ՀԱՅԿԱԿԱՆ ՍՍՌ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ՏԵՂԵԿԱԳԻՐ
ИЗВЕСТИЯ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР
Ֆիզիկա-մաթեմատ. գիտություններ XIV, № 4, 1961 Физико-математические науки

НАУЧНЫЕ ЗАМЕТКИ

Р. А. Казарин

Оценка энтропии армянского текста

§ 1. Введение

Осмысленный текст представляет собой последовательность неравновероятных и взаимозависимых символов S_i. В общем случае, когда взаимосвязи распространяются как угодно далеко, энтропия текста, рассматриваемого как источник информации, равна

$$H = \lim_{N\to\infty} F_N,$$

где

$$F_N = -\sum_{S_1=1}^{n}\sum_{S_2=1}^{n}\cdots\sum_{S_N=1}^{n} P(S_1, S_2, \dots S_N)\log P_{S_1, S_2 \dots S_{N-1}}(S_N).$$

Наиболее грубое приближение к H получим в предположении, что символы (буквы) текста равновероятны и взаимонезависимы т. е.

$$F_0 = \log n.$$

При этом $F_0 = H_{\max}$ есть максимальная энтропия источника с заданным алфавитом n.

Величина $\mu = H/H_{\max}$ есть нижняя грань коэффициента сжатия текста, а величина $1-\mu = R$ оценивает избыточную информацию.

Следующее приближение к H получим учитывая распределение простых вероятностей

$$F_1 = -\sum_{i=1}^{n} P(S_i)\log P(S_i).$$

F_2 — учитывает влияние одного предшествующего символа

$$F_2 = -\sum_{i,j} P(S_i\, S_j)\log P_{S_i}(S_j)$$

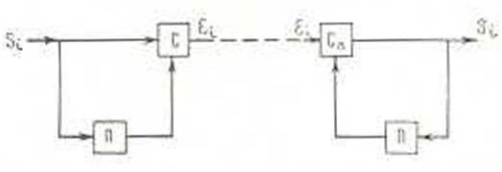
F_3 — учитывает влияние двух предшествующих символов и т. д.

Такой путь оценки H опирается на знание статистик многобуквенных сочетаний, получение же последних представляет громоздкую, а для четырех- и более-буквенных сочетаний, трудноосуществимую задачу.

Однако, можно сравнительно просто оценить энтропию H печатного текста способом, предложенным Шенноном [1]. В основе способа лежит утверждение о том, что эффективность предсказания есть мера организованности системы, т. е. эффективное предсказание возможно там, где имеется избыточная информация.

Система связи, в которой используется предсказание, строится по следующей блок схеме.

Символы сообщения поступают на сравнивающее устройства C куда одновременно подаются их предсказанные значения с выхода идеального предсказателя П, опирающегося в своем предсказании на прошлое передаваемого сообщения. На выходе C получается сокращенное сообщение (сигнал ошибки), содержащее меньшую избыточную информацию, поскольку взаимосвязи между его символами ослаблены в результате предсказания.



Фиг. 1.

Возможны два способа построения сокращенного текста. В первом способе сравнивающее устройство не выдает символа, если он предсказан точно и выдает истинный символ, при ошибочном предсказании. Во втором способе попытки предсказания каждого последующего символа повторяются до получения правильного результата. В канал, вместо данного символа, поступает число, соответствующее количеству попыток отгадывания. Идеальный предсказатель отгадывает в порядке убывающей условной вероятности. Очевидно, алфавиты исходного и сокращенного текстов содержат одинаковое число букв (одинаковое основание кода).

На приемном конце используется идентичный предсказатель. Поступающее на вход C_n число запускает предсказатель П приемника. Символ, полученный после соответствующего числа отгадываний, совпадает с исходным символом.

Обратимость сокращенного текста означает отсутствие потерь информации при таком преобразовании сообщения.

Поскольку взаимозависимость между символами сокращенного текста ослаблена, то энтропию сокращенного текста с достаточной точностью можно оценить пользуясь лишь одномерным распределением новых символов, т. е.

$$H_{\text{сокр}} = -\sum_{i=1}^{n} P(\varepsilon_i) \log P(\varepsilon_i),$$

где $P(\varepsilon_i)$—вероятность i-го символа в сокращенном тексте.

Иными словами, исходная k-связная цепь Маркова приближенно оценивается нуль-связной цепью Маркова, но уже не исходного, а сокращенного текста.

В опыте Шеннона предсказателем служит человек, думающий на данном языке. Инстиктивно используя статистику языка, такой оператор дает предсказание, близкое к идеальному.

§ 2. Некоторые свойства предсказания

Пусть S_i $(i=1, n)$ символы исходного сообщения, представляющего k-связную Марковскую цепь.

Тогда

$$P(S_i)=\sum_i P(B_i^{k-1}, S_j), \tag{2.1}$$

где B_i^{k-1} — последовательность $k-1$ предшествующих символов. Образуем сокращенный текст по второму из указанных выше способов. Тогда естественно обозначить символы сокращенного текста натуральными числами.

Символ „1" в сокращенном тексте появляется всякий раз, когда за последовательностью символов B_i^{k-1} следует такой символ S_j, для которого $P_{B_i^{k-1}}(S_j)$ максимальна.

Следовательно

$$P(1)=\sum_{i,j} P(B_i^{k-1}, S_j),$$

где суммирование ведется по всем i и по тем j, для которых $P_{B_i^{k-1}}(S_j)$ максимальна*.

Соответственно

$$P(2)=\sum_{i,j} P(B_i^{k-1}, S_j),$$

где суммирование ведется по всем i и по тем j, для которых $P_{B_i^{k-1}}(S_j)$ вторая по величине.

В общем случае

$$P(m)=\sum_{i,j} P(B_i^{k-1}, S_j), \tag{2.2}$$

где S_j — такие символы, для которых $P_{B_i^{k-1}}(S_j)$ по величине „m"-я.

Переписав (2.2) в виде

$$P(m)=\sum_{i,j} P(B_i^{k-1}) P_{B_i^{k-1}}(S_j)$$

видим, что предсказание представляет собой операцию усреднения. Известно [2], что усреднение ведет к возрастанию энтропии. Однако, при идеальном предсказании возрастание энтропии минимально, что и дает основание оценивать энтропию исходного текста посред-

* Следует подчеркнуть, что символ „1" в сокращенном тексте соответствует различным символам исходного текста.

ством энтропии сокращенного текста. Выясним некоторые особенности сокращенного текста полагая, что исходное сообщение представляет собой односвязную (простую) цепь Маркова. Это не ограничивает общности выводов, так-как многосвязную цепь Маркова можно свести к односвязной, применяя понятие многомерной случайной величины.

а) *Разброс простых вероятностей в сокращенном тексте больше, нежели в исходном.*

Действительно, пусть наибольшую вероятность в исходном тексте имеет символ S_1, а наименьшую—S_n. Тогда [см. (2.1) и (2.2)]

$$P(S_1)=\sum_i P(S_i, S_1),$$

$$P(1)=\sum_{i,j} P(S_i, S_j).$$

Количество слагаемых в обеих суммах одинаково, однако в то время, как во второй сумме берутся только максимальные условные вероятности, в первой сумме слагаемые не обязательно максимальны. Т. е.

$$P(1) \geq P(S_1).$$

С другой стороны

$$P(n) \leq P(S_n)$$

поскольку в выражении $P(n)$ суммируются только минимальные вероятности. Из последних двух неравенств следует высказанное выше утверждение.

б) *Взаимосвязи между символами в сокращенном тексте слабее, нежели в исходном.* Иными словами, в сокращенном тексте

$$P_l(k) \approx P(k). \tag{2.3}$$

Простоты ради ограничимся случаем, когда простые вероятности в исходном тексте примерно одинаковы

$$P(S_i) \approx \frac{1}{n}. \tag{2.4}$$

Это условие означает, что вся избыточная информация в сообщении обязана взаимосвязям между символами, а это, с точки зрения применения предсказания, наиболее интересный случай. В самом деле, если распределение простых вероятностей в исходном сообщении резко неравномерно, то именно оно и обусловливает избыточность. Но в этом случае предсказание неэффективно.

Учитывая (2.4), а также то, что для односвязной цепи Маркова

$$P_{S_i S_j}(S_t) = P_{S_j}(S_t),$$

получим

$$\frac{P_l(k)}{P(k)}=\frac{\sum\limits_{i,j,t} P(S_i S_j S_t)}{\sum\limits_{i,j} P(S_i S_j) \sum\limits_{j,t} P(S_j S_t)}=\frac{\sum\limits_{i,j,t} P(S_i S_j) P_{S_j}(S_t)}{\sum\limits_{i,j} P(S_i S_j) \sum\limits_{j,t} P(S_j) P_{S_j}(S_t)} \approx 1.$$

Число членов во всех суммах одинаково и равно n, т. к.

$$\sum_{j} P_{S_i}(S_j) = P_{S_i}(S_j).$$

Утверждение (2.3) выполняется точно, если условие (2.4) заменяется равенством. Последнее же в свою очередь означает, что матрица переходных вероятностей содержит идентичные строки.

в) *Предсказание неэффективно если*

$$P(S_j) = P(m) \quad (j,\ m = \overline{1,\ n}),$$

т. е.,

$$\sum_{i} P(S_i,\ S_j) = \sum_{i,\,j} P(S_i,\ S_j).$$

В частном случае двухбуквенного алфавита предсказание бессильно, если

$$\begin{vmatrix} P_a(a) > P_a(b) \\ P_b(a) > P_b(b) \end{vmatrix} \quad \text{или} \quad \begin{vmatrix} P_a(a) < P_a(b) \\ P_b(a) < P_b(b) \end{vmatrix}$$

где a и b буквы алфавита. Действительно, для первой матрицы

$$P(1) = P(aa) + P(ba) = P(a),$$
$$P(2) = P(ab) + P(bb) = P(b),$$

для второй же

$$P(1) = P(ab) + P(bb) = P(b),$$
$$P(2) = P(aa) + P(ba) = P(a).$$

Развив пример дальше, убедимся, что и матрицы переходных вероятностей сокращенного текста тождественно совпадут с исходными.

В заключение приведем пример двухбуквенного текста с матрицей вида

$$\begin{vmatrix} P_a(a) = \dfrac{1}{3}, & P_a(b) = \dfrac{2}{3} \\ P_b(a) = \dfrac{4}{5}, & P_b(b) = \dfrac{1}{5} \end{vmatrix}.$$

Из $P_a P_a(b) = P_b P_b(a)$ и $P_a + P_b = 1$ находим

$$P(a) = \frac{6}{11} = 0{,}545, \qquad P(b) = \frac{5}{11} = 0{,}455,$$

а также

$$P(aa) = \frac{2}{11}, \qquad P(ba) = \frac{4}{11},$$

$$P(ab) = \frac{4}{11}, \qquad P(bb) = \frac{1}{11}.$$

Кстати убеждаемся, что $\sum\limits_{i,\,j} P(i,\ j) = 1$.

Учитывая, что для односвязной цепи Маркова

$$P_{aa}(a) = P_{ba}(a) = P_a(a) \text{ и т. д.}$$

получим

$$P(aaa) = \frac{2}{33}, \quad P(abb) = \frac{4}{55}, \quad P(bba) = \frac{4}{55},$$

$$P(aab) = \frac{4}{33}, \quad P(baa) = \frac{4}{33}, \quad P(bbb) = \frac{1}{55},$$

$$P(aba) = \frac{16}{55}, \quad P(bab) = \frac{8}{33}, \quad \text{опять} \quad \sum_{i,j,k} P(i,j,k) = 1.$$

Перейдем к сокращенному тексту.

$$P(1) = P(ab) + P(ba) = \frac{8}{11} = 0{,}727,$$

$$P(2) = P(aa) + P(bb) = \frac{3}{11} = 0{,}273.$$

Далее

$$P(1,1) = P(aba) + P(bab) = \frac{88}{165}.$$

Аналогично

$$P(1,2) = P(2,1) = \frac{32}{165} \quad \text{и} \quad P(2,2) = \frac{13}{165}.$$

Учитывая, что

$$P_i(k) = \frac{P(i,k)}{P(i)}$$

находим

$$P_1(1) = \frac{11}{15} = 0{,}733, \qquad P_1(2) = \frac{4}{15} = 0{,}267,$$

$$P_2(1) = \frac{32}{45} = 0{,}711, \qquad P_2(2) = \frac{13}{45} = 0{,}289.$$

Видим, что

$$P(1) > P(a), \; P(2) < P(b)$$

и

$$P_1(1) \approx P_2(1) \approx P(1),$$
$$P_1(2) \approx P_2(2) \approx P(2).$$

Энтропия исходного текста

$$H = F_2 = -P(aa)\log P_a(a) - P(ab)\log P_a(b) - P(ba)\log P_b(a) -$$
$$- P(bb)\log P_b(b) = 0{,}827 \textit{ дв. ед. инф.}$$

Если пренебречь взаимосвязями, то получим

$$F_1 = -P(a)\log P(a) - P(b)\log P(b) = 0{,}994 \textit{ дв. ед. инф.},$$

т. е. преувеличенное значение.

Энтропия же сокращенного текста, при пренебрежении взаимосвязями,

$$F_{1_1} = -P(1)\log P(1) - P(2)\log P(2) = 0{,}844 \textit{ дв. ед. инф.}$$

близка к энтропии H.

Таким образом, если по каналу связи передавать посимвольно (т. е. пренебрегая взаимосвязями) в одном случае исходный, а в другом — сокращенный тексты, то в первом случае канал будет перегружен избыточной информацией, равной 0,15 *дв. ед. инф.*. Данные таблиц, составленных на основании предсказания (см. § 3), используются для получения верхней и нижней оценок энтропии текста. Верхней оценкой, на основании предшествующего [см. пояснения к выражению (2.2)], может служить ряд значений величины

$$H^k_{\max} = -\sum_{i=1}^{n} q^k_i \log q^k_i, \tag{2,5}$$

где q^k_i — частота символа i в k-ом столбце таблицы I (или II).

В качестве нижней грани энтропии текста взяты значения величины

$$H^k_{\min} = \sum_{i=1}^{n} i\,(q^k_i - q^k_{i+1})\log i, \tag{2,6}$$

предложенной Шенноном. Для пояснения (2.6) заметим, что величину q^k_i можно представить в виде следующей суммы:

$$q^k_i = q^k_i - q^k_{i+1} + q^k_{i+1} - q^k_{i+2} + q^k_{i+2} + \cdots + q^k_n - q^k_{n+1} = \sum_{i=1}^{n}(q^k_i - q^k_{i+1})$$

и составить таблицу (для k-го столбца)

$$\xrightarrow{\qquad} Y$$

$$\begin{array}{ll}
q_1 = & q_1 - q_2 + q_2 - q_3 + q_3 - q_4 + \cdots + q_i - q_{i+1} + \cdots + q_n - q_{n+1} \\
q_2 = & \qquad\quad q_2 - q_3 + q_3 - q_4 + \cdots + q_i - q_{i+1} + \cdots + q_n - q_{n+1} \\
\cdots & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\downarrow X \quad q_i = & \qquad\qquad\qquad\qquad\qquad\quad q_i - q_{i+1} + \cdots + q_n - q_{n+1} \\
\cdots & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
q_n = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad q_n - q_{n+1}
\end{array}$$

Очевидно

$$\sum_{i=1}^{n} q_i = \sum_{i=1}^{n}(q^k_i - q^k_{i+1}) = 1.$$

Каждый i-тый столбец (Y — задано) содержит i равновероятных членов. Следовательно

$$H_{Y=i}(x) = \log i.$$

Вероятность же i-го столбца $P(Y=i) = i\,(q_i - q_{i+1})$, отсюда средняя условная энтропия

$$H_Y(x) = \sum_{i=1}^{n} i\,(q_i^k - q_{i+1}^k) \log i.$$

С другой стороны

$$H_Y(x) \le H(x),$$

где

$$H(x) = -\sum_i q_i^k \log q_i^k.$$

§ 3. Результаты опыта

Получены оценки для армянского художественного и технического текстов. С этой целью из смешанного современного художественного текста* наудачу были выбраны 100 штук 20-буквенных отрезков (включая пробелы), которые, вторым из описанных выше способов, были преобразованы в числовые последовательности. Иными словами, предсказывающему лицу предлагалось букву за буквой отгадывать данный 20-буквенный фрагмент текста. При неверном отгадывании данной буквы, ему предлагалось повторять попытки до получения правильного результата. Под данной буквой записывалось число, соответствующее количеству попыток, приведших к правильному результату. После этого предсказывающее лицо, опираясь на уже отгаданные предшествующие буквы, переходило к отгадыванию последующей буквы. Во время отгадывания операторы располагали таблицами частот букв армянского алфавита [3], частот начал слов, а также словарями.

Результаты предсказания всех ста 20-буквенных фрагментов сведены в таблицу 1. Столбцы в этих таблицах соответствуют числу предшествующих букв, использованных при отгадывании последующей буквы плюс единица, строки же означают число попыток, приведших к правильному отгадыванию. Так в таблице 1 число 10, стоящее на пересечении 5-го столбца и 3-й строки, означает, что при известных четырех предшествующих буквах правильный результат с третьей попытки получился 10 раз из 100. То же самое проделано для 150-и 20-буквенных фрагментов технического текста**. Результаты сведены

* Были использованы номера журнала «Սովետական գրականություն» за 1958, 1959 годы.

** Тексты выбирались в основном из электрорадиотехнической литературы, а предсказание вели в основном студенты старших курсов факультета физики ЕГУ (специализация „радиофизика“).

ТАБЛИЦА I

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.7	32	41	55	61	62	64	65	65	66	54	64	53	61	70	72	72	71	82	73
2	11.4	16	21	12	11	9	15	15	11	10	14	9	14	14	9	13	12	10	10	13
3	8.3	15	12	7	10	6	6	6	10	7	12	5	9	7	9	6	8	6	2	7
4	7.5	6	5	5	5	5	4	4	5	5	5	3	3	3	2	4	3	3		1
5	5.4	6	6	6	4	5	2	4	2	5	1	4	5	3	2	1	1	2	3	
6	4.8	10	6	2	1	3	3	2	3		3	4	1		2		2		1	2
7	3.3	5	1		2		3	3	1		1		1	4	2	1				2
8	3.0	1	3	3	1	2	1			2		3	3	1	1			2		
9	2.5	2	1	1	1		1		1			1	2		1	1		2		
10	2.4	1	1	2	1						2	1	1	2				1	1	1
11	2.4	1	1	1			1			2				1						1
12	2.0	1		1	2	1						4	2		1					
13	1.7	2						1		1	3		1	1			1	1		
14	1.6			1					1		2									
15	1.6	1	2	1		2					1						1	1	1	
16	1.5									1				2	1					
17	1.4	1				1					1		1							
18	1.4											2								
19	1.3			1		2							1							
20	1.2					1			1	1						1				
21	1.2													1		1				
22	1.1																			
23	0.9																			
24	0.9				1								1							
25	0.9																			
26	0.8			1		1							1							
27	0.8										1							1		
28	0.7																			
29	0.7																			
30	0.5																			
31	0.4																			
32	0.4																			
33	0.4																			
34	0.4																			
35	0.4																			
36	0.3																			
37	0.2																			
38	0.1																			
39	0.07																			
40	0.03																			
Hmx	4.16	3.03	2.61	2.52	2.09	2.18	1.87	1.76	1.84	1.87	2.27	2.06	2.52	2.09	1.70	1.50	1.47	1.68	1.04	1.41
Hmn		2.24	1.79	1.55	1.24	1.33	1.05	1.00	1.04	1.09	1.29	1.26	1.59	1.24	0.94	0.79	0.78	0.94	0.51	0.74

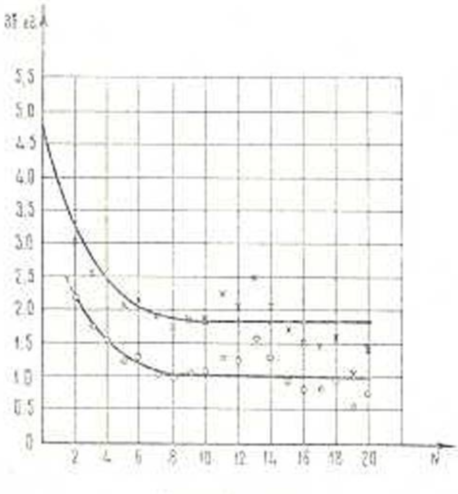
в таблицу 2. Первые столбцы в обеих таблицах получены непосредственно из таблицы частот букв армянского алфавита [3].

В процессе опытов выяснилось, что эффективность предсказания у отдельных лиц в среднем одинакова, что же касается технического текста, то у лиц некомпетентных в данной области, эффективность предсказания несколько (хотя и незначительно) падала.

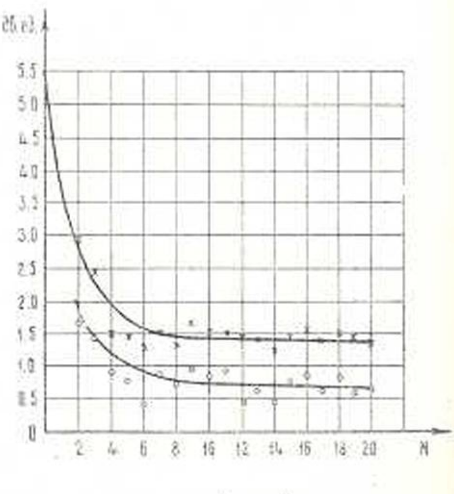
Результирующие кривые верхних и нижних оценок энтропии ($H^k_{\max}$ и $H^k_{\min}$), усредненные по методу минимизации среднеквадратической ошибки, приведены на фиг. 2 (художественная проза) и фиг. 3 (технический текст). Из них следует, что энтропия художественного текста H_x примерно равна 1,4 *дв. ед.* Поскольку $H_{\max} = \log 40 = 5{,}3$ *дв. ед.* (вместе с пробелом), то предельное сжатие текста, достижимое при применении идеального кода $\mu = 0{,}26$, а избыточность $R \approx 75\%$.

ТАБЛИЦА II

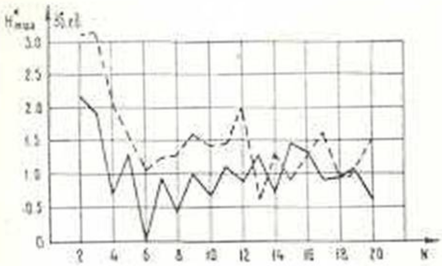
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		53	80	109	110	116	107	121	107	108	107	119	118	122	117	107	119	109	109	117
2		27	21	16	16	15	19	8	12	17	23	14	12	9	11	16	13	18	20	17
3		17	13	9	11	6	8	4	11	10	2	6	5	7	5	12	3	9	7	5
4		7	8	5	6	1	4	6	5	2	3	4	2	5	1	3	3	2	6	1
5		12	5	3	1	5	5	2	1	4	4	3	3	1	2	5	2	4	2	1
6		12	4		1	5	3	1	3	4	5	2	2	1	3	2	1	1		2
7		3	3	1	1			1	1		2	1			2	1	1	2	3	1
8		8	2	2	1				1	1	1	2	1		2	1	3		2	
9		1	2				1		1	1	1				1	1		2	1	1
10		2	3	1		1			2			1			2		1			1
11			4					1					1	1	1		2			2
12									3			1		1						
13		2	1				1	1					1			1	1	1		
14		2	2	1				1					2							
15		2	1		1		1	1			1	1	1	1				2		1
16				1					1											
17				2								1		1	1					
18			1			1														1
19		1			1				1						1					
20										2										
21									1						1					
22										1										
23																				
24																				
25																1				
26																				
27																				
28																				
29								1												
30													1	1			1			
31		1																		
32					1		1	1			1									
33																				
34																				
35													1							
36																				
37																				
38																				
39																				
40																				
Hmx	4.16	2.94	2.48	1.54	1.49	1.28	1.56	1.34	1.70	1.54	1.52	1.46	1.37	1.21	1.45	1.57	1.39	1.50	1.44	1.30
Hmn		1.68	1.46	0.87	0.75	0.41	0.83	0.73	0.9	0.79	0.86	0.44	0.63	0.48	0.73	0.84	0.63	0.78	0.60	0.69



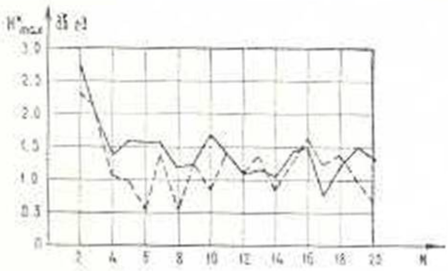
Фиг. 2.



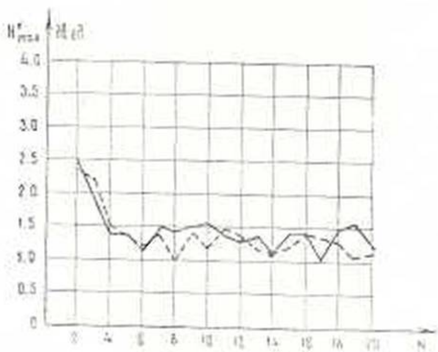
Фиг. 3.



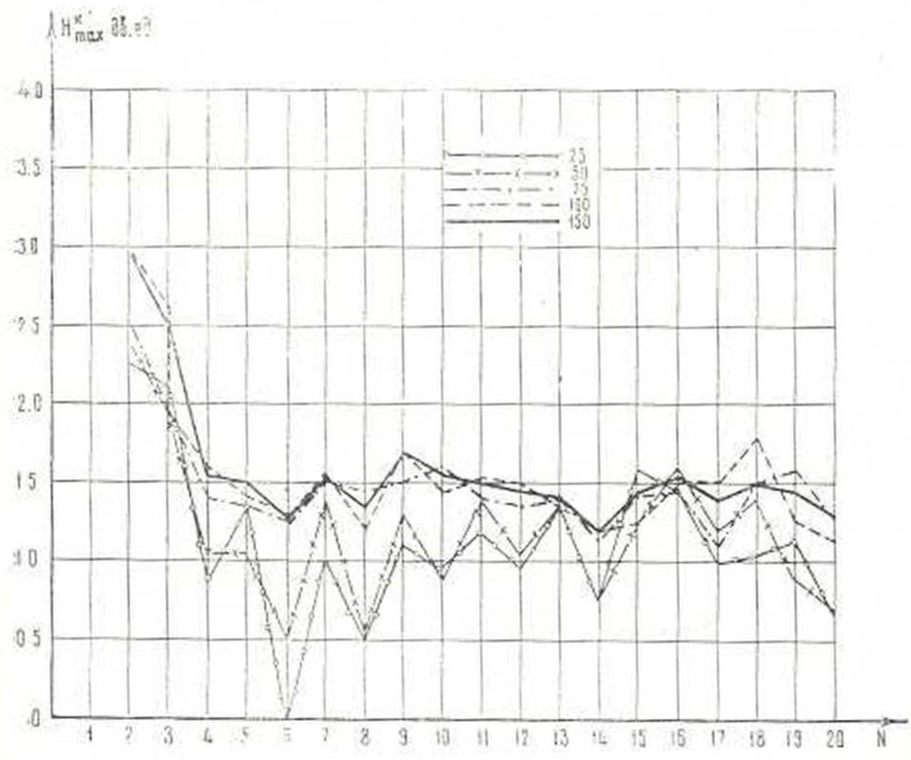
Фиг. 4а.



Фиг. 4б.



Фиг. 4в.



Фиг. 5.

Аналогично для технического текста $H_T = 1{,}0 \div 1{,}1$ *дв. ед.*, $\rho \approx \approx 0{,}20$, $R \approx 80\%$.

Следует отметить, что при использовании более длинных отрезков текста (скажем 50 или 100 букв), энтропия заметно понижается. Однако, как показывает ход кривых фиг. 2 и фиг. 3, существенное сокращение избыточности получается при учете 5—6 предшествующих букв. Дальнейшее увеличение прошлого—малоэффективно.

Графики фиг. 4 и 5 приведены в целях оправдания объема привлеченной статистики. Каждая кривая фиг. 4а получена на основе 25 отрезков, выбранных наудачу. При этом кривые не содержат одинаковых отрезков. Аналогично, кривые фиг. 4б соответствуют 50, а фиг. 4в—75 отрезкам текста. Уменьшение флуктуаций с ростом числа отрезков очевидно. На фиг. 5 совмещены кривые для 25, 50, 75, 100, 150 отрезков. Сравнение их показывает, что существенное уменьшение флуктуации происходит до кривой, соответствующей 75 отрезкам. Дальнейшее добавление числа исследуемых отрезков, незначительно сглаживает кривую значений $H^k_{\max}$.

Считаю своим приятным долгом выразить глубокую благодарность всем лицам, бескорыстно помогавшим в сборе и обработке статистического материала, а также Г. А. Амбарцумян и Д. С. Лебедеву за ряд полезных замечаний, поправок и содержательных бесед.

Ереванский государственный
университет

Поступила 31 X 1960

Ռ. Ա. Ղազարյան

ՀԱՅԵՐԵՆ ՏԵՔՍՏԻ ԷՆՏՐՈՊԻԱՅԻ ԳՆԱՀԱՏՈՒՄԸ

ԱՄՓՈՓՈՒՄ

Ուսումնասիրված տեքստն իրենից ներկայացնում է փոխադարձաբար կապված առանձնահատկական սիմվոլների (տառերի) հաջորդականություն: Ընդհանուր դեպքում, երբ ստատիստիկական կապերը տարածվում են 3-ից ավելի սիմվոլների վրա, այդպիսի հաջորդականության էնտրոպիայի գնահատումը գործնականորեն անհնարին է, քանի որ դրա համար անհրաժեշտ է անհնար բազմաթիվ զուգակցությունների ստատիստիկան: Սակայն հնարավոր է համեմատաբար հեշտ կերպով գնահատել տեքստի էնտրոպիան, եթե տեքստը նախօրոք փոխակերպվի (կոդավորվի) այնպիսի հաջորդականության, որի միջսիմվոլային կապերը բացակայեն, կամ հնարավորին չափ թուլացված լինեն: Իսկ միջսիմվոլային կապերը քայքայող (դեկոդեր) գործողության կիրառվում է գուշակումը:

Գուշակումը կիրառող կոդի սխեմի (նկ. 1) կանոնով հաջորդվում է ոչ թե նախնական հաջորդականությունը, այլ փոխակերպվածը (կրճատված հաջորդագիր կամ սխալանքի ազդանշան):

Եթե կապի կանալով տառ առ տառ (այսինքն անտեսելով միջախմբային կապերը) հաղորդենք նախնական հաղորդագիրը, ապա կանալը կբեռնվի զգալի ավելցուկային ինֆորմացիայով, իսկ եթե նույնը կատարենք կրճատված հաղորդագրի հետ, ապա ավելցուկային ինֆորմացիան չնչին կլինի, քանի որ միջախմբային կապերը տվյալ դեպքում համարյա բացակայում են։

Ընդունման կետում վերականգնվում է նախնական հաղորդագիրը։ Շրջելիությունը նշանակում է, որ այդ գործողությունների ընթացքում ինֆորմացիայի կորուստ տեղի չի ունենում։ Մյուս կողմից՝ կրճատված հաղորդագրի միջախմբային կապերի թուլացման շնորհիվ նրա էնտրոպիայի գնահատման համար բավական է խմբավորել միայն տառերի (սիմվոլների) համախմբությունը։

Իդեալական գուշակը գուշակում է ըստ պայմանական հավանականությունների նվազման կարգի (տրված անցյալի դեպքում)։

Այնուհետև ընդունվում են գուշակման էֆեկտիվության պարամետրերը և հիմնավորվում հետևյալ հատկությունները՝ ա) սիմվոլների հավանականությունները կրճատված տեքստում ավելի ուժեղ են ցրված, քան նախնական տեքստում, և բ) միջախմբային կապերը գուշակման շնորհիվ թուլանում են։

Նկարագրվող էքսպերիմենտում (Շեննոնի մեթոդը) գուշակիչ է ծառայում տվյալ լեզվով մտածող մարդը։ Ենթագիտակցաբար կիրառելով լեզվի ստատիստիկան, նա տալիս է իդեալականին մոտ գուշակում։ Կրճատված տեքստը կառուցվում է հետևյալ կերպ՝ գուշակչին առաջարկվում է գուշակել հաջորդ տառը։ Եթե նա սխալվում է, ապա նրան առաջարկվում է փորձել նորից։ Փորձումները կատարվում են մինչև ճիշտ արդյունքի ստացումը։ Այդ տառի փոխարեն գրվում է փորձումների քանակին համապատասխան թիվ։

Ուսումնասիրվել են ժամանակակից խոսք գեղարվեստական (արձակ) և էլեկտրառադիոտեխնիկական տեքստերը։ Յուրաքանչյուրից պատահական կերպով ընտրվել են 100 (տեխնիկականի համար 150) հատ 20-տառանի հատվածներ, որոնք վերը նկարագրված ձևով վեր են ածվել թվային հաջորդականությունների։ Այդ տվյալների հիման վրա կառուցված են № 1 (արձակ) և № 2 (տեխնիկական) աղյուսակները։

Այսպես՝ № 1 աղյուսակի 5-րդ սյան և 3-րդ տողի հատման բջջում գրված 10 թիվը նշանակում է, որ երբ հայտնի են 4 հատ նախորդ տառ, ապա ճիշտ գուշակումը 3 փորձումից հաջողվել է 100-ից 10 անգամ։

Աղյուսակների հիման վրա ստացվել են տեքստի էնտրոպիայի վերին և ստորին գնահատականները (նկ. 2-արձակի համար, և նկ. 3-տեխնիկականի)։

Պարզված է, որ արձակի ավելցուկությունը մոտ 75% է, իսկ տեխնիկականինը՝ 80%, այսինքն՝ կիրառելով իդեալական տիպի կոդեր, հնարավոր է կրճատել տեքստային հաղորդումները 4–5 անգամ։ Նկ. 4-ում և նկ. 5-ում պատկերված կորերը արտահայտում են օգտագործված ստատիստիկայի ծավալը։

ЛИТЕРАТУРА

1. *Shannon C. E.* Prediction and Entropy of Printed English. „BSTY“, January, 1951.
2. *Elias Peter.* Predictive Coding. „Transaction IRE, PGIT—1“, March, 1955.
3. *Казарян Р. А.* Частоты букв в армянском тексте. „Известия АН Армянской ССР, серия физ.-мат. наук“, 14, № 1, 1961.