А. А. Вылиток

Операции над формальными языками

Формальный язык L над алфавитом Σ — это подмножество множества всех цепочек в алфавите Σ, $L \subseteq \Sigma^*$.

Над языками можно выполнять обычные теоретико-множественные операции (объединение, пересечение, разность). Кроме теоретико-множественных, введем еще несколько специальных операций над языками.

Пусть L, L_1, L_2 — языки над алфавитом Σ. Определим операции[1]:

1) $L_1 \cup L_2$ — *объединение* L_1 и L_2;
2) $L_1 \cap L_2$ — *пересечение* L_1 и L_2;
3) $L_1 - L_2$ — *разность* L_1 и L_2;
4) $\overline{L} = \Sigma^* - L$ — *дополнение* L (до Σ^*);
5) $L_1 \cdot L_2 = \{\varphi\psi \mid \varphi \in L_1, \psi \in L_2\}$ — *сцепление*[2] L_1 и L_2;
6) $L^i = \begin{cases} \{\varepsilon\}, & i = 0, \\ L = L^{i-1} \cdot L, & i \geq 1 \end{cases}$ — *i-я степень* L;
7) $L^* = \bigcup_{n=0}^{\infty} L^n$ — *итерация* L;
8) $L^+ = \bigcup_{n=1}^{\infty} L^n$ — *усеченная итерация* L;
9) *Подстановка* L_1, ..., L_n в L вместо a_1, ..., $a_n \in \Sigma$, где $L_i \subseteq \Sigma_i^*$ есть операция $Subst(L;\ a_1,\ ...,\ a_n \mid L_1,\ ...,\ L_n) = \{\ \chi_{i_1},\ ...,\ \chi_{i_k} \mid a_{i_1} \ldots a_{i_k} \in L,\ \chi_{i_j} \in L_j\} \cup L'$, где $L' = \{\varepsilon\}$, если $\varepsilon \in L$, и $L' = \varnothing$, если $\varepsilon \notin L$.

Операции позволяют конструировать новые языки из набора заданных. Пусть, например, $L_1 = \{a,\ ab,\ bb\}$, $L_2 = \{aa,\ ba,\ ab,\ b\}$, $L_3 = \{abab,\ aa\}$. Тогда выражение $(L_1 \cap L_2)^* - L_3$ задает язык $L = \{(ab)^n \mid n \geq 0,\ n \neq 2\}$.

Множество языков называют *классом*. Например, класс контекстно-свободных (КС) языков состоит из всех языков, порождаемых КС-грамматиками. Говорят, что класс языков $\mathcal{C}$ *замкнут относительно операции* θ, если для любых языков из $\mathcal{C}$ результат операции θ также принадлежит $\mathcal{C}$.

Покажем, что класс КС-языков замкнут относительно операций объединения, сцепления и итерации. Для каждой из этих операций достаточно указать способ — как из КС-грамматик, порождающих исходные языки, построить КС-грамматику, порождающую язык, являющийся результатом операции. Будем считать, что множества нетерминалов исходных грамматик не пересекаются. Этого всегда можно добиться переименованием нетерминальных символов (имена нетерминалов на порождаемый грамматикой язык не влияют).

Пусть $G_1 = (T_1,\ N_1,\ P_1,\ S_1)$, $G_2 = (T_2,\ N_2,\ P_2,\ S_2)$ — КС-грамматики, $N_1 \cap N_2 = \varnothing$. Построим грамматику G, такую что:

---

[1] Знаки равенства во всех пунктах этого определения следует читать как «равно по определению».

[2] Эту операцию называют также конкатенацией, иногда — (прямым) произведением языков. Знак операции «·» может быть опущен.

а) $L(G) = L(G_1) \cup L(G_2)$:

$G = (T_1 \cup T_2, N_1 \cup N_2 \cup \{S\}, P_1 \cup P_2 \cup \{S \to S_1 \mid S_2\}, S), S \notin N_1 \cup N_2;$

б) $L(G) = L(G_1)L(G_2)$:

$G = (T_1 \cup T_2, N_1 \cup N_2 \cup \{\mathrm{S}\}, P_1 \cup P_2 \cup \{S \to S_1S_2\}, S), S \notin N_1 \cup N_2;$

в) $L(G) = L(G_1)^*$:

$G = (T_1, N_1 \cup \{S\}, P_1 \cup \{S \to S_1S \mid \varepsilon\}, S), S \notin N_1.$

Класс КС-языков не замкнут относительно операции пересечения. Например, язык $\{a^n b^n c^n \mid n \geq 0\}$, не являющийся контекстно-свободным, может быть получен пересечением двух КС-языков: $\{a^k b^k c^i \mid i, k \geq 0\}$ и $\{a^m b^j c^j \mid j, m \geq 0\}$. Класс КС-языков не замкнут относительно операции вычитания (разности). Это нетрудно доказать, учитывая, что пересечение выражается через объединение и разность.

Задача. Показать, что класс КС-языков замкнут относительно операции подстановки.

Пример. С помощью операций над языками построим КС-грамматику G, порождающую все непустые цепочки в алфавите $\{a, b\}$, кроме цепочки bb.

Заданный язык можно представить так: $L = \{a, b\}^+ - \{bb\}$. Рассмотрим вспомогательные языки, и КС-грамматики, их порождающие.

$L_1 = \{a, b\}^+$	G_1:	$S_1 \to aS_1 \mid bS_1 \mid a \mid b$	$L(G_1) = L_1$
$L_2 = \{a, b\}$	G_2:	$S_2 \to a \mid b$	$L(G_2) = L_2$
$L_3 = \{aa, ab, ba, bb\}$	G_3:	$S_3 \to aa \mid ab \mid ba \mid bb$	$L(G_3) = L_3$
$L_4 = L_3L_1$	G_4:	$S_4 \to S_3S_1$	$L(G_4) = L_4$
$L_5 = \{aa, ab, ba\}$	G_5:	$S_5 \to aa \mid ab \mid ba$	$L(G_5) = L_5$

Язык L_1 представляет собой множество всех непустых цепочек в алфавите $\{a, b\}$, L_2 — множество всех однобуквенных цепочек, L_3 — множество всех двухбуквенных цепочек, L_4 — множество всех цепочек длины не меньшей трех, L_5 — множество всех двухбуквенных цепочек, кроме bb. Язык L является объединением языков L_2, L_5,L_4: $L = L_2 \cup L_5 \cup L_4$. Построим грамматику G для L, используя уже построенные грамматики для вспомогательных языков.

G:
$$\begin{aligned} S &\to S_2 \mid S_5 \mid S_4 \\ S_2 &\to a \mid b \\ S_5 &\to aa \mid ab \mid ba \\ S_4 &\to S_3S_1 \\ S_3 &\to aa \mid ab \mid ba \mid bb \\ S_1 &\to aS_1 \mid bS_1 \mid a \mid b \end{aligned}$$

Для языка L можно построить и регулярную грамматику:

$$\begin{aligned} S &\to a \mid b \mid aS_1 \mid bS_2 \mid aS_3 \mid bS_3 \\ S_1 &\to a \mid b \\ S_2 &\to a \\ S_3 &\to aS_4 \mid bS_4 \\ S_4 &\to aS_4 \mid bS_4 \mid a \mid b \end{aligned}$$