Задание 10

Преобразование Контекстно-Свободных языков

Ключевые слова [1]:*язык, контекстно-свободный язык, магазинный автомат, грамматика, морфизм, метод математической индукции.*

1 Теорема Хомского-Шютценберже

Обозначим D_n язык правильных скобочных выражений (язык Дика) с n типами скобок. Язык D_n определён над размеченым алфавитом $\Sigma = \Sigma_n \cup \bar{\Sigma}_n$ – в Σ_n входят открывающие скобки, в $\bar{\Sigma}_n$ закрывающие.

Будем говорить, что КС-язык $L \subseteq \Delta^*$ задан в представлении Хомского-Шютценберже, если определены язык Дика D_n, регулярный язык $R \subseteq \Sigma^*$ и морфизм $h : \Sigma^* \to \Delta^*$ и $L = h(R \cap D_n)$.

Теорема (Хомский, Шютценберже, 1963)**.** Язык $L \subseteq \Delta^*$ является контекстно-свободным тогда и только тогда, когда существуют такое n, регулярный язык R и морфизм $h : \Sigma^* \to \Delta^*$, что $h(D_n \cap R) = L$.

Доказательство. Пусть КС-язык задан грамматикой $G = (N, \Sigma, P, S)$, где N – множество нетереминалов, Σ – алфавит, P – правила вывода, а S – аксиома. Без ограничения общности будем считать, что G не содержит ε-правил, кроме быть может $S \to \varepsilon$, причём тогда нетерминал S больше не входит в правые части правил.

Зафиксируем язык D_n, где $n = |N| + |\Sigma|$. Поставим в соответствие каждому элементу X из $N \cup \Sigma$ пару скобок $[_X$ и $_X]$. Неформально опишем язык R, описав все регулярные события[2], которые допустимы в R. Сначала опишем вспомогательную конструкцию. Для каждого нетерминала A, определим множество R_A, в которое, для каждого правило $A \to X_1X_2 \ldots X_n$ грамматики G, входит соответствующее слово $w = [_{X_n}[_{X_{n-1}} \ldots [_{X_{n-i}} \ldots [_{X_1}$.

$$R_A = \{w = [_{X_n}[_{X_{n-1}} \ldots [_{X_{n-i}} \ldots [_{X_1} \mid A \to X_1X_2 \ldots X_n \in P\}$$

[1] минимальный необходимый объем понятий и навыков по этому разделу)

[2] Свойства слов, проверяемые ДКА.

Теперь опишем регулярные события, задающие R.

- Любое слово из R начинается со скобки $[_S$;
- После закрывающей скобки $]_A$ идёт слово из R_A, где $A \in N$;
- После открывающей скобки $[_\sigma$ может идти закрывающая скобка $_\sigma]$;
- После закрывающей скобки $]_X$ может идти закрывающая скобка $_Y]$, где $X, Y \in N \cup T$.

Таким образом, если слово w лежит в языке $D_n \cap R$, то оно является кодированием левого вывода некоторого слова u в грамматике G: по подслову $]_A[_{X_n}[_{X_{n-1}} \dots [_{X_{n-i}} \dots [_{X_1}\ _{X_1}]$ однозначно восстанавливается правило $A \to X_1X_2 \dots X_n$. Осталось определить морфизм h, который и даёт переход от слова $w \in R \cap D_n$ к слову $u \in L$. Он устроен следующим образом: $h(_\sigma]) = \sigma\ \forall \sigma \in \Sigma$, иначе $h(_X]) = \varepsilon\ \forall X \in N$ и $h([_X) = \varepsilon\ \forall X \in N \cup \Sigma$. Таким образом мы показали, что любой КС-язык представим в форме ХШ и привели эффективный алгоритм построения по языку формы ХШ.

Пусть теперь язык задан в представлении ХШ (D_n, R, h). Построим по нему МП-автомат, который будет недетерминировано угадывать прообраз $h^{-1}(w)$ слова w и проверять удовлетворяет ли он регулярному ограничению R и является ли он правильным скобочным выражением. Поскольку такой автомат можно построить, то язык, заданный в форме ХШ является КС-языком.

$\square$

Упражнение 1. В доказательстве есть тонкое место про кодирование левого вывода. Для достаточной строгости это утверждение надо доказать по индукции. Проведите это доказательство.

Аккуратно и понятно это сделать не очень просто. См. доказательства в следующем разделе, которые я постарался сделать законченными – идея в доказательстве упражнения такая же, как и в обосновании корректности алгоритма построения грамматики по МП-автомату.

2 От МП-автоматов к КС-грамматикам

Опишем алгоритм построения КС-грамматики G по N-автомату[3] $M = (\Sigma, \Gamma, Q, q_0, Z_0, \delta, \emptyset)$. $G = (N, T, P, S)$, причём

[3]допускающему по пустому стеку

- $T = \Sigma$;
- $N = \{[qZp] \mid q, p \in Q, Z \in \Gamma\}$
- Если $\delta(q, u, Z) \vdash (q_1, Y_1, Y_2, \ldots Y_n)$, то P содержит правила $[qZp] \to u[q_1Y_1r_1][r_1Y_2r_2] \ldots [r_{n-2}Y_{n-1}r_{n-1}][r_{n-1}Y_np]$, $u \in \Sigma \cup \varepsilon$, для всевозможных наборов состояний $r_1, r_2, \ldots, r_{n-1} \in Q$. Если $\delta(q, u, Z) = (p, \varepsilon)$, то P содержит правило $[qZp] \to u$.
- $\forall p \in Q \; S \to [q_0Z_0p] \in P$.

Обратите внимание, что слова u и v которые будут фигурировать дальше в правилах либо буквы, либо пустые слова. Идея алгоритма состоит в том, что левый вывод слова w в грамматике G соответствует успешной последовательности конфигураций на входе w автомата M. Состояние q в первом слева нетерминале и есть состояние, в котором находится автомат при обработке слова. Если автомат находясь в состоянии q, видя на верхушке стека Z, переходит в состояние q_1, и при этом кладёт что-то в стек, то в выводе грамматики нетерминал $[qZp]$ раскрывается как $u[q_1Y_1r_1][r_1Y_2r_2] \ldots [r_{n-2}Y_{n-1}r_{n-1}][r_{n-1}Y_np]$, таким образом в выводе грамматики слева опять оказывается текущее состояние автомата q_1, а также в нетерминалах закодировано содержимое его стека. Если же, автомат выталкивает символ Y_1 из стека читая v и переходит из состояния q_1 в r_1, то в грамматике есть правило $[q_1Y_1r_1] \to v$, таким образом $[qZp] \Rightarrow_L uv[r_1Y_2r_2] \ldots [r_{n-2}Y_{n-1}r_{n-1}][r_{n-1}Y_np]$ и опять таки в промежуточном шаге вывода закодировано текущее состояние автомата и содержание его стека. Вторые состояния в кодировке нетерминалов $[qZ\underline{p}]$ нужны для того, чтобы обеспечить корректность кодировки протокола переходов от одной поверхностной конфигурации к другой.

Перейдём к формальной части доказательства. Покажем что $L(M) \subseteq L(G)$.

Утверждение 1. *Если* $(q_0, uv, Z_0) \vdash^* (q, v, Y_1Y_2 \ldots Y_n)$*, тогда для всевозможных состояний* $r_1, r_2, \ldots r_n, p \in Q$ *справедливо*

$$S \Rightarrow_L^* u[qY_1r_1][r_1Y_2r_2] \ldots [r_{n-2}Y_{n-1}r_{n-1}][r_{n-1}Y_np].$$

Доказательство. Докажем индукцией по числу тактов k работы автомата M.

<u>База:</u> $k = 1$. Тогда, $(q_0, uv, Z_0) \vdash (q, v, Y_1 Y_2 \dots Y_n)$, переход происходит за один такт работы. Построим соответствующий вывод. Сначала применим правило $S \to [qZp]$, а далее раскрываем нетерминал $[qZp]$ – соответствующее правило есть в G по алгоритму построения.

<u>Переход:</u> пусть утверждение верно для $k = n$ – покажем, что оно верно для $k = n+1$. Пусть $(q_0, uv, Z_0) \vdash^* (q, v, Y_1 Y_2 \dots Y_n)$ и переход выполнен за $n+1$ такт работы автомата M. Рассмотрим конфигурацию, соответствующую n-ому такту автомата. Пусть она имеет вид $(q_1, u_l v, Z_1 Z_2 \dots Z_N)$ и при этом $(q_1, u_l v, Z_1 Z_2 \dots Z_N) \vdash (q, v, Y_1 Y_2 \dots Y_n)$. По предложению индукции,

$$S \Rightarrow_L^* u_1 \dots u_{l-1} [q_1 Z_1 r'_1][r'_1 Z_2 r'_2] \dots [r'_{n-2} Z_{n-1} r'_{n-1}][r'_{n-1} Z_N p]$$

За такт работы автомат раскрывает ровно один нетерминал, таким образом автомат делает переход $\delta(q_1, u_l, Z_1) \vdash (q, Y_1 Y_2 \dots Y_m)$, а $Y_{m+1} = Z_2, Y_{m+2} = Z_3, \dots$ Но тогда в грамматике по построению есть правило

$$[q_1 Z_1 r'_1] \to u_l [q Y_1 r_1][r_1 Y_2 r_2] \dots [r_{m-1} Y_m r_m][r'_1 Z_2 r'_2]$$

Тогда в результате переобозначения нетерминалов и в силу произвольности всех состояний, кроме q получаем, что

$$\begin{aligned} S &\Rightarrow_L^* u_1 \dots u_{l-1} [q_1 Z_1 r'_1][r'_1 Z_2 r'_2] \dots [r'_{n-2} Z_{n-1} r'_{n-1}][r'_{n-1} Z_N p] \Rightarrow_L \\ &\Rightarrow_L u_1 \dots u_{l-1} u_l [q Y_1 r_1][r_1 Y_2 r_2] \dots [r_{n-2} Y_{n-1} r_{n-1}][r_{n-1} Y_n p]. \end{aligned}$$

Переход доказан. $\square$

Из доказанного утверждения в частности следует, что если $(q_0, w, Z_0) \vdash^* (q, \varepsilon, \varepsilon)$, то $S \Rightarrow_L w$. Таким образом, мы показали, что $L(M) \subseteq L(G)$.

Доказательство обратного включения строится в том же духе, что и доказательства приведённых выше утверждений, поэтому его я оставляю в качестве упражнения.

Упражнение 2. Доказать, что $L(G) \subseteq L(M)$.

3 Задачи

Задача 1. Приведите грамматику G к нормальной форме Хомского. Все построения должны быть выполнены строго по алгоритмам. Эквивалентные преобразования допустимы, но если вы их делаете, то вы должны их явно указать и обосновать их корректность. Алгоритмы, аналогичные разобранным на семинаре, есть в книжке Хопкрофта Мотвани и Ульмана. Грамматика G задана правилами:

$$S \to A \mid B \mid C \mid E \mid AG$$
$$A \to C \mid aABC$$
$$B \to bABaF \mid aFCbDaGb \mid \varepsilon$$
$$C \to BaAbC \mid aGD \mid \varepsilon$$
$$F \to aBFaaCbA \mid aGE$$
$$E \to AF$$

Задача 2*. Проверьте по алгоритму Кока-Янгера-Касами порождает ли грамматика G из предыдущей задачи слово $aabaab$.

Задача 3. Язык L задан в ХШ-представлении: (D_2, Σ^*, φ), где D_2 – язык Дика с двумя типами скобок, регулярное ограничение Σ^* означает, что на слова не накладывается регулярное ограничение, морфизм φ определим следующим образом $\varphi\ :\ [_1 \to\ a;\ _1] \ \to\ b; [_2 \to\ b;\ _2] \ \to\ a$. Докажите или опровергните, что $L = \{w \mid |w|_a = |w|_b\}$.

Задача 4. Возьмите любой детерминированный МП-автомат, допускающий по пустому стеку, как минимум с двумя состояниями, распознающий КС-язык, не являющийся регулярным. Можете взять язык из примера в задании 7. Постройте по МП-автомату КС-грамматику, сделайте из неё приведённую грамматику. Будет ли она однозначна?