ФПМИ МФТИ
Теория формальных языков
Алгоритм LR(k)

LR(k) алгоритм или заклинание, которое позволяет изгнать строгого преподавателя на экзамене.

Далее стрелочка означает вывод за один шаг, а стрелочка со звёздочкой за произвольное. Все выводы правосторонние.

Определение. Грамматика является LR(k) грамматикой, если из условий:

1) $S' \Rightarrow^* \alpha A w \Rightarrow \alpha\beta w$
2) $S' \Rightarrow^* \gamma B x \Rightarrow \alpha\beta y$
3) $First_k(w) = First_k(y)$

Следует, что $\alpha A y = \gamma B x$, то есть $\alpha = \gamma$, A=B, x=y.

Обратите внимание, что не обязательно $w = y$, так как выводы могут быть для разных слов. Более того, $x = y$ может содержать нетерминалы, если последнее правило второго вывода было $B \to \eta_1\beta\eta_2$.

Более неформально, определение говорит, что какое бы слово не выводилось, по стеку и следующим k буквам можно однозначно восстановить один шаг вывода.

Теорема 5.9

Грамматика является LR(k) грамматикой т. и т.т. когда следующие две ситуации не бывают одновременно допустимы для некоторого активного префикса $\alpha\beta$: $A \to \beta\cdot$, u и $A_1 \to \beta_1 \cdot \beta_2$, v и $u \in EFF_k(\beta_2, v)$. Напомним, что EFF_k — это такие префиксы выводимых слов, что если первый символ является нетерминальным, то он не раскрывается как пустое слово. При этом $beta_2$ может быть пустым (тогда эта ситуация порождает конфликтный reduce).

Доказательство

Необходимость. Пусть всё-таки две ситуации различны и оказались допустимы. Тогда есть два вывода:

$S' \Rightarrow^* \alpha A w \Rightarrow \alpha\beta w$
$S' \Rightarrow^* \alpha_1 A_1 x \Rightarrow \alpha_1\beta_1\beta_2 x$

и $First_k(w) = u$, $First_k(x) = v$, $\alpha\beta = \alpha_1\beta_1$. Пусть $\beta_2 x \Rightarrow^* uy$ для некоторого слова y (хвост слова, следующий за фиксированными k буквами), причём если β_2 начинается с нетерминала, то он не открывается как пустое слово. Покажем, что грамматика не LR(k).

Если $\beta_2 = \epsilon$, то u=v и выводы имеют вид:

$S' \Rightarrow^* \alpha A w \Rightarrow \alpha\beta w$
$S' \Rightarrow^* \alpha_1 A_1 x \Rightarrow \alpha_1\beta_1 x$

и $First_k(w) = u = First_k(x) = v$. Тогда, в силу различности ситуаций, либо $A_1 \neq A$, либо $\beta \neq \beta_1$. Оба случая явно противоречат определению.

Если $\beta_2 = z$ состоит только из терминальных символов, то выводы имеют вид:

$S' \Rightarrow^* \alpha A w \Rightarrow \alpha\beta w$
$S' \Rightarrow^* \alpha_1 A_1 x \Rightarrow \alpha_1\beta_1 z x$

и $\alpha\beta = \alpha_1\beta_1$, $First_k(zx) = u$. В определении требуется $\alpha A z x = \alpha_1 A_1 x$. Но z непусто (пустой случай рассмотрен выше), значит строки не равны, так как A_1 — нетерминал.

Случай 3. Уберите особо впечатлительных от экрана.

Пусть в β_2 есть нетерминал. Рассмотрим, как раскрывается β_2 в течении вывода: $\beta_2 \Rightarrow^* u_1 B u_3 \Rightarrow u_1 u_2 u_3$, причём $u_1 u_2 \neq \epsilon$, так как первому нетерминалу запрещено раскрываться в пустое слово в этом выводе (иначе противоречие с условием теоремы, вторая ситуация допустима с словом u только при условии, что оно лежит в $EFF_k(\beta_2 v)$). Тогда первый вывод не меняется, а второй можно расширить: $S' \Rightarrow^* \alpha_1 A_1 x \Rightarrow \alpha_1\beta_1\beta_2 x \Rightarrow \alpha_1\beta_1 u_1 B u_3 x \Rightarrow \alpha_1\beta_1 u_1 u_2 u_3 x$,

что $\alpha_1\beta_1 = \alpha\beta$ и $u_1 u_2 u_3 x = uy$ (то есть мы явно написали, как из $\beta_2 x$ выводится uy. Рассмотрим определение LR(k) грамматики (для символов A и B, то есть для непосредственно предпоследних шагов в выводах). В нём требуется $\alpha A u_1 u_2 u_3 x = \alpha_1\beta_1 u_1 B u_3 x$ (действительно, в терминах определения $\gamma = \alpha_1\beta_1 u1$, $y = u_1 u_2 u_3 x$ — не перепутайте его с y из теоремы, мы его тоже ввели). Тогда сократим последние две строки и получим $\alpha A u_1 u_2 = \alpha_1\beta_1 u_1 B$. Подставим $\alpha\beta = \alpha_1\beta_1$ (активный префикс один и тот же), сократим α и получим, что $A u_1 u_2 = \beta u_1 B$, что невозможно, так как $u_1 u_2$ непусто (заканчивается на терминал, точно не равный нетерминалу B). Заметим, что противоречие получается именно благодаря тому, что используется не First, а EFF. Действительно, если бы перенос определялся по First, то он был бы возможен одновременно со свёрткой (и читалась бы буква из другого поддерева).

Необходимость установлена.

Для установления достаточности получим следующий факт. Если грамматика не является $LR(k)$ грамматикой, то существуют два вывода из определения, причём во втором выводе последним правилом является $B \to \delta$ и $\gamma\delta x = \alpha\beta y$ю Ключевой момент заключается в том, что $|\gamma\delta| \geq |\alpha\beta|$, но $\gamma B x \neq \alpha A y$ (верно чуть более сильное чем определение свойство).

Пусть мы нашли контрпример в котором $|\gamma\delta| < |\alpha\beta|$. Так как $\gamma\delta x = \alpha\beta y$ и $|\gamma\delta| < |\alpha\beta|$, то для некоторого слова z верно $\alpha\beta = \gamma\delta z$ (сократим лишние буквы в равенстве выше). Таким образом есть два вывода:

$S' \Rightarrow^* \gamma B x \Rightarrow \gamma\delta x$ и $S' \Rightarrow^* \alpha A w \Rightarrow \alpha\beta w = \gamma\delta z w$. Так как z это то, что осталось от сокращения букв, то $x = zy$. Так как $First_k(w) = First_k(y)$, то $First_k(x) = First_k(zw)$. Если верно LR(k) определение, то $\alpha A w = \gamma B z w$ (следите за порядком, в котором написаны выводы тут и в определении). Тогда, сокращая w получим $\alpha A = \gamma B z$ и $\alpha A y = \gamma B z y = \gamma B x$ $(zy = x)$. Но это неверно, так как изначально рассмотренная пара выводов была контрпримером к определению. Таким образом, построенная пара выводов является контрпримером, для которого условие леммы выполнено (с точностью до переименования строк. Следите за порядком, в котором поданы выводы, он совпадает с условием теоремы, просто строки и буквы переназваны).

Тогда в силу доказанной леммы построим пару выводов так, чтобы $|\gamma\delta| \geq |\alpha\beta|$. Пусть $\alpha_1 A_1 y_1$ — последняя строка в выводе $S' \Rightarrow^* \gamma B x$ (напомним, что вывод —

последовательность строк, где каждая отличается раскрытием одного и только одного нетерминала), что длина $|\alpha_1 A_1| < |\alpha\beta| + 1$. Тогда второй вывод из пары контрпримеров можно записать как $S' \Rightarrow^* \alpha_1 A_1 y_1 \Rightarrow \alpha_1\beta_1\beta_2 y \Rightarrow \alpha_1\beta_1 y$, где $\alpha_1\beta_1 = \alpha\beta$. В силу выбора строки $\alpha_1 A_1 y_1$ имеем $|\alpha_1| \leq |\alpha\beta| \leq |\gamma\delta|$ и на шаге вывода $\beta_2 y_1 \Rightarrow^* y$ не участвует правило $B \to \epsilon$, где $B = \beta_2[0]$ (самая левая буква). Если бы это правило применялось последним, то выбранное слово $\alpha_1 A_1 y_1$ было бы не последним в выводе $S' \Rightarrow^* \gamma\beta x$ с длиной части до открывающегося нетерминала $< |\alpha\beta| + 1$ (так как последним является правило $B \to \epsilon$, то его применение не удлиняет часть правой части до нетерминала, добавляя к нему пустое слово, значит эта строка тоже подходит и находится на два шага дальше в выводе). Значит $u = First_k(y)$ попадает в $EFF_k(\beta_2 y_1)$ и ситуация $A_1 \to \beta_1 \cdot \beta_2, v$ допустима для $\alpha\beta$, $v = First_k(y_1)$. Ситуация $A \to \beta\cdot, u$ тоже допустима, так как есть соответствующий вывод. Если эти ситуации равны, то два соответствующих им вывода (построенных выше) совпадают, что противоречит тому, что они выбирались разными.

Осталось доказать теорему 5.10 о том, что алгоритм построения $V_k^G(\gamma)$ действительно строит все допустимые ситуации, что является несложным техническим результатом, требующим аккуратной расписки выводов. Остаётся на самостоятельное изучение.