В работе исследуются деревья вывода высоты t, порождаемые согласованной стохастической КС-грамматикой, при $t \to \infty$.

Стохастической КС-грамматикой [3] называется четвёрка $G = \langle V_N, V_T, R, s \rangle$, где V_N и V_T — конечные алфавиты нетерминальных и терминальных символов (терминалов и нетерминалов), $s \in V_N$ — аксиома, $R = \cup_{i=1}^{n} R_i$, где $n = |V_N|$ и R_i — конечное множество правил вывода r_{ij} вида:

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij} (j = 1, 2, \ldots, n_i),$$

где $A_i \in V_N$, $\beta_{ij} \in (V_N \cup V_T)^*$, и p_{ij} — вероятность применения правила r_{ij}, причём $0 < p_{ij} \leq 1$ и $sum_{j=1}^{n} p_{ij} = 1$.

Применение правила r_{ij} грамматики к слову $\alpha \in (V_N \cup V_T)^*$ состоит в замене какого-либо вхождения нетерминала A_i в α на слово β_{ij}. Язык L_G, порождаемый грамматикой G, содержит все слова из алфавита V_T, которые можно получить из аксиомы s последовательным применением правил вывода.

Каждому слову α из L_G соответствует последовательность $\omega(\alpha) = (r_1, r_2, \ldots, r_k)$ правил вывода, с помощью последовательного применения которых α можно получить из аксиомы s. Такая последовательность правил называется выводом слова α. Выводу слова соответствует дерево вывода [1] d, вероятность $p(d)$ которого определяется как произведение вероятностей правил, образующих вывод: $p(d) = \prod_{i=1}^{k} p(r_i)$. Одному и тому же слову $\alpha \in L_G$ может соответствовать более одного дерева вывода. Вероятность слова $\alpha \in L_G$ определяется как сумма вероятностей всех порождающих его деревьев.

Грамматика называется согласованной, если сумма вероятностей всех конечных деревьев вывода равна 1. Согласованная стохастическая грамматика G задаёт распределение вероятностей на множестве слов порождаемого ею языка L_G. В работе рассматриваются согласованные грамматики.

По стохастической КС-грамматике строится матрица A первых моментов. Её элемент a_j^i определяется как $\sum_{k=1}^{n_i} p_{ik} s_{ik}^j$, где величина s_{ik}^j равна числу нетерминальных символов A_j в правой части правила вывода r_{ik}. Перронов корень [2] матрицы A обозначим через r. Известно, что для согласованной грамматики $r \leq 1$.

Будем обозначать $A_i \to A_j$, если в грамматике имеется правило вывода вида $A_i \xrightarrow{p_{ij}} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_N \cup V_T)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$. Будем обозначать $A_i \leftrightarrow_* A_j$, если одновременно $A_i \to_* A_j$ и $A_j \to_* A_i$. Множество V_N нетерминалов разбивается на классы эквивалентности $K_1, K_2, \ldots, K_m$ по отношению $\leftrightarrow_*$. Будем обозначать $K_i \prec K_j$, если существуют $A_{k_i} \in K_i$ и $A_{k_j} \in K_j$, такие что $A_{k_i} \to A_{k_j}$. Рефлексивное замыкание $\prec$ обозначим $\prec_*$.

Случай $r < 1$ (докритический случай) рассматривался Л.П. Жильцовой в [4] и других работах. А.Е. Борисов обобщил [5] полученные результаты на случай $r \leq 1$ для разложимой грамматики, содержащей два класса нетерминалов.

Пусть классы нетерминалов пронумерованы таким образом, что $i \leq j$ для любых $K_i \prec_* K_j$. Матрица A первых моментов грамматики в этом

случае имеет вид:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1,m} \\ 0 & A_{22} & \cdots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{m,m} \end{pmatrix}.$$

Для каждого класса K_i матрица A_{ij} неразложима. Обозначим через r_i перронов корень матрицы A_{ii}. Очевидно, $r = \max_i\{r_i\}$. Обозначим $J = \{i : r_i = r\}$.

Для пары классов K_i и K_j рассмотрим всевозможные цепочки $K_{i_1} \prec K_{i_2} \prec \ldots \prec K_{i_k}$, где $i_1 = i$ и $i_k = j$. Обозначим через s_{ij} максимальное число классов с номерами из J в такой цепочке. Будем также обозначать $s_i = \max_j\{s_{ij}\}$.

Теорема 1 *Пусть матрица A первых моментов стохастической КС-грамматики G имеет перронов корень, не превосходящий 1. Тогда математическое ожидание числа применений правила r_{ij} в случайном дереве вывода высоты t при $t \to \infty$ имеет следующий вид:*

$$M_{ij}(t) \sim d_i \cdot p_{ij} \cdot t^{\left(\frac{1}{2}\right)^{s_1 - s_{1l} - 1}},$$

где p_{ij} — вероятность правила r_{ij}, d_i — некоторая константа, и $A_i \in K_l$.

Теорема 2

$$M_i(t) \sim d_i \cdot t^{\left(\frac{1}{2}\right)^{s_1 - s_{1l} - 1}}$$

Обозначим через $q_i(t)$ число нетерминалов A_i в случайном дереве вывода высоты t, порождённом грамматикой.

Теорема 3 *Для любой пары нетерминалов $A_i \in K_h$, $A_j \in K_l$, такой что $s_{1h} = s_{1l}$, при $t \to \infty$ выполняется условие:*

$$D\left(\frac{q_i(t)}{q_j(t)} - \frac{d_i}{d_j}\right) \to 0,$$

где $q_i(t)$, $q_j(t)$ — число нетерминалов A_i и A_j в случайном дереве вывода высоты t, d_i и d_j — некоторые константы.

Таким образом, соотношение числа нетерминалов в деревьях вывода высоты t становится всё ближе к фиксированному значению при $t \to \infty$.

Список литературы

[1] Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции — М.: МИР, 1978

[2] Гантмахер Ф.Р. Теория матриц. — М.: ФИЗМАТЛИТ, 2010

[3] К. Фу. Структурные методы в распознавании образов. М.: МИР, 1977

[4] Жильцова Л.П. Закономерности применения правил грамматики в выводах слов стохастического контекстно-свободного языка // Математические вопросы кибернетики. Вып. 9. М.: Наука, 2000. С. 101-126

[5] Борисов А.Е. Закономерности в словах стохастических контекстно-свободных языков, порождённых грамматиками с двумя классами нетерминальных символов. Вопросы экономного кодирования.