Контекстно-свободные грамматики для описания человеческих языков и как расширить их возможности

Обозначения: $A^* = \{a_1 \dots a_n \mid n \geq 0, a_i \in A\}$ — множество конечных строк, символы которых берутся из множества A (включая строку длины 0, обозначаемую Λ); $A^+ := A^* \setminus \{\Lambda\}$. Если $\alpha \in A^*$, то α^n — это строка, получающаяся n-кратным копированием строки α.

Определение 1. *Формальная грамматика* — структура $Gr = \langle N, \Sigma, P, S \rangle$, где N и Σ — конечные множества (алфавиты нетерминальных и терминальных символов соответственно), $S \in N$ — стартовый символ, P — конечное множество правил вида $\alpha \to \beta$, где $\alpha, \beta \in (\Sigma \cup N)^*$, причем в α есть хотя бы один нетерминальный символ.

Определение 2. Говорят, что грамматика $Gr = \langle N, \Sigma, P, S \rangle$ непосредственно выводит строку ψ из строки φ, если $\varphi = \gamma\alpha\delta$, $\psi = \gamma\beta\delta$ и $\alpha \to \beta \in P$. Обозначение: $\varphi \Rightarrow_{Gr} \psi$ (или просто $\varphi \Rightarrow \psi$). Транзитивное рефлексивное замыкание $\Rightarrow$ обозначаем $\Rightarrow^*$; $\alpha \Rightarrow^* \beta$ произносим как "β выводится из α".

Определение 3. Язык, задаваемый грамматикой $Gr = \langle N, \Sigma, P, S \rangle$, — множество строк из терминальных символов, которые можно вывести из S; то есть $L(Gr) = \{\alpha \in \Sigma^* \mid S \Rightarrow^* \alpha\}$.

1 Контекстно-свободные грамматики

Определение 4. Контекстно-свободная грамматика — грамматика, в которой все правила имеют вид $A \to \alpha$, где $A \in N$ — нетерминальный символ, $\alpha \in (\Sigma \cup N)^*$ — строка из терминальных и нетерминальных символов. Язык, задаваемый контекстно-свободной грамматикой, называется контекстно-свободным.

Замечание. Термины "алфавит", "символ" могут немного путать при разговоре о естественных языках. К примеру, если мы рассмотрим грамматику с правилами $S \to$ *Вася* VP, $VP \to$ *спит*, то в ней все слово *Вася* (а не отдельно буквы *В*, *а*, *с* и *я*) будет являться одним терминальным символом. Кстати, VP — тоже один символ, только нетерминальный (лингвистически мы хотим, чтобы все, что из него выводится, было глаголом со своими зависимыми, скажем, *ест*, или *любит Машу*, или *третьи сутки готовится к зачетной сессии*; на согласование мы внимание не обращаем).

Теорема 1. *Пусть L — контекстно-свободный язык над алфавитом Σ. Тогда найдётся такое натуральное число p, что для любого слова $w \in L$ длины не меньше p найдутся слова u, v, x, y, $z \in \Sigma^*$, для которых верно $uvxyz = w$, $vy \neq \Lambda$ (то есть $v \neq \Lambda$ или $y \neq \Lambda$), $|vxy| \leq p$ и $uv^ixy^iz \in L$ для всех $i \in \mathbb{N}$.*

1.1 Задайте контекстно-свободной грамматикой а) язык $L_1 = \{ab^ncd^ne \mid n \geq 0\}$; b) L_2 — язык правильных скобочных последовательностей (пример: (()())()() — правильная, а (()()))(() — неправильная). Какие конструкции в русском языке устроены похожим образом с этими формальными языками?

1.2 Докажите, что язык $\{ww \mid w \in \{a, b\}^*\}$ нельзя задать контекстно-свободной грамматикой. Где в русском (или каком-нибудь еще) языке встречаются конструкции аналогичной структуры?

Пример. В голландском языке имеются конструкции с так называемыми cross-serial dependencies:

dat	Jan	Marie	Pieter	Arabisch	laat	zien	schrijven
что	Ян	Мария	Петер	арабский	предложить	посмотреть	писать

Перевод: "что Ян предложил Марии посмотреть на Петера, пишущего по-арабски".

2 Контекстно-зависимые грамматики

Определение 5. Контекстно-зависимая грамматика — грамматика, в которой все правила имеют вид $\eta A\theta \to \eta\alpha\theta$, где $A \in N$ — нетерминальный символ, $\eta, \theta \in (\Sigma \cup N)^*$ — строки из терминальных и нетерминальных символов (левый и правый контекст), а $\alpha \in (\Sigma \cup N)^+$ — непустая строка.

Определение 6. Неукорачивающая грамматика — грамматика, в которой все правила имеют вид $\alpha \to \beta$, где $|\alpha| \leq |\beta|$.

Теорема 2. *Любой язык, задаваемый неукорачивающей грамматикой, задается и контекстно-зависимой грамматикой, и наоборот.*

2.1 Докажите Теорему 2.

2.2 Задайте контекстно-зависимой грамматикой язык а) $\{ww \mid w \in \{a, b\}^*\}$; b) $\{a^nb^nc^n \mid n > 0\}$; c) $\{a^{2^n} \mid n \geq 0\}$.

3 Вершинные грамматики

Определение 7. Вершинная грамматика — структура $Gr = \langle N, \Sigma, P, S\rangle$, где N и Σ — алфавиты нетерминальных и терминальных символов, $S \in N$ — стартовый символ, а P — конечное множество правил одного из двух видов:

1. $A \to W(\sigma_1, \sigma_2)$;
2. $A \to C_{i,n}(\sigma_1, \ldots, \sigma_n)$, где $i \le n$.

Здесь $\sigma_i \in N \cup (\Sigma^* \times \Sigma^*)$. Пара строк (u, v) из $(\Sigma^* \times \Sigma^*)$ будет обозначаться нами как $u_\uparrow v$. Само множество $(\Sigma^* \times \Sigma^*)$ будет обозначаться как Π.

- Функция $W : \Pi^2 \to \Pi$ ("wrapping function", сворачивающая функция) действует следующим образом: $W(u_\uparrow v, x_\uparrow y) = ux_\uparrow yv$.
- Функция $C_{i,n} : \Pi^n \to \Pi$ ("concatenation", конкатенация) действует следующим образом:

$$C_{i,n}(u_{1\uparrow}v_1, \ldots, u_{n\uparrow}v_n) = u_1v_1 \ldots u_{i\uparrow}v_i \ldots u_nv_n.$$

Отношение $\Rightarrow^*$ определяется следующим образом:

- $\sigma \Rightarrow^* \sigma$ для всех $\sigma \in \Pi$.
- Пусть $A \to f(\sigma_1, \ldots, \sigma_n) \in P$ и пусть $\sigma_i \Rightarrow^* u_{i\uparrow}v_i \in \Pi$. Тогда $A \Rightarrow^* f(u_{1\uparrow}v_1, \ldots, u_{n\uparrow}v_n)$.

Наконец, дадим определение языка, задаваемого данной грамматикой: $L(Gr) := \{uv \mid S \Rightarrow^* u_\uparrow v\}$.

Замечание. Оригинальное определение вершинной грамматики немного иное (в нем строка не делится на две; вместо этого в каждой строке выделяется один символ, называемый вершиной). Несмотря на это, наборы задаваемых языков изначального и модифицированного формализмов совпадают, а данное определение проще.

3.1 Покажите, что любой контекстно-свободный язык задается некоторой вершинной грамматикой.

3.2 Задайте вершинной грамматикой язык а) $\{ww \mid w \in \{a, b\}^*\}$; b) $L_2 = \{a^n b^n c^n d^n \mid n \ge 0\}$.

3.3 Можно ли задать язык $\{a^{2^n} \mid n \ge 0\}$ вершинной грамматикой?

3.4 Покажите, что любой язык (без пустой строки), задаваемый вершинной грамматикой, задается некоторой контекстно-зависимой грамматикой.

Теорема 3. *Существует* полиномиальный *алгоритм (сложность $O(n^3)$, см. алгоритм Кока-Янгера-Касами), который по данным ему* контекстно-свободной грамматике Gr *и слову w возвращает 1, если $w \in L(Gr)$, и 0 иначе.*

Теорема 4. *Задача проверки по данным* контекстно-зависимой грамматике Gr *и слову w того, принадлежит ли $w \in L(Gr)$, является* PSPACE-*полной.*

Теорема 5. *Существует* полиномиальный *алгоритм, который по данным ему* вершинной грамматике Gr *и слову w возвращает 1, если $w \in L(Gr)$, и 0 иначе.*

Мораль: вершинные грамматики

1. задают контекстно-свободные языки,
2. задают интересный для лингвистов язык $\{ww \mid w \in \{a, b\}^*\}$,
3. могут быть относительно эффективно запрограммированы.