинающаяся с правила вида $S \rightarrow \alpha$, в результате применения которых порождается слово w . Под применением правила $\alpha \rightarrow \beta$, понимается, что подслово α заменяется на подслово β .

Грамматика называется *контекстно-свободной* (тип 2 по Хомскому), если все правила грамматики имеют вид $A \rightarrow \alpha$, где A – нетерминал, а α – цепочка, которая может состоять как из терминалов, так и нетерминалов.

При записи правил так же используются вспомогательное обозначение $A \rightarrow \alpha | \beta$, которое означает, что есть два правила: $A \rightarrow \alpha$ и $A \rightarrow \beta$.

¹минимальный необходимый объем понятий и навыков по этому разделу)

Пример 1. Грамматика G задана правилами:

$$S \rightarrow aAB$$

$$A \rightarrow aA|a$$

$$B \rightarrow bB|b$$

Слово $aabb$ выводится грамматикой. Последовательность применений правил вывода такая:

$$S$$

$$S \rightarrow aAB$$

$$aAB$$

$$A \rightarrow a$$

$$aaB$$

$$B \rightarrow bB$$

$$aabB$$

$$B \rightarrow b$$

$$aabb$$

1.2 Вывод, левый вывод, дерево разбора

Выводом цепочки α называется такая последовательность применений правил с указанием раскрываемого нетерминала, что применяя правила из неё начиная с аксиомы получается цепочка α . Если цепочка α не содержит нетерминалов, то α принадлежит языку, порождаемому КС-грамматикой. Нам будет удобно пользоваться такими понятиями как левый вывод и правый вывод. *Левым выводом* называют такой вывод, что на каждом его шаге раскрывается самый левый нетерминал в промежуточной цепочке. Вывод в примере 1 является левым. Правый вывод определяется аналогично.

Также мы будем использовать деревья вывода. С формальным определением дерева вывода вы можете познакомиться, например, в книге Хопкрофта, Мотвани и Ульмана, а мы рассмотрим пример деревьев вывода и затем дадим неформальное описание этого понятия.

Пример 2. Грамматика G , $T = \{a, +, *\}$ задана правилами:

$$S \rightarrow E$$

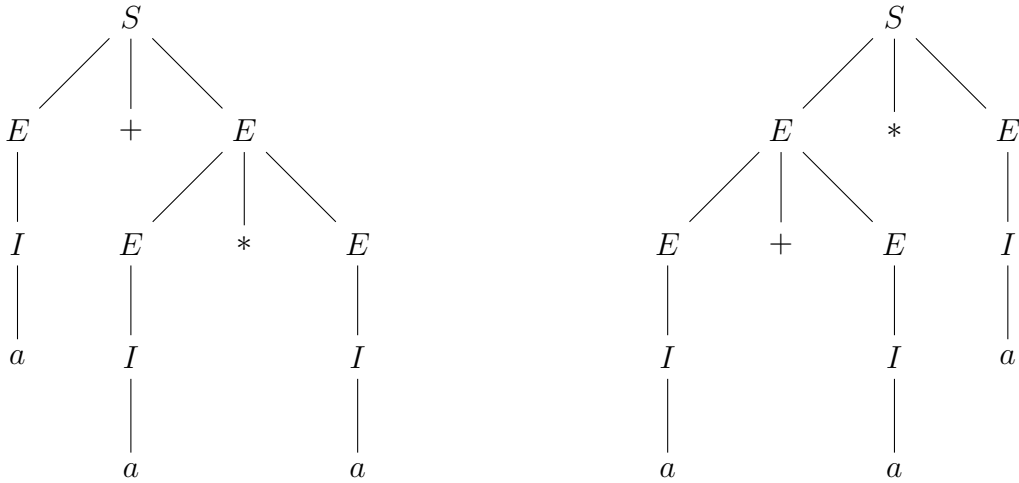
$$E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow I$$

$$I \rightarrow a$$

Построим деревья вывода для слова $a + a * a$:



Теперь опишем понятие дерева вывода, которое также называется деревом разбора. Зафиксируем грамматику G . *Деревом вывода* для слова w называется упорядоченное дерево, в корне которого находится аксиома S , каждая вершина помечена нетерминалом, терминалом или пустым словом, если вершина помечена терминалом или ε , то эта вершина является листом, если же вершина помечена нетерминалом A , то для некоторого правила $A \rightarrow X_1 X_2 \dots X_n \in P$ ($X_i \in N \cup T$) вершины-дети A помечены символами X_1, X_2, \dots, X_n слева направо. Листья дерева вывода образуют слово w .

Как мы видим, синтаксически деревья вывода принципиально разные: в первом случае выражение интерпретируется как $a + (a * a)$, а во втором как $(a + a) * a$, что приводит к непредсказуемому результату при выполнении стандартных операций!

Эта проблема приводит нас к новому важному понятию – неоднозначности. Грамматика G является *неоднозначной*, если хотя бы для одного слова существует два различных дерева вывода.

Левый и правый вывод помимо удобства важны тем, что они фактически задают порядок обхода дерева вывода, поэтому каждому левому выводу соответствует ровно одно дерево разбора, а каждому дереву разбора соответствует ровно один левый вывод.

Упражнение 1. Доказать, что грамматика G является однозначной тогда и только тогда, когда каждое слово, порожаемое G имеет ровно один левый(правый) вывод.

2 От КС-грамматик к МП-автоматам

Приведём алгоритм построения МП-автомата по описанию КС-грамматики и докажем его корректность.

Для каждой грамматики G введём вспомогательное отношение \Rightarrow на множестве $(N \cup T)^* \times (N \cup T)^*$. будем говорить, что $\alpha \Rightarrow \beta$, если существует правило, которое переводит цепочку α в цепочку β . Отношение \Rightarrow_L определим аналогично, только с условием, что цепочка β имеет левый вывод из α . Если цепочка β выводится из α за k шагов, будем обозначать это как $\alpha \Rightarrow^k \beta$. Будем также использовать транзитивное замыкание \Rightarrow^* отношения \Rightarrow . Неформально, $\alpha \Rightarrow^* \beta$, если для некоторого k выполняется $\alpha \Rightarrow^k \beta$.

Построение. По КС-грамматике $G = (N, T, P, S)$, построим N-автомат² $M = (T, N \cup T, q_0, q_0, S, \delta, \emptyset)$. Как видно из описания автомата, он содержит ровно одно состояние. Осталось описать только функцию переходов δ . Она устроена следующим образом:

- $A \rightarrow \alpha \in P \Rightarrow \delta(q_0, \varepsilon, A) \vdash (q_0, \alpha)$;
- $\forall \sigma \in T : \delta(q_0, \sigma, \sigma) \vdash (q_0, \varepsilon)$;

Корректность. Покажем, что $L(G) \subseteq L(M)$. Поясним неформально принцип работы автомата. В начале работы автомата, на дне стека лежит аксиома S . Если слово w выводится грамматикой G , то существует левый вывод данного слова. Автомат будет имитировать левый вывод

²МП-автомат, допускающий по пустому стеку

следующим образом. Если на вершшке стека находится нетерминал A , то необходимо продолжить левый вывод и применить очередное правило. Если же на вершшке стека находится последовательность терминалов, то она является префиксом необработанной части слова w , а значит считывая терминалы из стека и «сокращая» их с терминалами w автомат добирается до ближайшего левого нетерминала в выведенной цепочке и продолжает левый вывод или же добирается до дна стека и в случае совпадения содержимого стека с необработанной частью w , слово w будет принято, в противном случае, автомат аварийно завершит работу.

Теперь формально.

Утверждение 1. *Если на вершшке стека лежит нетерминал A , $A \Rightarrow_L^* u$, при этом автомат M находится в конфигурации $(q_0, uv, A\alpha)$, то $(q_0, uv, A\alpha) \vdash^* (q_0, v, \alpha)$.*

Доказательство. Проведём доказательство по индукции, параметром индукции будет длина k левого вывода \Rightarrow_L^k .

База: $k = 1$. Тогда существует правило $\delta(q_0, \varepsilon, A) \vdash (q_0, u)$, а далее необработанный префикс u сокращается со словом u на вершшке стека, следовательно $(q_0, uv, A\alpha) \vdash^* (q_0, v, \alpha)$.

Переход: пусть утверждение верно для $k = n$, покажем что оно верно для $n + 1$. Пусть первый вывод в цепочке левого вывода u из A будет $A \rightarrow u_1 B \beta$, $u_1 \in T^*$. Тогда $u = u_1 u_2$ и $(q_0, u_1 u_2 v, A\alpha) \vdash^* (q_0, u_2 v, B \beta \alpha)$ – если $u_1 \neq \varepsilon$, то необработанный префикс u_1 сокращается с u_1 в стеке. Но $B \beta \Rightarrow_L^k u_2$, а отсюда следует, что $u_2 = u_3 u_4$, и $B \Rightarrow_L^l u_3$, причём $l \leq n$, а значит $(q_0, u_2 v, B \beta \alpha) \vdash^* (q_0, u_3 v, \beta \alpha)$ по предположению индукции. Продолжая сокращать префикс u_i с терминалами в стеке и пользоваться тем фактом, что очередной префикс u_i выводится из нетерминала в цепочке β за не большее чем n число выводов, получим, что β удаляется из стека, то есть $(q_0, u_3 v, \beta \alpha) \vdash^* (q_0, v, \alpha)$. \square

В частности, из утверждения 1 следует, что если $S \Rightarrow_L w$, то $(q_0, w, S) \vdash^* (q_0, \varepsilon, \varepsilon)$, что и означает, что $w \in L(M)$.

Покажем теперь, что $L(M) \subseteq L(G)$. Каждой успешной³ последовательности конфигураций на входе w автомата M соответствует левый вывод w в грамматике G . А именно каждому переходу $\delta(q_0, \varepsilon, A) \vdash (q_0, \alpha)$ соответствует применению правила $A \rightarrow \alpha$ на шаге вывода.

³Под «успешной» мы понимаем, что на этой последовательности конфигураций слово w было принято автоматом.

Докажем это.

Упражнение 2. Доказать, что если $(q_0, u, \alpha) \vdash^* (q_0, \varepsilon, \varepsilon)$, то $\alpha \Rightarrow_L^* u$.

Отсюда следует, что если $w \in L(M)$, то $w \in L(G)$, т.к. первое влечёт $(q_0, w, S) \vdash^* (q_0, \varepsilon, \varepsilon)$.

3 Задачи

Задача 1. Даны языки $L_1 = \{a^n b^n c^m \mid n, m \geq 0\}$, $L_2 = \{f^n a^m b^m \mid n, m \geq 0\}$. Построить однозначную КС-грамматику, порождающую язык $L = L_1 \cup L_2$ и детерминированный МП-автомат, распознающий язык L . За однозначность грамматики и детерминированность автомата дополнительные очки. Если не получается их построить, стройте неоднозначную КСГ/недетерминированный МП-автомат.

Задача 2. Построить КС-грамматику или МП-автомат для языка $\{x y \mid x \neq y, x, y \in \{a, b\}^*\}$.

Расширенным МП-автоматом называется МП-автомат, который может извлекать из стека за такт работы не обязательно один, а возможно несколько символов, но при этом не более чем некоторое константное число c . Таким образом, в нём допустимы правила вида $\delta(q, a, \alpha) \vdash (p, \beta)$, где $\alpha < c$.

Задача 3. Доказать, что если M является расширенным МП-автоматом, то существует МП-автомат M' , такой что $L(M) = L(M')$. Докажите, что если M детерминированный автомат, то найдётся и детерминированный автомат M' .

На семинаре я вас обманул, сказав что МП-автомат с одним состоянием можно легко получить взяв за алфавит стека $Q \times \Gamma$. Для обычных МП-автоматов это неверно.

Задача 4[†]. Докажите, что для расширенного МП-автомата M есть расширенный МП-автомат M' с одним состоянием, алфавит стека которого есть $Q_M \times \Gamma_M$.

МП-автоматом с предпросмотром назовём МП-автомат, который может считывать с входной ленты не более, чем константное число символов. При этом, ему на вход подаётся слово $w\$$, где $\$$ – маркер конца слова, $w \in \Sigma^*$, $\$ \notin \Sigma$, алфавит автомата – $\Sigma \cup \{\$\}$. Таким образом, в нём

допустимы правила вида $\delta(q, u, x) \vdash (p, z)$, $\delta(q, v\$, x) \vdash (p, z)$ где $|u| < c$ и $|v\$| < c$.

Задача 5*. Доказать, что если M является МП-автоматом с предпросмотром, то существует МП-автомат M' , такой что $L(M) = L(M')$.