1 Предварительные сведения

Рассматривается множество D^t деревьев вывода высоты t. Известны асимптотические выражения для вероятностей таких деревьев, а также математические ожидания числа применений правила r_{ij} в них ($M_{ij}(t)$):

$$\begin{aligned} P_j(t) &= U^{(j)} \frac{c_j}{t^{1+\left(\frac{1}{2}\right)^{q_j-1}}} \cdot (1+o(1)) \\ M_{ij}(t) &= d_i p_{ij} t^{\left(\frac{1}{2}\right)^{q_l^*-1}} \cdot (1+o(1)) \end{aligned}, \tag{1}$$

где $c_j, d_i > 0$ и

$$q_l^* = \begin{cases} q_l - 1, & \text{класс } K_l \text{ — критический} \\ q_l, & \text{класс } K_l \text{ — некритический} \end{cases} \tag{2}$$

2 Энтропия

Пусть L^t — множество слов языка L_G, которым соответствуют деревья вывода из D^t. Будем рассматривать грамматики с однозначным выводом, то есть, положим что каждому слову из L^t соответствует единственное дерево вывода из D^t.

По определению, энтропия языка L^t есть

$$H(L^t) = -\sum_{\alpha \in L^t} p_t(\alpha) \log p_t(\alpha), \tag{3}$$

где $p_t(\alpha) = p(\alpha | \alpha \in L^t) = \frac{p(\alpha)}{P(L^t)}$. Используя это выражение для $p_t(\alpha)$, получаем:

$$\begin{aligned} H(L^t) = -\frac{1}{P(L^t)} \sum_{\alpha \in L^t} p(\alpha) \left(\log p(\alpha) - \log P(L^t) \right) = \\ = \frac{\log P(L^t)}{P(L^t)} \sum_{\alpha \in L^t} p(\alpha) - \frac{1}{P(L^t)} \sum_{\alpha \in L^t} p(\alpha) \log p(\alpha) = \\ = \log P(L^t) - \frac{1}{P(L^t)} \sum_{\alpha \in L^t} p(\alpha) \log p(\alpha). \end{aligned} \tag{4}$$

Выразим вероятность слова α через вероятности правил вывода r_{ij}. Будем рассматривать грамматику с однозначным выводом и считать, что каждому слову α из L^t соответствует единственное дерево $d(\alpha)$ из D^t, и, следовательно, единственный левый вывод $\omega(\alpha) = r_1 \cdot r_2 \cdot \ldots \cdot r_\mu$. Получаем:

$$p(\alpha) = p(r_1) \cdot p(r_2) \cdot \ldots \cdot p(r_\mu) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} p_{ij}^{q_{ij}(\alpha)}, \tag{5}$$

где $q_{ij}(\alpha)$ — число применений правила r_{ij} при выводе слова α (в грамматике с однозначным выводом это число определяется единственным образом). Тогда:

$$\sum_{\alpha \in L^t} p(\alpha) \log p(\alpha) = \sum_{\alpha \in L^t} p(\alpha) \sum_{i=1}^{k} \sum_{j=1}^{n_i} q_{ij}(\alpha) \log p_{ij} = \\ = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log p_{ij} \sum_{\alpha \in L^t} q_{ij}(\alpha) p(\alpha) \tag{6}$$

Пользуясь определением $M(S_{ij}(t))$, получаем:

$$\sum_{\alpha \in L^t} p(\alpha) \log p(\alpha) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log p_{ij} M(S_{ij}(t)) P(L^t) \tag{7}$$

Подставляя это выражение в (4), получаем:

$$H(L^t) = \log P(L^t) - \sum_{i=1}^{k} \sum_{j=1}^{n_i} M(S_{ij}(t)) \log p_{ij} \tag{8}$$

По определению, $P(L^t) = P_1(t) = O(t^{-1-\frac{1}{2}^{q_j - 1}})$, и $\log P(L^t) = O(\log t)$. Подставляя выражение для $M(S_{ij}(t)) = M_{ij}(t)$ в (8), получаем:

$$H(L^t) = O(\log t) - \sum_{i=1}^{k} \sum_{j=1}^{n_i} p_{ij} log p_{ij} d_i t^{\frac{1}{2}^{q_l^* - q}} = \\ = \sum_{i=1}^{k} \sum_{j=1}^{n_i} H(R_i) d_i t^{\frac{1}{2}^{q_l^* - q}} (1 + o(1)), \tag{9}$$

где $H(R_i) = -\sum_{j=1}^{n_i} p_{ij} \log p_{ij}$ — энтропия множества R_i правил вывода.

Обозначим $l' = maxl : l \in J$ — номер последнего критического класса. Элементы суммы при $i \in I_{l'}$ имеют вид $O(t^2)$, остальные имеют вид $o(t^2)$. Поэтому:

$$H(L^t) = \sum_{i \in I_{l'}} \sum_{j=1}^{n_i} d_i H(R_i) t^2 (1 + o(1)) \tag{10}$$

Сформулируем теорему:

Теорема 1 *Энтропия языка L^t, состоящего из слов длины t, порождаемых разложимой стохастической контекстно-свободной грамматикой, имеющей вид "цепочки выражается формулой*

$$H(L^t) \sim \sum_{i \in I_{l'}} \sum_{j=1}^{n_i} d_i H(R_i) \cdot t^2, \tag{11}$$

где $d_i > 0$, $H(R_i) = \sum_{j=1}^{n_i} p_{ij} \log p_{ij}$ — энтропия множества R_i правил вывода с нетерминалом A_i в левой части, и l' — номер критического класса, наиболее удалённого от начала цепочки.