Лекція 11-12.

ЙМОВІРНІСНІ ГРАМАТИКИ

Регулярная (автоматная) грамматика может быть задана набором правил как **левая** или **правая** регулярная грамматика.
правая регулярная грамматика - все правила могут быть в одной из следующих форм:

1. $A \to a$
2. $A \to aB$
3. $A \to \varepsilon$

левая регулярная грамматика - все правила могут быть в одной из следующих форм:

1. $A \to a$
2. $A \to Ba$
3. $A \to \varepsilon$

где

- заглавные буквы (A, B) обозначают нетерминалы из множества N
- строчные буквы (a, b) обозначают терминалы из множества Σ
- ε - пустая строка, т.е. строка длины 0

Для каждой цепочки x из языка L вероятность $p(x)$ определяется так, что $\sum_{x \in L} p(x) = 1$ и $0 < p(x) \leqslant 1$. Значит, функция $p(x)$ есть плотность вероятности, заданная на L, и ее можно использовать для оценки неопределенности и случайности в L. Кроме того, эвристически построенная грамматика может порождать «нежелательные» цепочки, т. е. цепочки, которые не представляют никаких объектов в заданном классе. В этом случае для эффективного использования грамматики нежелательным цепочкам можно приписать очень малые вероятности.

Есть два естественных способа расширения понятия формального языка до стохастического языка. Они состоят в рандомизации выбора правила подстановки в грамматике либо выбора следующего состояния в распознающем автомате.

Определение 5.1. Стохастической порождающей грамматикой (или просто стохастической грамматикой) называется четверка

$G_s = (V_N, V_T, P_s, S)$, где V_T и V_N — конечные множества основных и вспомогательных символов; $S \in V_N$ — начальный символ [1]); P_s — конечное множество стохастических правил подстановки, каждое из которых имеет вид

$$\alpha_i \xrightarrow{p_{ij}} \beta_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k, \tag{5.1}$$

где

$$\alpha_i \in (V_N \cup V_T)^* V_N (V_N \cup V_T)^*, \quad \beta_{ij} \in (V_N \cup V_T)^*,$$

а p_{ij} — вероятность, связанная с применением этого правила подстановки и удовлетворяющая условиям [2])

$$0 < p_{ij} \leqslant 1 \text{ и } \sum_{j=1}^{n_i} p_{ij} = 1. \tag{5.2}$$

Предположим, что правило $\alpha_i \overset{p_{ij}}{\longrightarrow} \beta_{ij}$ принадлежит P_S. Тогда цепочка $\xi = \gamma_1\alpha_1\gamma_2$ может быть заменена на $\eta = \gamma_1\beta_{ij}\gamma_2$ с вероятностью p_{ij}. Будем записывать такой вывод как

$$\xi \overset{p_{ij}}{\Rightarrow} \eta$$

и будем говорить, что цепочка ξ прямо порождает η с вероятностью p_{ij}. Если существует последовательность цепочек $\omega_1, \ldots, \omega_{n+1}$, такая, что

$$\xi = \omega_1,\ \eta = \omega_{n+1},\ \omega_i \overset{p_i}{\Rightarrow} \omega_{i+1},\quad i = 1, \ldots, n,$$

то будем говорить, что цепочка ξ порождает η с вероятностью $p = \prod_{i=1}^{n} p_i$, и будем обозначать этот вывод как

$$\xi \overset{p}{\underset{*}{\Rightarrow}} \eta.$$

Вероятность этого вывода равна произведению вероятностей стохастических правил подстановки, используемых в выводе. Ясно, что $\overset{p}{\underset{*}{\Rightarrow}}$ есть рефлексивное и транзитивное замыкание отношения $\overset{p}{\Rightarrow}$.

Стохастический язык, порождаемый грамматикой G_s, есть

$$L(G_s) = \{(x,\ p(x)) \mid x \in V_T^*,\ S \overset{p_j}{\underset{*}{\Rightarrow}} x,$$

$$j = 1, \ldots, k \text{ и } p(x) = \sum_{j=1}^{k} p_j\}, \qquad (5.3)$$

где k — число различных выводов [1]) цепочки x из S, а p_j — вероятность j-го вывода x.

Поскольку у стохастических грамматик вид правил подстановки такой же, как и у обычных порождающих грамматик, за исключением введенного распределения вероятностей, семейство языков, порождаемых стохастическими грамматиками, совпадает с семейством языков, порождаемых обычными грамматиками.

В общем случае стохастический язык $L(G_s)$ определяется парой (L, p), где L — формальный язык, а p — определенное на L распределение вероятностей. Язык L стохастического языка $L(G_s) = (L, p)$ называется характеристическим языком $L(G_s)$.

Определение 5.2. Для любой стохастической грамматики G_s соответствующая характеристическая грамматика, обозначаемая $\overline{G}_s$, получается исключением вероятностей применения из стохастических правил подстановки, принадлежащих G_s. Если $P_s = (P, D)$, где D — вероятностная мера на множестве правил подстановки P, то очевидно, что $\overline{G}_s = (V_N, V_T, P, S)$ — нестохастическая грамматика.

Используя определение характеристической грамматики, мы будем называть стохастическую грамматику G_s стохастической грамматикой типа 0 (без ограничений на правила вывода), или типа 1 (непосредственно составляющих), или типа 2 (бесконтекстной), или типа 3 (автоматной) тогда и только тогда, когда $\overline{G}_s$ является соответственно грамматикой типа 0, 1, 2 или 3. Язык, порождаемый грамматикой $\overline{G}_s$ и обозначаемый $L(\overline{G}_s)$, есть в точности характеристический язык языка $L(G_s)$. В общем случае стохастическая грамматика G_s эквивалентна некоторой стохастической грамматике G'_s тогда и только тогда, когда $L(G_s) = L(G'_s)$.

Теорема 5.1. I. *Если* L_s *— стохастический автоматный язык, то он одновременно и стохастический бесконтекстный язык.*

II. *Если* L_s *— λ-свободный стохастический бесконтекстный язык* [2]*, то* L_s *есть стохастический язык непосредственно составляющих.*

III. *Если* L_s *— стохастический язык непосредственно составляющих, то* L_s *есть стохастический язык типа* 0.

Теорема 5.2. (Нормальная форма Хомского). *Любая стохастическая бесконтекстная грамматика эквивалентна некоторой стохастической бесконтекстной грамматике* G_s*, в которой все правила подстановки имеют вид* $A \xrightarrow{p} BC$ *или* $A \xrightarrow{p} a$*, где* A*,* B *и* C *— вспомогательные символы, а* a *— основной символ.*

Теорема 5.3. (Нормальная форма Грейбах). *Любая стохастическая бесконтекстная грамматика эквивалентна некоторой стохастической бесконтекстной грамматике* G_s*, у которой все правила подстановки имеют вид* $A \xrightarrow{p} a\gamma$*, где* A *— вспомогательный символ,* a *— основной символ, а* γ *— цепочка вспомогательных символов.*

Пример 5.1. Рассмотрим автоматную грамматику $G_s = (V_N, V_T, P_s, S)$, где $V_N = \{A_1, \ldots, A_k\}$, $V_T = \{a_n, \ldots, a_m\}$, $S = A_1$ и P_s для $i = 1, 2, \ldots, k$ состоит из правил

$$A_i \xrightarrow{p_{ij}^1} a_1 A_1, \qquad A_i \xrightarrow{p_{i0}^1} a_1,$$
$$\vdots \qquad\qquad \vdots$$
$$A_i \xrightarrow{p_{i1}^2} a_2 A_1, \qquad A_i \xrightarrow{p_{i0}^l} a_l,$$
$$\vdots \qquad\qquad \vdots$$
$$A_i \xrightarrow{p_{ij}^l} a_l A_j, \qquad A_i \xrightarrow{p_{i0}^m} a_m,$$
$$\vdots$$
$$A_i \xrightarrow{p_{ik}^m} a_m A_k.$$

Некоторые из вероятностей p_{ij}^{l}, p_{i0}^{l} могут быть равны нулю. Далее,

$$\sum_{j=1}^{k}\sum_{l=1}^{m} p_{ij}^{l} + \sum_{l=1}^{m} p_{i0}^{l} = 1.$$

В частности, пусть $V_N = \{S, A, B\}$, $V_T = \{0, 1\}$ и P_s:

$$S \xrightarrow{1} 1A, \qquad B \xrightarrow{0,3} 0,$$

$$A \xrightarrow{0,8} 0B, \qquad B \xrightarrow{0,7} 1S,$$

$$A \xrightarrow{0,2} 1.$$

Типичным выводом будет, например, вывод

$$S \Rightarrow 1A \Rightarrow 10B \Rightarrow 100, \qquad p\,(100) = 1 \times 0{,}8 \times 0{,}3 = 0{,}24.$$

Грамматика G_s порождает следующий стохастический язык $L\,(G_s)$:

Порожденная цепочка x	$p\,(x)$
11	0,2
100	0,24
$(101)^n11$	$0{,}2 \times (0{,}56)^n$
$(101)^n100$	$0{,}24 \times (0{,}56)^n$

Заметим, что

$$\sum_{x \in L(G_s)} p\,(x) = 0{,}2 + 0{,}24 + \sum_{n=1}^{\infty} (0{,}2 + 0{,}24)\,(0{,}56)^n = 1.$$

Определение 5.3. Если в стохастической грамматике G_s

$$\sum_{x \in L(G_s)} p(x) = 1, \tag{5.4}$$

то говорят, что G_s — согласованная стохастическая грамматика.

В этом разделе приводятся условия, необходимые для того, чтобы стохастические линейные и стохастические бесконтекстные грамматики были согласованы [2, 6, 10]. Является ли согласованной при каких-либо условиях стохастическая грамматика непосредственно составляющих, до сих пор неизвестно. Каждая стохастическая грамматика типа 3 согласована.

Пусть G_s — стохастическая линейная грамматика с $V_N = \{A_1, \ldots, A_k\}$, $S = A_1$, а $\overline{G_s} = (V_N, V_T, P, S)$. Для каждого $A_i \in V_N$ определим класс эквивалентности C_{A_i} как

C_{A_i} = {все правила подстановки из P с посылкой (левой частью) A_i}.

Таким образом, множество правил подстановки

$$P = \bigcup_{i=1}^{k} C_{A_i}.$$

Пусть p_{ij} — вероятность, связанная с каждым правилом подстановки вида $A_i \to uA_jv$, $u,v \in V_T$, p_{ir} — вероятность, связанная с правилом подстановки вида $A_i \to u_r$, $u_r \in V_T$. Тогда $\sum_j p_{ij} + \sum_r p_{ir} = 1$. Процесс порождения может быть представлен как марковский процесс с $k+1$ состояниями, которые соответствуют множеству $V_N + \{T\}$. Поглощающим состоянием будет T, а все остальные состояния, если G_s согласована, будут переходными.

С каждой линейной грамматикой можно связать некоторый граф. Вершины этого графа соответствуют элементам V_N, а одна из них — дополнительная вершина — соответствует T. Вершины A_i и A_j связаны дугой тогда и только тогда, когда в множестве P существует правило подстановки вида $A_i \to uA_jv$. С другой стороны, дуга между A_i и T существует тогда и только тогда, когда в P есть правило вида $A_i \to p$. Начальная вершина — это вершина, соответствующая $S = A_i$. Для любой цепочки $x \in L(\overline{G}_s)$ в графе существует путь из начальной вершины в поглощающую вершину T, соответствующий некоторому выводу $S \overset{*}{\Rightarrow} x$. Заметим, что если G_s согласована, то вероятность достижения вершины T из начальной вершины должна быть равна 1. В противном случае G_s не согласована.

Чтобы проверить, согласована ли грамматика G_s, определим матрицу переходов марковского процесса $(k+1) \times (k+1)$ следующим образом:

$$M = [m_{ij}],$$

где

$$m_{ij} = \begin{cases} P_{ij}, & 1 \leqslant i \leqslant k, \quad 1 \leqslant j \leqslant k+1, \\ 0, & i = k+1, \quad 1 \leqslant j \leqslant k, \\ 1, & i = k+1, \quad j = k+1. \end{cases}$$

Каждый элемент m_{ij} представляет вероятность одношагового вывода или применения одного правила в выводе цепочки в $L(G_s)$. Элемент $m_{1,\,k+1}$ представляет вероятность всех цепочек из $L(G_s)$, вывод которых требует применения лишь одного правила подстановки. Пусть

$$M^n = \underbrace{M \times M \times \ldots \times M}_{n \text{ раз}} = [m_{ij}(n)].$$

Элемент $m_{ij}(n)$ представляет вероятность всех цепочек из $L(G_s)$, вывод которых требует применения не больше чем n правил подстановки. Мы приходим к следующей теореме.

Теорема 5.4. *Пусть G_s — стохастическая линейная грамматика. Грамматика G_s согласована тогда и только тогда, когда*

$$\lim_{n\to\infty} m_{1,\,k+1}(n) = 1. \tag{5.5}$$

Пример 5.2. Дана стохастическая линейная грамматика $G_s = (V_N, V_T, P_s, S)$, где $V_N = \{S, A, B\}$, $V_T = \{0, 1\}$ и P_s:

$$S \xrightarrow{p_1} 0A1, \qquad A \xrightarrow{1-p_2} 00,$$

$$S \xrightarrow{1-p_1} 1B1, \qquad B \xrightarrow{p_3} 1A1,$$

$$S \xrightarrow{p_2} 0A0, \qquad B \xrightarrow{1-p_3} 1.$$

Тогда

$$M = \begin{array}{c} \\ S \\ A \\ B \\ T \end{array} \begin{array}{c} \begin{array}{cccc} S & A & B & T \end{array} \\ \begin{bmatrix} 0 & p_1 & (1-p_1) & 0 \\ 0 & p_2 & 0 & (1-p_2) \\ 0 & p_3 & 0 & (1-p_3) \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Пусть

$$M_1 = \begin{bmatrix} 0 & 0{,}4 & 0{,}6 & 0 \\ 0 & 0{,}5 & 0 & 0{,}5 \\ 0 & 0{,}6 & 0 & 0{,}4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Тогда

$$\lim_{n\to\infty} M_1^n = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Таким образом, грамматика G_s с $p_1=0{,}4$, $p_2=0{,}5$ и $p_3=0{,}6$ согласована, поскольку

$$\lim_{n\to\infty} m_{1.4}(n)=1.$$

Если допустить, что

$$M_2=\begin{bmatrix} 0 & 0{,}4 & 0{,}6 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0{,}6 & 0 & 0{,}4 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

то

$$\lim_{n\to\infty} M_2^n = \begin{bmatrix} 0 & 0{,}76 & 0 & 0{,}24 \\ 0 & 1 & 0 & 0 \\ 0 & 0{,}6 & 0 & 0{,}4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Отсюда грамматика G_s с $p_1 = 0{,}4$, $p_2 = 1$ и $p_3 = 0{,}6$ не согласована, поскольку

$$\lim_{n\to\infty} m_{1,4}(n) = 0{,}24 \neq 1.$$

В бесконтекстных грамматиках все правила подстановки имеют вид

$$A \to \alpha, \quad A \in V_N, \quad \alpha \in V^+.$$

В этом случае вспомогательный символ в левой части правила может порождать некоторое конечное число, возможно 0, вспомогательных символов. Для изучения порождения языка в стохастической бесконтекстной грамматике можно применить теорию ветвящихся процессов Гальтона — Уотсона [11, 12]. Нулевой уровень процесса порождения соответствует начальному символу S.

В качестве первого уровня примем β_1, где β_1 — цепочка, порожденная правилом подстановки $S \to \beta_1$. Второй уровень соответствует цепочке β_2, которая получается из β_1 в результате применения подходящих правил подстановки к каждому вспомогательному символу β_1. Если β_1 не содержит ни одного вспомогательного символа, то процесс заканчивается. При помощи этого процесса мы можем определить цепочку j-го уровня β_j как цепочку, полученную из цепочки β_{j-1} при помощи применения соответствующих правил подстановки ко всем вспомогательным символам β_{j-1}. Поскольку при переходе от $(j-1)$-го уровня к j-му все вспомогательные символы рассматриваются одновременно, достаточно учесть только вероятности p_{ir} всех правил $A_i \to \alpha_r$. Значения этих вероятностей обрабатываются так же, как это делалось для стохастических линейных грамматик. Итак, пусть

$$P = \bigcup_{i=1}^{k} C_{A_i}.$$

Для каждого класса эквивалентности C_{A_i}

$$\sum_{C_{A_i}} p_{ir} = 1.$$

Определение 5.4. Для каждого класса C_{A_i}, $i = 1, \ldots, k$, определим k-местную производящую функцию следующим образом:

$$f_i(s_1, \ldots, s_k) = \sum_{C_{A_i}} p_{ir} s_1^{\mu_{i1}(\alpha_2)} s_2^{\mu_{i2}(\alpha_2)} \ldots s_k^{\mu_{ik}(\alpha_2)}, \qquad (5.6)$$

где $\mu_{il}(\alpha_r)$ обозначает число вхождений вспомогательного символа A_l в цепочку α_r правила $A_i \to \alpha_r$, а $S = A_1$.

Определение 5.5. Производящая функция j-го уровня определяется рекурсивно следующим образом:

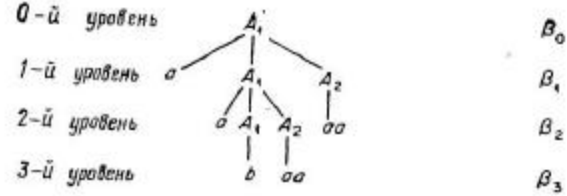
$$\begin{aligned} &F(s_1, \ldots, s_k) = s_1, \\ &F_1(s_1, \ldots, s_k) = f_1(s_1, \ldots, s_k), \\ &F_j(s_1, \ldots, s_k) = F_{j-1}(f_1(s_1, \ldots, s_k), \ldots, f_k(s_1, \ldots, s_k)). \end{aligned} \qquad (5.7)$$

Пример 5.3. Грамматика $G_s = (V_N, V_T, P_s, S)$, где $V_N = \{A_1, A_2\}$, $V_T = \{a, b\}$, $S = A_1$ и P_s:

$$A_1 \xrightarrow{p_{11}} aA_1A_2, \qquad A_2 \xrightarrow{p_{21}} aA_2A_2,$$

$$A_1 \xrightarrow{p_{12}} b, \qquad A_2 \xrightarrow{p_{22}} aa.$$

Здесь процесс порождения цепочки $aabaaaa$, записанный по уровням, выглядит так:



Производящие функции для C_{A_1} и C_{A_2} есть соответственно

$$f_1(s_1, s_2) = p_{11}s_1s_2 + p_{12}, \quad f_2(s_1, s_2) = p_{21}s_2^2 + p_{22}.$$

Производящие функции j-го уровня при $j = 0, 1, 2$ таковы:

$$F_0(s_1, s_2) = s_1,$$

$$F_1(s_1, s_2) = f_1(s_1, s_2) = p_{11}s_1s_2 + p_{12},$$

$$\begin{aligned} F_2(s_1, s_2) &= F_1(f_1(s_1, s_2), f_2(s_1, s_2)) = \\ &= p_{11}f_1(s_1, s_2) f_2(s_1, s_2) + p_{12} = \\ &= p_{11}^2 p_{21}s_1s_2^3 + p_{11}^2 p_{22}s_1s_2 + p_{11}p_{12}p_{21}s_2^2 + p_{11}p_{12}p_{22} + p_{12}. \end{aligned}$$

Рассмотрев предыдущий пример, мы можем выразить $F_j(s_1, \ldots, s_k)$ в виде

$$F_j(s_1, \ldots, s_k) = G_j(s_1, \ldots, s_k) + K_j,$$

где $G_j(s_1, \ldots, s_k)$ обозначает полином от $s_1, \ldots, s_k$ без постоянного члена. Постоянный член K_j соответствует вероятности всех цепочек, которые могут быть выведены не дальше j-го уровня.

Мы приходим к следующей теореме:

Теорема 5.5. *Стохастическая бесконтекстная грамматика согласована тогда и только тогда, когда*

$$\lim_{j \to \infty} K_j = 1. \tag{5.8}$$

Заметим, что если вышеупомянутый предел не равен 1, то существует ненулевая вероятность того, что процесс порождения никогда не закончится. Отсюда следует, что вероятностная мера, определенная на всем L, меньше 1, и, следовательно, грамматика G_s не согласована. С другой стороны, если предел равен 1, то не существует вывода, который нельзя было бы закончить на некотором шаге, поскольку этот предел представляет вероятность появления всех цепочек, для порождения которых требуется конечное число применений правил. Следовательно, грамматика G_s согласована. Проблема проверки согласованности данной стохастической бесконтекстной грамматики может быть решена с помощью следующей процедуры проверки, разработанной для ветвящихся процессов.

Определение 5.6. Математическое ожидание числа появлений вспомогательного символа A_j в множестве правил подстановки C_{A_i} есть

$$e_{ij} = \frac{\partial f_i(s_1, \ldots, s_k)}{\partial s_j}\Big|_{s_1, \ldots, s_k = 1}. \tag{5.9}$$

Определение 5.7. Матрица первых моментов E порождающегося процесса, соответствующего стохастической бесконтекстной грамматике G_s, определяется как

$$E = [e_{ij}], \quad 1 \leqslant i, \quad j \leqslant k, \tag{5.10}$$

где k есть число вспомогательных символов в G_s.

Теорема 5.6. *Если для данной стохастической бесконтекстной грамматики G_s собственные значения матрицы первых моментов или корни ее характеристического уравнения $\rho_1, \ldots, \rho_k$ расположены в порядке убывания их абсолютных величин, т. е.*

$$|\rho_i| \geqslant |\rho_j|, \quad \text{если} \quad i < j, \tag{5.11}$$

то грамматика G_s согласована при $\rho_1 < 1$ и не согласована при $\rho_1 > 1$.

Пример 5.4. Для стохастической бесконтекстной грамматики из примера 5.3

$$e_{11} = \frac{\partial f_1(s_1, s_2)}{\partial s_1}\Big|_{s_1, s_2=1} = p_{11},$$

$$e_{12} = \frac{\partial f_1(s_1, s_2)}{\partial s_2}\Big|_{s_1, s_2=1} = p_{11},$$

$$e_{21} = \frac{\partial f_2(s_1, s_2)}{\partial s_1}\Big|_{s_1, s_2=1} = 0,$$

$$e_{22} = \frac{\partial f_2(s_1, s_2)}{\partial s_2}\Big|_{s_1, s_2=1} = 2p_{21}.$$

Таким образом,

$$E = \begin{bmatrix} p_{11} & p_{11} \\ 0 & 2p_{21} \end{bmatrix}.$$

Характеристическое уравнение для матрицы E есть

$$\Phi(x) = (x - p_{11})(x - 2p_{21}).$$

Следовательно, грамматика G_s будет согласованной до тех пор, пока $p_{21} < 1/2$.