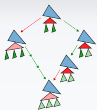


Кодирующие КС-языки



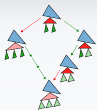
Теория формальных языков
2021 г.



Кодировка путей праволинейной грамматики

Рассмотрим путь вывода произвольного слова $a_1 \dots a_n$ в праволинейной грамматике. Он имеет вид $S \rightarrow a_1 A_1; A_1 \rightarrow a_2 A_2; \dots A_n \rightarrow a_n$. Применим к нему обратный гомоморфизм $h(A_i; A_i \rightarrow) = \varepsilon$ и сотрём префикс $S \rightarrow$, получим искомое слово.

Алфавит: $\Sigma \cup N \cup \{; , \rightarrow\}$. Описание языка:
 $\{S \rightarrow a_i (A_i; A_i \rightarrow a_j)^*\}$.



Кодировка путей праволинейной грамматики

Рассмотрим путь вывода произвольного слова $a_1 \dots a_n$ в праволинейной грамматике. Он имеет вид $S \rightarrow a_1 A_1; A_1 \rightarrow a_2 A_2; \dots A_n \rightarrow a_n$. Применим к нему обратный гомоморфизм $h(A_i; A_i \rightarrow) = \varepsilon$ и сотрём префикс $S \rightarrow$, получим искомое слово.

Алфавит: $\Sigma \cup N \cup \{; , \rightarrow\}$. Описание языка:

$\{S \rightarrow a_i (A_i; A_i \rightarrow a_j)^*\}$.

Описание языка привязано к множеству нетерминалов в рассматриваемой RLG.



Теорема Хомского–Шутценбергера

Пусть PAREN_n — язык из $4 * n$ элементов
 $\{[1,]_1, \dots, [n,]_n, (1,)_1, \dots, (n,)_n\}$.

Теорема

Любой CF-язык получается гомоморфизмом из языка
 $L' = \text{PAREN}_n \cap R$, где R — регулярный.

Пусть G — грамматика L в нормальной форме Хомского.
Пронумеруем правила G и поставим им в соответствие следующие.



Теорема Хомского–Шутценбергера

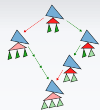
Пусть PAREN_n — язык из $4 * n$ элементов
 $\{[1,]_1, \dots, [n,]_n, (1,)_1, \dots, (n,)_n\}$.

Теорема

Любой CF-язык получается гомоморфизмом из языка
 $L' = \text{PAREN}_n \cap R$, где R — регулярный.

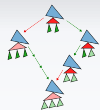
Пусть G — грамматика L в нормальной форме Хомского.
Пронумеруем правила G и поставим им в соответствие следующие.

- ❶ Если правило n имеет вид $A \rightarrow BC$, тогда порождаем правило $A \rightarrow [{}_n B]_n ({}_n C)_n$.
- ❷ Если правило n имеет вид $A \rightarrow a$, тогда порождаем правило $A \rightarrow [{}_n]_n ({}_n)_n$.



Свойства языка $L(G')$

- Все $]_n$ строго предшествуют $(_n$.



Свойства языка $L(G')$

- Все $]_n$ строго предшествуют $(_n$.
- Ни одна $)_n$ не предшествует непосредственно левой скобке.



Свойства языка $L(G')$

- Все $]_n$ строго предшествуют $(_n$.
- Ни одна $)_n$ не предшествует непосредственно левой скобке.
- Если правило n — это $A \rightarrow BC$, тогда $[_n$ непосредственно предшествует некоторой $[_p$, так же как и $(_n$.



Язык R

$R = \{x \in$
 $\{[j,]_j, (j,)_j\}^* \mid x \text{ начинается с } [{}_n \text{ для некоторого правила } n :$
 $A \rightarrow \dots \& \text{ все }]_n \text{ предшествуют } ({}_n\}.$



Язык R

$R = \{x \in \{[j,]_j, (j,)_j\}^* \mid x \text{ начинается с } [{}_n \text{ для некоторого правила } n : A \rightarrow \dots \& \text{ все }]_n \text{ предшествуют } ({}_n\}.$

Можно убедиться, что $L' = R \cap \text{PAREN}_n.$

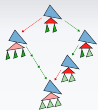


Язык R

$R = \{x \in \{[j,]_j, (j,)_j\}^* \mid x \text{ начинается с } [{}_n \text{ для некоторого правила } n : A \rightarrow \dots \& \text{ все }]_n \text{ предшествуют } ({}_n\}.$

Можно убедиться, что $L' = R \cap \text{PAREN}_n.$

Осталось определить h . Если n — нефинальное правило, то $h([{}_n) = h(]_n) = h(({}_n) = h()_n) = \varepsilon$. Иначе $h([{}_n) = a$, для остальных скобок так же.



Значение теоремы Х.-Ш.

Возможно разделить парсинг любого КС-языка на две стадии: лексический анализ (проверка условия R) и разбор правильных скобочных структур.

Замечание: поскольку гомоморфизм h не обязан быть инъективным, разбор ПСП не всегда можно определить однозначно. Пример: $\{a^n b^n\} \cup \{a^n b^{2n}\}$ (полностью неоднозначность устранить нельзя, т.к. этот язык не является детерминированным). Однако Т.Х.Ш. даёт подсказку, как строить КС-грамматики: надо найти в языке все скрытые «скобочные структуры».



Построение грамматики по X.-Ш.

Построить КС-грамматику для языка $\{a^n b^m c^k \mid n = 2 * m - k\}$.

Ищем возможную скобочную структуру. Для этого сначала избавимся от вычитания: $n + k = 2 * m$. Значит, буквы a должны балансироваться буквами b справа (т.е. буквы b являются «закрывающими скобками» для a), а буквы c — буквами b слева (т.е. буквы b являются «открывающими» для c). Возможны два случая: n и k оба чётны либо оба нечётны. Построим соответствующие им разбиения: $\{a^{2*n'} b^{n'} b^{k'} c^{2*k'}\}$ и $\{a a^{2*n'} b^{n'} b b^{k'} c^{2*k'} c\}$. Дальнейшее построение грамматики уже очевидно. Заметим, что гомоморфизм подразумевает минимум четыре вида скобок: пара $(_2a,)_b$, пара $(_b,)_{2c}$, внешняя пара $[_a,]_c$ (для нечётного варианта) и $[_b,]_\varepsilon$ для него же, чтобы породить внутреннюю букву b .



Построение грамматики по Х.-Ш.

Построить КС-грамматику для языка $\{a^n b^m c^k \mid n = 2 * m - k\}$.

Как итог, получаем язык, гомоморфно порождаемый языком Дика над $\{(2_a,)_b, (}_b,)_{2c}, [}_b,]_\epsilon, [}_a,]_c\}$ со следующим лексером:

- ❶ До $(2_a$ может идти лишь единственная $]_a$.
- ❷ После $)_b$ распознаётся одна $]_b$, если распозналась $]_a$.
- ❸ После $)_b$ или $]_\epsilon$ не может идти ничего другого, кроме $(}_b$ или $]_c$ (последняя — только после $]_\epsilon$).
- ❹ После $)_{2c}$ не может быть ничего, кроме $)_{2c}$ или $]_c$.

Дополнительное условие на существование $]_a$ уже не требуется — оно следует из сбалансированности ПСП.

Конструкция выше отличается от используемой в доказательстве теоремы — в целях экономии, в ней почти нет скобок, гомоморфно отображаемых в пустое слово.



Дополнительный пример

Построить КС-грамматику для $L_{\neq} = \{w_1cw_2 \mid w_i \in \{a, b\}^+ \text{ \& } w_1 \neq w_2\}$.

Классический пример грамматики с не-КС дополнением. Чтобы расшифровать неравенство, раскроем его в дизъюнкцию: «слово w_1 короче, чем w_2 ; либо w_2 короче, чем w_1 ; либо существует такое i , что w_1 и w_2 различаются в i -й позиции». Здесь условия не взаимоисключающие: достаточно одного из них, чтобы слово принадлежало L_{\neq} , но могут выполняться и два сразу.

Перепишем первое условие: $w_1cw'_1w_2$, где $|w_2| > 0$ и $|w_1| = |w'_1|$.

Очевидно, что «открывающими скобками» будут буквы из w_1 , «закрывающими» — из w'_1 , а «скобки» для w_2 замкнуты на самом w_2 .

Чтобы обеспечить четыре вида соответствий букв по счёту, придётся ввести четыре пары скобок для w_1 и w'_1 : $\{(a,)_a, (b,)_b, [a,]_b, [b,]_a\}$. И две пары скобок для w_2 : $\{\{a, \}_\varepsilon, \{b, \}'_\varepsilon\}$ и пара скобок для порождения c : $(c \text{ и })_\varepsilon$ (в нижних индексах — гомоморфные образы скобок).



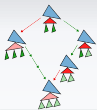
Дополнительный пример

Построить КС-грамматику для $L_{\neq} = \{w_1cw_2 \mid w_i \in \{a, b\}^+ \text{ \& } w_1 \neq w_2\}$.

Осталось построить регулярные условия. Для языка $w_1cw'_1w_2$, где $|w_2| > 0$ и $|w_1| = |w'_1|$, их можно описать следующим образом:

- ❶ После $(_a, ({}_b, [_a, [_b$ всегда идёт либо опять одна из таких скобок, либо $(_c$. Скобка $(_c$ единственна.
- ❷ После $)_c$, а также скобок $)_a,)_b,]_b,]_a$, могут идти либо $)_a,)_b,]_b,]_a$, либо фигурные скобки.
- ❸ В каждом слове есть хотя бы одна фигурная скобка. После открывающей фигурной скобки обязательно сразу идёт закрывающая, и после первой встреченной фигурной скобки все остальные скобки — тоже фигурные.

Представление Хомского–Шутценбергера для языка $w_1w_2cw'_1$, где $|w_2| > 0$ и $|w_1| = |w'_1|$, строится симметрично.



Дополнительный пример

Построить КС-грамматику для $L_{\neq} = \{w_1 c w_2 \mid w_i \in \{a, b\}^+ \text{ \& } w_1 \neq w_2\}$.

Осталось разобрать $\{w_1 t_1 w_2 c w_3 t_2 w_4 \mid |w_1| = |w_3| \text{ \& } t_1 \neq t_2\}$. Очевидно, что в нём «открывающими» будут элементы w_1 , закрывающими — элементы w_3 , t_1 и t_2 — уникальные скобки, отображающиеся в разные элементы алфавита, а w_2 и w_4 закрываются сами собой (как w_2 в предыдущем языке). Для $t_1 + t_2$ -скобок назначим пары $[_a^t,]_b^t$ и $[_b^t,]_a^t$, остальные обозначения сохраним те же. Лексер языка:

- ❶ После $(_a, (_b, [_a, [_b$ идёт либо опять одна из таких скобок, либо $[_a^t$ -скобка. $[_a^t$ единственна, за ней следует либо $\{_a$, либо $\{'_b$, либо $(_c$.
- ❷ После открывающей фигурной скобки сразу идёт закрывающая, и после первой встреченной фигурной скобки все остальные скобки — тоже фигурные, до конца строки либо до чтения скобки $(_c$.
- ❸ После $)_\varepsilon$, а также скобок $)_a,)_b,]_b,]_a$, могут идти либо $)_a,)_b,]_b,]_a$, либо скобка $]_a^t$. За $]_a^t$ следует EOL или $\{_a$ или $\{'_b$.

Чтобы получить прообраз языка L_{\neq} , объединим все три лексера.



Язык Грейбах

Здесь ε -free вариант. D — язык сбалансированных скобочных структур над $\{ (,), [,] \}$.

$$L_0 = \{ x_1 c y_1 c z_1 d \dots d x_n c y_n c z_n d \mid y_1 \dots y_n \in eD \text{ \& } z_i, x_i \text{ не содержат } e \text{ \& } y_1 \in e\{ (,), [,] \}^* \text{ \& } y_{i+1} \in \{ (,), [,] \}^* \}$$



Язык Грейбах

Здесь ε -free вариант. D — язык сбалансированных скобочных структур над $\{ (,), [,] \}$.

$$L_0 = \{ x_1 c y_1 c z_1 d \dots d x_n c y_n c z_n d \mid y_1 \dots y_n \in eD \text{ \& } z_i, x_i \text{ не содержат } e \text{ \& } y_1 \in e\{ (,), [,] \}^* \text{ \& } y_{i+1} \in \{ (,), [,] \}^* \}$$

Утверждение

Если L — CFL, тогда существует $h \in \text{Hom}$ такой, что $h^{-1}(L_0) = L$.



Гомоморфизм Грейбах

Пусть G — GNF грамматика для L . Пронумеруем нетерминалы G так, чтобы стартовый был первым. Построим вспомогательную функцию ξ :

- для правил $A_i \rightarrow a$ положим $\xi(i) =]^i$
- для правил $A_i \rightarrow aA_{j_1} \dots A_{j_n}$ положим $\xi(i) =]^i)([^m(\dots ([^1($
- если $i = 1$, тогда дополнительно припишем префикс $e([($.

Пусть терминалом a начинаются левые части правил k_1, \dots, k_m . Тогда $h(a) = c\xi(k_1)c \dots c\xi(k_m)d$.