Стационарная модель и энтропия языка

Опр (Стационарная модель открытого текста)

Стационарная модель открытого текста — это последовательность случайных величин $x_1, x_2, \cdots$, таких что

- x_i распределена на Σ,
- $\forall w \in \Sigma^k, \forall i, j \in \mathbb{N} : P(x_{i+1}, x_{i+2}, \cdots, x_{i+k} = w) = P(x_{j+1}, x_{j+2}, \cdots, x_{j+k})$
- Вероятности появления символов/биграмм/н-грамм не зависят от их позиции в тексте, т.е. $\forall k$ на Σ^k задано распределение вероятностей, работающее для любых подпоследовательностей длины k

Замечание 1

$\forall i, j, k \in \mathbb{N} : H(x_{i+1}, x_{i+2}, \cdots, x_{i+k}) = H(x_{j+1}, x_{j+2}, \cdots, x_{j+k})$

Доказательство

Подставим в определение энтропии эти вероятности.
$H(X_{i+1}X_{i+2}...X_{i+t}) = -\sum w \in \Sigma^t P(X_{i+1}X_{i+2}...X_{i+t} = \omega) \log P(X_{i+1}X_{i+2}...X_{i+t} = \omega)$

■

Замечание 2

$\forall i, j, k, s \in \mathbb{N}$:

$$H(x_{i+s+1}, x_{i+s+2}, \cdots, x_{i+s+k} \mid x_{i+1}, x_{i+2}, \cdots, x_{i+s}) =$$
$$H(x_{j+s+1}, x_{j+s+2}, \cdots, x_{j+s+k} \mid x_{j+1}, x_{j+2}, \cdots, x_{j+s})$$

Доказательство

Переименуем:

$$X = x_{i+s+1}, \cdots, x_{i+s+k}, \quad Y = x_{i+1}, \cdots, x_{i+s}$$

$$Z = x_{j+s+1}, \cdots, x_{j+s+k}, \quad E = x_{j+1}, \cdots, x_{j+s}$$

По цепному правилу:

$$H(X|Y) = H(X, Y) - H(Y), \quad H(Z|E) = H(Z, E) - H(E)$$

По замечанию 1 $H(X, Y) = H(Z, E), H(Y) = H(E) \Rightarrow H(X|Y) = H(Z|E)$ ■

Можем ввести обозначение $x_{i+1}, \cdots, x_{i+k} = x^k$, т.к. распределение не зависит от i, j, только от k

Опр (Условная взаимная информация)

$$I(X \leftrightarrow Y \mid Z) = H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z)$$

Теорема

$$I(X \leftrightarrow Y \mid Z) \geq 0$$

Доказательство

Как в теореме о взаимной информации. ■

Теорема для стационарного источника открытого текста

1. $H(x \mid x^n) = H(x_{i+n+1} \mid x_{i+1}, \cdots, x_{i+n}) \searrow$
2. $H_n(x) = \frac{H(x^n)}{n} \searrow$
3. $H_n(x) \geq H(x \mid x^{n-1})$
4. $\lim\limits_{n\to\infty} H_n(x) = \lim\limits_{n\to\infty} H(x \mid x^n)$

Доказательство 1

Из неравенства $I(X \leftrightarrow Y|Z) \geq 0$:
$H(x \mid x^{n-1}) - H(x \mid x^n) \geq 0$
■

Доказательство 2

$$H(x_1, \cdots, x_n) = H(x_1) + H(x_2 \mid x_1) + \cdots + H(x_n \mid x_1, \cdots, x_{n-1})$$

Перепишем через стационарность:

$$\geq n \cdot H(x_n \mid x_1, \cdots, x_{n-1})$$

$$\Rightarrow \frac{H(x^n)}{n} \geq H(x \mid x^{n-1})$$

■

Доказательство 3

То же самое, только снизу — значит, $H_n(x) \geq H(x \mid x^{n-1})$
■

Доказательство 4

$H_n(x)$ убывает, ограничена снизу, значит предел существует (по Вейерштрассу).
Показываем обе границы через теорему. Получаем равенство пределов.
■

Опр (Энтропия языка)

Энтропия языка L:

$$H_L = \lim_{n \to \infty} H_n(x)$$

Пусть:

$$H_0(x) = \log(|\Sigma|), \quad H_1(x) = -\sum_{i=0}^{l-1} p_i \log(p_i)$$

	H_0	H_1	H_2		H_L
Русск	5	4,35	3,52	...	1,37
Англ	4,7	4,15	3,62		1,5

Опр (Избыточность языка)

$$R_L = 1 - \frac{H_L}{\log(|\Sigma|)} = 1 - \frac{H_L}{H_0}$$

Это доля неиспользуемой выразительной способности каждой буквы.

- Для русского языка $R_L = 73\%$
- Для английского языка $R_L = 68\%$