Формальні граматики

Формальна граматика або просто **граматика** в теорії формальних мов — спосіб опису формальної мови, тобто виділення деякої підмножини з множини всіх слів деякого скінченного алфавіту. Розрізняють *породжувальні* і *аналітичні* граматики — перші ставлять правила, за допомогою яких можна побудувати будь-яке слово мови, а другі дозволяють по даному слову визначити, входить воно в мову чи ні. Формальні граматики були введені американським вченим, математиком та філософом, Н. Хомським у 50-тих роках XX сторіччя.

Зміст

Поняття алфавіту

Алфавіт — скінченна множина символів. ε — порожній ланцюжок, слово, послідовність. Алфавітом є об'єднання алфавітів, перетин, різниця алфавітів. Нехай T — алфавіт, тоді:

- T^+ — множина усіх можливих послідовностей, що складені з елементів цього алфавіту крім порожньої послідовності ε.
- T* — множина усіх можливих послідовностей, що складені з елементів цього алфавіту, будь-якої довжини. Отже: $T^* = T^+ \cup \{\varepsilon\}$
- T^k — множина усіх можливих послідовностей,що складені з елементів цього алфавіту, довжини не більше k.

Мова в алфавіті T це множина ланцюжків скінченної довжини в цьому алфавіті. Зрозуміло, що кожна мова в алфавіті T є підмножиною множини T*.

Формальна граматика

Формальна граматика - це четвірка $G = \{N, T, P, S\}$. Де:

- **T** — алфавіт термінальних символів, терміналів (від англ. terminate - завершитись). Термінальні символи є алфавітом мови.
- **N** — алфавіт нетермінальних символів, нетерміналів. $T \cap N = \emptyset$; Нетермінали не входять в алфавіт мови.
- **S** — аксіома, спеціально виділений нетермінальний символ з якого починається опис граматики.

$$S \in N$$

- **P** — скінченна підмножина множини $(T \cup N)^+ \times (T \cup N)^*$. Інколи визначають так: $\alpha \in (T \cup N)^* \times N \times (T \cup N)^*, \beta \in (T \cup N)^*$.

Елемент (α,β) з множини P називається *правилом виводу* і записується у вигляді $\alpha \to \beta$. Таким чином, ліва частина правила не може бути порожньою. Правила з однаковою лівою частиною записують: $\alpha \to \beta_1 \mid \beta_2 \mid \ldots \mid \beta_n,\ \beta_i, \mathbf{i = 1,2,..,n}$ - називаються альтернативами правила виводу з ланцюжка α.

Виведення в формальних граматиках

Ланцюжок $\beta \in (T \cup N)^*$ назвемо **безпосередньо виведеним** з ланцюжка $\alpha \in (T \cup N)^+$ в граматиці $G = \{T, N, S, P\}$ (позначається $\alpha \Rightarrow \beta$), якщо $\alpha = \xi_1\gamma\xi_2, \beta = \xi_1\delta\xi_2 : \xi_1, \xi_2, \delta \in (T \cup N)^*, \gamma \in (T \cup N)^+, \gamma \to \delta \in P$.

Ланцюжок $\beta \in (T \cup N)^*$ назвемо **виведеним** з ланцюжка $\alpha \in (T \cup N)^+$ в граматиці $G = \{T, N, S, P\}$ (позначається $\alpha \Rightarrow^* \beta$), якщо

$$\exists \gamma_0, \gamma_1, \ldots, \gamma_n, n \geq 0 : \alpha = \gamma_0 \Rightarrow \gamma_1 \Rightarrow \ldots \Rightarrow \gamma_n = \beta$$

Термінальний рядок називається **правильним**(або сентенціальною формою) відносно граматики G якщо він виводиться з аксіоми цієї граматики.

Неоднозначності, та стратегії виводу

Граматика G називається **неоднозначною**, якщо існує декілька варіантів виводу слова ω в G (ω ∈ L(G)).

Приклад. Розглянемо таку граматику G = <N, Σ, P, S> зі схемою P.

```
S → S +S |S *S |a
```

Покажемо, що для ланцюжка ω = a + a + a існує щонайменше дв варіанти виводу:

1. `S ⇒ S + S ⇒ S + S + S ⇒ a + S + S ⇒ a + a + S ⇒ a + a + a`
2. `S ⇒ S + S ⇒ a + S ⇒ a + S + S ⇒ a + a + S ⇒ a + a + a`

В теорії граматик розглядається декілька стратегій виведення ланцюжка ω в G.

Лівостороння стратегія виводу ланцюжка ω в G — це послідовність кроків безпосереднього виводу, при якій на кожному кроку до уваги береться перший зліва направо нетермінал.

Правостороння стратегія виводу ω в G протилежна лівосторонній стратегії. З виводом ω в G пов'язане синтаксичне дерево, яке визначає синтаксичну структуру програми.

Формальні мови

Формальна мова породжена граматикою G - це множина термінальних рядків, що виводяться з аксіоми, тобто множина усіх правильних рядків.

$$L(G) = \{\alpha \in T^* \mid S \Rightarrow^* \alpha\}$$

З іншого боку, множина термінальних рядків, таких, що вони можуть бути виведені в граматиці G, називається мовою, що може бути розпізнана в даній граматиці, або допускається даною граматикою.

- Граматики G_1 та G_2 називаються **еквівалентними**, якщо $L(G_1) = L(G_2)$.
- Граматики G_1 та G_2 називаються **майже еквівалентними**, якщо $L(G_1) \cup \{\varepsilon\} = L(G_2) \cup \{\varepsilon\}$, тобто мови, ними породжувані, відрізняються не більш ніж на ε.

Граматика може бути **породжуючою** та **розпізнавальною**, це залежить від "напрямку" застосування правил. Для породжуючих граматик виведення починається з аксіоми і закінчується термінальним рядком (рядком, що складається тільки з терміналів). А розпізнавальні аналізують вхідний термінальний рядок на правильність, чиможна такий рядок вивести в цій граматиці. Коротко кажучи - породжуючі це "згори вниз", а розпізнавальні - "знизу вгору". Остання задача є практичнішою і гострою.

Класифікація граматик

Хомський також запропонував класифікацію граматик. Він виділив чотири типи граматик.

Класифікація за Хомським

Тип 0. Граматики Типу 0
: це найзагальніший клас граматик. На правила не накладаються ніяких обмежень, окрім зазначених у визначенні. Вони еквівалентні Машині Тюринга. Доведено існування мов, які породжуються граматиками типу 0, але не граматиками типу 1, але невідомі прості приклади таких мов.

Тип 1. Граматики Типу 1
: нескорочуюча або контекстно-залежна(КЗ) граматика. Вибір означення не впливає на множину мов, породжуваних граматиками цього класу, оскільки доведено, що множина мов, породжуваних граматиками, що не укорочують, збігається із множиною мов, породжуваних КЗ-граматиками.

- *Нескорочуючі граматики.* Всі правила мають вигляд

  $\alpha \to \beta$
  $\alpha \in (T \cup N)^+, \beta \in (T \cup N)^*$
  $|\alpha| \leq |\beta|$
  де $|\alpha|$ означає кількість символів у рядку α.

- *Контекстно-залежні граматики.* Всі правила мають вигляд:

  $\alpha \to \beta$
  $\alpha = \xi_1 A \xi_2, \beta = \xi_1 \gamma \xi_2 : A \in N, \gamma \in (T \cup N)^+, \xi_1, \xi_2 \in (T \cup N)^*$

Тип 2. Граматики Типу 2. Контекстно-вільні граматики(КВ)
: Всі правила мають вигляд
  $\alpha \to \beta$
  $\alpha \in N, \beta \in (T \cup N)^*$
  Тобто в усіх правилах цього виду зліва стоїть тільки один нетермінал. Тому вони і контекстно вільні, бо на використання правила для даного нетерміналу не впливають символи, що оточують його. Ці символи називають **контекстом**.

Тип 3. Граматики Типу 3. Регулярні граматики. А-граматики
: можна визначити як праволінійну або ліволінійну граматику, або як змішану. Також ці мови називають скінченно-автоматними, бо вони еквівалентні скінченним автоматам, тобто клас мов, що породжуються граматиками типу 3, збігається з класом мов, які розпізнаються скінченими автоматами. Також ці граматики є основою регулярних виразів.

- Праволінійна граматика. Всі правила мають вигляд:

  $\alpha \to \beta$
  $\alpha \to \omega\beta, \beta \in N, \omega \in T^*$
  або

$\alpha \to \omega, \omega \in T^*$

- Ліволінійна граматика. Всі правила мають вигляд:

$\alpha \to \beta$
$\alpha \to \beta\omega, \beta \in N, \omega \in T^*$
або
$\alpha \to \omega, \omega \in T^*$

Доведено, що праволінійні та ліволінійні граматики еквівалентні, і існує алгоритм для переведення правил граматики одного типу в інший тип.

Співвідношення між типами граматик

1. Будь-які граматики типу 3 є граматиками типу 2.
2. Будь-які КВ-граматики є граматиками типу 1 (КЗ або неукорочуючі), з точністю до ε.
3. Будь-які граматики типу 1 є граматиками типу 0.

Мова L(G) є *мовою типу k*, якщо її можна описати граматикою типу *k*.

Співвідношення між типами мов

1. кожна регулярна мова є КВ-мовою, але існують КВ- мови, що не є регулярними (наприклад, $L = \{a^n b^n \mid n > 0\}$);
2. кожна КВ- мова є КЗ- мовою, але існують КЗ-мови, що не є КВ-мовами (наприклад, $L = \{a^n b^n c^n \mid n > 0\}$);
3. кожна КЗ-мова є мовою типу 0.

Варто підкреслити, що якщо мова задана граматикою типу k, то це не означає, що не існує граматики типу k' (k'>k), що описує ту ж мову. Тому, коли говорять про мову типу k, звичайно мають на увазі максимально можливе k.

Використання формальних граматик

Формальні граматики це дуже потужній математичний інструмент, який використовується в математичній та комп'ютерній лінгвістиці, описі мов програмування, розробці компіляторів в теорії програмування. Найбільш вживаними є КВ-граматики і регулярні, бо їх найлегше описати алгоритмічно.

Сама по собі концепція формальних граматик доволі гнучка, тому не дивно, що з'явилося багато інших інструментів, що розширюють використання та потужність КВ-граматик і граматик третього типу. Наприклад, атрибутні граматики, LL-k граматики, скінченні автомати, регулярні вирази та множини.

Див. також

- Нотація Бекуса-Наура
- Синтаксичний аналіз
- Ієрархія Чомскі

Посилання

- *Серебряков В.А., Галочкин М.П., Гончар Д.Р., Фуругян М.Г.* Теория и реализация языков программирования // М.: МЗ-Пресс, 2006 г, 2-е изд. ISBN 94073-094-9. (рос.)
- Подборка ссылок на литературу по направлению. (рос.)