Упражнение 12

1 Синтактични дървета на извод

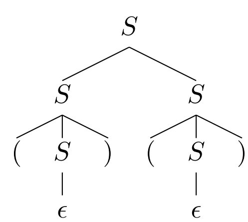
Нека G е контекстно-свободна граматика. Както вече видяхме, една дума $w \in L(G)$ може да има няколко извода по G. Да разгледаме по-прост пример отново използващ граматиката, генерираща думите от балансирани леви и десни скоби — думата ()() има два различни извода, именно,

$$S \Rightarrow SS \Rightarrow (S)S \Rightarrow ()S \Rightarrow ()(S) \Rightarrow ()()$$

и

$$S \Rightarrow SS \Rightarrow S(S) \Rightarrow (S)(S) \Rightarrow (S)S \Rightarrow ()()$$

В действителсност, тези два извода са в някакъв смисъл "еднакви". Използваните правила са едни и същи и се прилагат на едни и същи позиции в междинните низове. Единствената разлика е в *реда*, в който тези правила се прилагат. Интуитивно, и двата извода могат да се получат чрез обхождане на следното дърво.



Неформално, една такава картинка наричаме **синтактично дърво на извод**. Точките наричаме **върхове**; всеки връх има етикет, който е елемент на $\Sigma \cup V$. Най-горния връх наричаме **корен**, а най-долните върхове наричаме **листа**. Всички листа са етикетирани с терминали или с ϵ. Конкатенирайки етикетите на листата в ред от ляво надясно, получаваме изведената дума, която още ще наричаме **продукт** на дървото. По-формално, за дадена контекстно-свободна граматика $G = (V, \Sigma, R, S)$, дефинираме синтактично дърво на извод заедно с неговите корен, листа и продукт индуктивно, както следва.

1. $\circ\ \sigma$

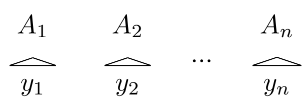
Това е синтактично дърво за всяко $\sigma \in \Sigma$. Единственият връх на това дърво е едновременно негов корен и единствено листо. Продукта на дървото е σ.

2. Ако $A \to \epsilon$ е правило в R, то

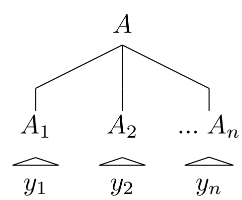
$$\begin{array}{c} A \\ | \\ \epsilon \end{array}$$

е синтактично дърво на извод; корена му е върхът с етикет A, единственото му листо е върхът с етикет ϵ, и продукта му е ϵ.

3. Ако



са синтактични дървета, за $n \geq 1$, с корени етикетирани $A_1, ..., A_n$ съответно, и продукти $y_1, ..., y_n$, и $A \to A_1 ... A_n$ е правило в R, то



е синтактично дърво на извод. Корена му е новият връх етикетиран A, листата му са листата на съставящите го синтактични дървета, и продукта му е $y_1 ... y_n$.

2 Езици, които не са контекстно-свободни

Множеството на контекстно-свободните езици над дадено множество от терминали и нетерминали $\Sigma \cup V$ също е изборимо безкрайно. За жалост нямаме толкова подходящо описание на контекстно-свободните езици, каквито бяха регулярните изрази за регулярните езици, с помощта на което да съобразим този факт. Достатъчно е обаче да забележим, че множеството на всички възможни правила над $\Sigma \cup V$ е $V \times (\Sigma \cup V)^*$, което е изброимо безкрайно, по следните съображения:
— V е крайно;
— Σ е крайно;
— $\Sigma \cup V$ е крайно м-во, като обединение на крайни м-ва;
— $(\Sigma \cup V)^*$ е изброимо безкрайно;
— декартовото произведение на крайно множество с избоимо безкрайно такова е изброимо безкрайно.

Правилата на една граматика са *крайно* подмножество на изброимо безкрайното множество $V \times (\Sigma \cup V)^*$. Тоест, множеството от всички възможни множества от правила е множеството от всички *крайни* подмножества на изброимо безкрайно множество. Като такова, то е изброимо безкрайно (всяко крайно подмножество може да се кодира в разлагане на прости множители на някое естествено число; това кодиране е съобразено с даденото ни изброяване). Възможните множества от правила образуват изброимо безкрайно множество. Остана само да съобразим, че възможните избори на начална променлива образуват крайно множество и можем да заключим, че има изброимо много контекстно-свободни граматики при фиксирано множество от терминали и нетерминали, а от тук и контекстно-свободните езици са изброимо много.

Щом множеството на контекстно-свободните езици е изброимо безкрайно, а множеството на всички езици е неизброимо безкрайно, то съществуват езици, които не са контекстно-свободни. За да открием критерий за това, кога един език не е контекстно-свободен отново е смислено да търсим свойство, което всички безкрайни контекстно-свободни езици притежават.

Нека $G = (V, \Sigma, R, S)$ е контекстно-свободна граматика. С $\phi(G)$ означаваме най-големият брой символи в дясната страна на кое да е правило в R. **Път** в синтактично дърво на извод е редица от върхове свързани със линия в дървото на извод; първия връх е коренът, а послелния е листо. **Дължината** на този път е броят на свързващите линии в него. **Височината** на едно синтактично дърво е дължината на най-дългия път в него.

Лема 1. Продуктът на дадено синтактично дърво на G с височина h има дължина не повече от $\phi(G)^h$.

Доказателството на горната лема е проста индукция по $h \geq 1$. Лемата ни казва, че синтактичното дърво на коя да е дума $w \in L(G)$ с $|w| > \phi(G)^h$ трябва да има път по-дълъг от h. Това е ключово за доказателството на следната теорема, представляваща търсения от нас критерии за контекстно-свободност.

Теорема 1(лема за разрастването за контекстно-свободни езици). Нека L е контекстно-свободен език. Тогава съществува естествено число $p \geq 1$, такова че всяка дума $w \in L$ с $|w| \geq p$ може да се презапише във вида $w = vwxyz$, като
(1) $|wy| \geq 1$,
(2) $|wxy| \leq p$ и
(3) $(\forall i \in \mathbb{N})[vw^i xy^i z \in L]$.

Нека $G = (V, \Sigma, R, S)$ е такава контекстно-свободна граматика, че $L(G) = L$. Ясно е, че един възможен избор за p е именно $\phi(G)^{|V|} + 1$ — ако w е дума от $L(G)$, такава че $|w| \geq \phi(G)^{|V|} + 1$, то всяко синтактично дърво за w ще има продукт с дължина поне $\phi(G)^{|V|} + 1$, тоест по-голяма от $\phi(G)^{|V|}$. Съгласно **Лема 1**, всяко такова дърво трябва да има път по-дълъг от $|V|$, тоест с дължина поне $|V|+1$. Този път ще съдържа $|V|+2$ върха, от които точно един — листото — е етикетиран с терминал, а останалите $|V|+1$ са етикетирани с нетерминали. Тогава е ясно, че поне два върха по пътя ще са етикетирани с един и същ нетерминал. Ако v_1 и v_2 са два такива върха, то заменяйки "поддървото, вкоренено във v_1" с "поддървото, вкоренено във v_2" получаваме синтактично дърво на извод за vw^0xy^0z. Заменяйки i на брой пъти последователно "поддървото, вкоренено във v_2" с "поддървото, вкоренено във v_1" получаваме синтактично дърво на извод за $vw^{i+1}xy^{i+1}z$.

Пример 1. $L = \{a^n b^n c^n \mid n \in \mathbb{N}\}$ не е контекстно-свободен. За да докажем това, да допуснем, че L контекстно-свободен и нека p е въпросното число от **лемата за разрастването**. Съгласно лемата, думата $w = a^p b^p c^p$ може да се презапише във вида $w = vwxyz$ за $wy \neq \epsilon$ и $vw^ixy^iz \in L$ за всяко $i \in \mathbb{N}$. Има два случая, всеки от които води до противоречие.
1сл. wy съдържа появи на всяка от трите букви a, b и c. Тогава поне една измежду w и y трябва да съдържа появи на поне две от тези букви. Веднага се вижда, че редът на буквите в vw^2xy^2z е развален — има b преди a, или c преди a или b.
2сл. wy съдържа появи само на някои, но не всяка от буквите a, b и c. Тогава веднага можем да съобразим, че vw^2xy^2z ще има неравен брой a-та, b-та и c-та.

Пример 2. $L = \{a^n \mid n \geq 1 \text{ е просто число}\}$ не е контекстно-свободен. За да покажем това, да допуснем, че L е контекстно-свободен и нека p е въпросното число от **лемата за разрастването**. Нека p' е първото просто число по-голямо от p. Тогава думата $a^{p'}$ може да се презапише във вида $w = vwxyz$, където $wy \neq \epsilon$. Тогава $wy = a^q$ и $vxz = a^r$, където q и r са естествени числа и $q > 0$. Съгласно лемата, $vw^ixy^iz \in L$ за всяко $i \in \mathbb{N}$. Тоест, $r + iq$ е просто число за всяко $i \in \mathbb{N}$. За $i = q + r + 1$ обаче имаме, че

$$\begin{gathered} r + iq = \\ r + (q + r + 1)q = \\ r + q^2 + rq + q = \\ r(q+1) + q(q+1) = \\ (q+1)(r+q), \end{gathered}$$

което е произведение на две числа, всяко от които е по-голямо от нула и значи няма как да е просто число. Противоречие, породено от допускането ни, че L е контекстно-свободен език.

3 Задачи

Задача 1. Използвайте лемата за разрастването за да докажете, че следните езици не са контекстно-свободни.
(а) $\{a^{n^2} \mid n \in \mathbb{N}\}$
(б) $\{www \mid w \in \{a,b\}^*\}$
(в) $\{w \in \{a,b,c\}^* \mid w$ съдържа равен брой a-та, b-та и c-та $\}$

Задача 2. Използвайте лемата за разрастването за да докажете, че езикът $L = \{babaabaaab...ba^{n-1}ba^nb \mid n \geq 1\}$ не е контекстно-свободен.

Задача 3. Кои от следните езици са контекстно-свободни? Обосновете се.
(а) $\{a^mb^nc^p \mid m = n$ или $n = p$ или $m = p\}$
(б) $\{a^mb^nc^p \mid m \neq n$ или $n \neq p$ или $m \neq p\}$
(в) $\{a^mb^nc^p \mid m = n \;\&\; n = p \;\&\; m = p\}$

4 Решения

Задача 1. (а) Да допуснем, че L е контекстно-свободен. Разглеждаме думата $a^{p^2} \in L$. Съгласно лемата за разрастването $a^{p^2} = vwxyz$ за някои $v,w,x,y,z \in \{a\}^*$. Нека $w = a^k$ и $y = a^l$ съгласно свойства (1) и (2) от лемата, $1 \leq k + l \leq p$. От тук имаме, че $p^2 + 1 \leq p^2 + k + l \leq p^2 + p$. Тоест, $p^2 < p^2 + k + l \leq p^2 + p < p^2 + 2p + 1$. Значи $p^2 < p^2 + k + l < (p+1)^2$. Тоест $p^2 < |vw^2xy^2z| < (p+1)^2$, откъдето следва, че $|vw^2xy^2z|$ няма как да бъде точен квадрат. Следователно $vw^2xy^2z \notin L$. Намерихме число, за което прилагайки свойство (3) стигаме до дума извън L. Противоречие с лемата. Значи L не е контекстно-свободен.

(б) Разглеждаме думата $(a^pb)^3 \in L$. За wxy имаме следните два случая:
(1) $wxy = a^k$, аз някое $1 \leq k \leq p$. Тогава лесно може да се съобрази, че покачвайки нагоре ще излезем от езика.
(2) $wxy = a^kba^l$ за някои $1 \leq k + l + 1 \leq p$. Имаме следните подслучаи:

(2.1) Буквата b е в w или в y. Тогава думата vw^0xy^0z ще има две b-та и няма как да е в L.

(2.2) Буквата b е в x. Тогава както и да изберем да се възползваме от свойство (3) на лемата, ще стигнем до дума извън L, защото броят на a-тата няма да е един и същ преди всяко от трите b-та, а е ясно, че всяка разбивка на покачената дума на три еднакви части www ще има свойството, че w завършва на b.

(в) Разглеждаме думата $a^pb^pc^p \in L$. За wxy имаме следните два случая:
(1) $wxy = \sigma^k$ за някои $\sigma \in \{a,b,c\}$ и $1 \leq k \leq p$. Тук е ясно, че както и да покачим, излизаме от езика.
(2) wxy попада в интервал, застъпващ *точно* две различни букви. Неза-

висимо как са разпределени тези две букви между w и y, тъй като $|wy| \geq 1$, то покачвайки в коя да е посока ще получим дума, в която бройките на трите различни букви не са равни.

Задача 2. Да разгледаме думата $babaabaaab...ba^{p-1}ba^pb \in L$. Нека $vwxyz$ е нейно разбиване със свойствата от лемата за разрастването. Случаите, в които wxy попада изцяло или частично преди a^pb са ясни — покачвайки в която и да е посока, линейно нарастващата структура на думата се нарушава и попадаме извън езика. Да разгледаме случая, в който wxy попада изцяло в a^pb. Имаме следните подслучаи:
(1) $wxy = a^k$ за някое $1 \leq k \leq p$. Тук нещата пак са ясни — ако се опитаме да приложим св-во (3) веднага разваляме линейно нарастващата структура на a-тата.
(2) $wxy = a^kb$ за някое $1 \leq k \leq p$. Ако w и y са едновременно непразни, то в частност y съдържа b-то, а $w = a^l$, за някое $l \leq k$ и покачвайки нагоре добавяме a-та преди последното b в оригиналната дума, което разваля структурата. Ако само y е празна, то x съдържа b-то и отново имаме същия проблем. Ако и x и y са празни, то $w = a^kb$ и покачвайки нагоре получаваме думата $babaabaaab...ba^{p-1}ba^pba^kb$, която не е в L, тъй като $k \leq p$, тоест $k \neq p+1$. Ако w е празна, а y е непразна, то $y = a^lb$, за някое $l \leq k$ и покачвайки нагоре получаваме същия проблем. С това случаите се изчерпаха.

Задача 3. (а) Този език е $\{a^mb^mc^n \mid m,n \in \mathbb{N}\} \cup \{a^mb^nc^n \mid m,n \in \mathbb{N}\} \cup \{a^mb^nc^m \mid m,n \in \mathbb{N}\}$ и следователно е контекстно-свободен. Граматики ще дадем за първия и третия от тези операнди.
(1) $S \to Sc \mid A$
$A \to aAb \mid \epsilon$.
(2) $S \to aSc \mid B$
$B \to bB \mid \epsilon$.

(б) Този език е $\{a^mb^nc^p \mid m \neq n\} \cup \{a^mb^nc^p \mid n \neq p\} \cup \{a^mb^nc^p \mid m \neq p\}$. На свой ред той е равен на $[\{a^mb^n \mid m \neq n\} \circ \mathscr{L}(c^*)] \cup [\mathscr{L}(a^*) \circ \{b^nc^p \mid n \neq p\}] \cup \{a^mb^nc^p \mid m \neq p\}$. Всеки от тези езици знаем вече, че е контекстно-свободен, освен $\{a^mb^nc^p \mid m \neq p\}$. Контекстно-свободна граматика за този език е например:

$$\begin{gathered} S \to aA \mid Cc \\ A \to aA \mid aAc \mid B \\ C \to Cc \mid aCc \mid B \\ B \to bB \mid \epsilon \end{gathered}$$

Следователно и целият език е контекстно-свободен.

(в) Това е езикът $\{a^nb^nc^n \mid n \in \mathbb{N}\}$, за който вече показахме, че не е контекстно свободен.