А. И. Шкловский, Некоторые вопросы графической регулярности, *Зап. научн. сем. ЛОМИ*, 1971, том 20, 272–281

А.И.Шкловский

НЕКОТОРЫЕ ВОПРОСЫ ГРАФИЧЕСКОЙ РЕГУЛЯРНОСТИ *)

I. Задача экстраполяции последовательностей (см., например, [I]) и связанные с ней задачи распознавания образов (см., например, [2], стр. 94 и [3]) естественным образом приводят к необходимости выявления и изучения внутренних связей (регулярностей) в имеющейся информации.

В связи с этим мы рассматриваем следующий вопрос. Пусть имеется некоторое слово (например, экстраполируемая последовательность). В какой мере можно сократить это слово путем введения новых "атомарных понятий" для его регулярных частей, т.е. для многократно встречающихся в нем достаточно длинных подслов. Точнее, мы рассматриваем следующий метод выявления регулярности: вместо непересекающихся вхождений слова B в слово X подставляется новый знак α и перед получившимся словом приписывается слово $\alpha \to B *$, где $*$ - разделительный знак, отличный от α и букв слова X. Этот процесс выявления графической регулярности повторяется. Начальное слово X называется сократимым, если суммарная длина получившегося слова (включающего схемы подстановок) меньше длины X. В работе в основном исследуются характеристики наибольших по длине несократимых слов в зависимости от числа различных букв в этих словах. Использование нами аппарата порождающих бесконтекстных грамматик вызвано, во-первых, удобством его для решения поставленных задач и, во-вторых, имеющимся опытом использования этого аппарата для решения задач экстраполяции после-

*) Основные результаты настоящей заметки были доложены на Ленинградском семинаре по конструктивной математике 2 октября 1969 года.

довательностей (см. [I]). Методы выявления регулярностей, аналогичные нашим, предлагались М.М.Бонгардом (см. [2], стр.209).

Основным результатом работы является нахождение для бесконтекстной грамматики точной верхней оценки длины несократимых слов в зависимости от числа различных букв в них. Кроме того, найдена точная нижняя оценка возможной величины сокращения слова.

Если рассматривать наши методы сокращения слов как методы кодирования, то они не дают уменьшения количества информации (в обычном для теории связи понимании) для всех слов, т.к. кодирование вызывает расширение алфавита, однако на достаточно регулярных словах выигрыш получается столь значительным, что превышает потери на расширение алфавита. Поэтому некоторые результаты работы могут быть интерпретированы как результаты теории связи. Но, как уже говорилось, основной нашей целью является рассмотрение методов выявления регулярностей для решения задач экстраполяции последовательностей, а эти задачи существенно отличаются от задач теории связи.

2. Порождающая грамматика - это четверка вида (V_T, V_N, P, S) (будем обозначать ее через G), где V_T - основной алфавит, не пересекающийся со вспомогательным алфавитом V_N; P - список схем подстановок, т.е. пар вида (φ_i, ψ_i), где φ_i и ψ_i - некоторые слова в алфавите $V_N \cup V_T$; S - аксиома, т.е. непустое слово в $V_N \cup V_T$. Обычно требуется, чтобы слово S было однобуквенным, но мы не накладываем ограничений на длину S.

Большими латинскими буквами из начала алфавита, возможно с индексами, будем обозначать буквы из V_N. Если U - алфавит, то через U^* будем обозначать множество всех слов в алфавите U.

Порождающая грамматика G называется б е с к о н т е к с т н о й, если каждая схема из P имеет вид (A, ψ), где ψ - непустое слово из $(V_N \cup V_T)$. Запись $\varphi_1 \vdash \varphi_2$ означает, что φ_1 представимо в виде $w_1 A w_2$, а φ_2 представимо в виде $w_1 \psi w_2$, где $(A, \psi) \in P$, а w_1 и w_2 принадлежат к $(V_N \cup V_T)^*$. Запись $\varphi \vdash\!\!= \psi$ означает, что су-

ществует последовательность $\varphi_0, \ldots, \varphi_n$ (называемая в ы в о д о м ψ из φ такая, что $\varphi_0 = \varphi$, $\varphi_n = \psi$ и $\varphi_{i-1} \vdash \varphi_i$, $i = 1, 2, \ldots, n$. Множество $L(G) = \{X : X \in V_T^*$ и $S \models X\}$ будем называть я з ы к о м, порожденным грамматикой G. В дальнейшем, говоря о какой-либо грамматике, будем иметь в виду бесконтекстную порождающую грамматику, язык которой состоит из одного слова.

3. Следующим образом определим з а п и с ь $P : \lrcorner P \lrcorner = \varphi_l \psi_l^* \ldots\ldots\ldots \varphi_1 \psi_1^*$, где l – число схем подстановок в P. Через $|X|$ будем обозначать длину слова X.

$\tau(X)$ – мерой графической регулярности слова X – будем называть величину $\min\limits_{L(G)=X} | \lrcorner P \lrcorner S |$

Ту грамматику G, на которой достигается этот минимум, назовем о п т и м а л ь н о й для X. Очевидно, что $\tau(X) \leqslant |X|$ для всякого X.

Пусть P состоит из l пар $(A_1, \psi_1), \ldots, (A_l, \psi_l)$. Будем говорить, что P (или G) имеет п р и в е д е н н ы й в и д, если

(а) $A_i \neq A_j$ при $i \neq j$

(б) если $j \geqslant i$, то A_i не входит в ψ_j

(в) $|\psi_i| \geqslant 2$ $(1 \leqslant i \leqslant l)$

Теорема I. Д л я в с я к о г о с л о в а X с у щ е с т в у е т о п т и м а л ь н а я д л я н е г о г р а м м а т и к а, и м е ю щ а я п р и в е д е н н ы й в и д.

Замечание. Не всякая грамматика, имеющая приведенный вид, является оптимальной.

4. Назовем слово X с о к р а т и м ы м грамматикой G (грамматику G – с о к р а щ а ю щ е й слово X), если $L(G) = X$ и $|\lrcorner P \lrcorner S| < X$. Будем говорить, что слово X с о к р а т и м о, если существует сокращающая его грамматика.

Через Y^m будем обозначать слово $YY\ldots Y$, составленное из m слов Y.

Определим оператор P_τ (оператор проекции) на всех словах вида X/Y, где Y непусто, следующим образом. Положим $P_\tau(X/Y) = Y^m$, где m – наибольшее число, для которого существует представление X в виде $W_1 Y W_2 \ldots W_m Y W_{m+1}$.

Теорема 2. Слово X тогда и только тогда сократимо грамматикой, множество схем подстановок которой состоит только из одной пары когда

$$\frac{P_\tau(X/Y)}{|Y|} > \frac{|Y|+2}{|Y|-1} \qquad (*)$$

Доказательство. Пусть выполнено условие (*). Положим $P = \{(A, Y)\}$. Тогда $|\langle P \rangle| = |Y| + 2$.

Пусть $P_\tau(X/Y) = Y^m$. Тогда X можно представить в виде $W_1 Y W_2 \ldots W_m Y W_{m+1}$. Через S обозначим результат замены в этом слове всех выделенных вхождений Y на A. Нетрудно видеть, что

$$|S| = |X| - m(|Y|-1).$$

Отсюда и из (*) получаем:

$$|\langle P \rangle| + |S| = |Y| + 2 + |X| - m(|Y|-1) < |X|$$

Обратное утверждение доказывается с помощью аналогичных выкладок.

Теорема 3. Если слово X сократимо, то оно сократимо и некоторой грамматикой, множество схем подстановок которой состоит только из одной пары.

Доказательство. Пусть G – сокращающая X грамматика, оптимальная для X и имеющая приведенный вид. Пусть S – ее аксиома, а $(A_\ell, \psi_\ell), \ldots (A_1, \psi_1)$ – список ее схем подстановок. Через Z_0 обозначим S. Результат применения (A_1, ψ_1) к S, т.е. результат подстановки в S вместо всех вхождений

A_1 слова ψ_1, обозначим через Z_1. Результат применения (A_2, ψ_2) к Z_1 обозначим через Z_2 и т.д. Тогда $Z_\ell = X$. Найдем наибольшее j такое, что

$$|Z_j| + |A_\ell \psi_\ell * A_{\ell-1} \psi_{\ell-1} * \ldots A_{j+1} \psi_{j+1} *| < |X|.$$

Поскольку

$$|Z_{j+1}| + |A_\ell \psi_\ell * \ldots * A_{j+2} \psi_{j+2} *| \geqslant X,$$

то отсюда и из предыдущего неравенства получаем, что

$$|A_{j+1} \psi_{j+1}| + |Z_j| < |Z_{j+1}|.$$

Следовательно, должно быть выполнено условие (*) из теоремы 2 при $X = Z_{j+1}$, $Y = \psi_{j+1}$.

Применим к слову ψ_{j+1} всюду, где возможно, схему (A_{j+2}, ψ_{j+2}), затем схему (A_{j+3}, ψ_{j+3}) и т.д. до схемы (A_ℓ, ψ_ℓ) включительно. Получившееся в результате слово обозначим через $\overline{\psi_{j+1}}$. Очевидно, что $|\overline{\psi_{j+1}}| \geqslant |\psi_{j+1}|$ и всякому вхождению ψ_{j+1} в Z_{j+1} соответствует вхождение $\overline{\psi_{j+1}}$ в X. Отсюда видно, что условие (*) выполнено при $Y = \overline{\psi_{j+1}}$. Представим X в виде $W_1 \overline{\psi_{j+1}} W_2 \ldots \overline{\psi_{j+1}} W_{m+1}$ с наибольшим возможным m. В силу теоремы 2 слово X сократимо грамматикой, единственной схемой подстановок которой является $(A, \overline{\psi_{j+1}})$, а аксиомой - слово $W_1 A W_2 \ldots A W_{m+1}$. Теорема доказана.

5. Рассмотрим последовательность расширяющихся алфавитов $\Delta_1, \Delta_2, \ldots, \Delta_n, \ldots$, где $\Delta_1 = \{a_1\}$, $\Delta_{n+1} = \Delta_n \cup \{a_{n+1}\}$.

Теорема 4. Если $X \in \Delta_n^*$ и $|X| > 4n^2 + 3n + 1$, то X сократимо.

Доказательство. Определим функцию $\text{деф}_\ell(X)$, которую будем называть дефицитом ℓ-буквенных блоков в слове X, следующим равенством:

$$\text{деф}_\ell(X) = \max\{0, |X| - \ell + 1 - \frac{1}{\ell}\sum_{|z|=\ell} |P_z(X/z)|\}.$$

Понятие дефицита дает возможность учесть, например, различие между словами $X_1 = aaaaaaaaa$ и $X_2 = ababababa$, заключающееся в том, что из X_1 можно выделить только 4 непересекающихся вхождения подслова aa, а из X_2 – либо 4 вхождения подслова ab, либо 4 вхождения подслова ba. В дальнейшем $\text{деф}_2(X)$ будем называть д е ф и ц и т о м в слове X и обозначать через $\text{деф}(X)$.

(а) Докажем, что каково бы ни было слово X, $\text{деф}(X) = 0$ тогда и только тогда, когда X не содержит ни одного подслова вида a_i^3. Действительно, если X пустое, то $\text{деф}(X) = 0$, а если X непустое и не содержит ни одного такого подслова, то все двухбуквенные подслова, у которых вхождение последней буквы одного совпадает с вхождением первой буквы другого – различны. В этом случае $\sum_{|y|=2} |P_z(X/y)| = 2(|X| - 1)$ и, следовательно, $\text{деф}(X) = 0$.

Аналогично доказывается и обратное утверждение.

(б) Непосредственным вычислением нетрудно убедиться, что

$$\text{деф}(a_i^k) = \left[\frac{|k-1|}{2}\right].$$

(в) Пусть $X, Y \in \Delta_n^*$ и последняя буква слова X отлична от первой буквы слова Y. Всем, кроме одного, вхождениям двухбуквенных слов в слово XY соответствуют вхождения этих же слов либо в X, либо в Y. Единственное новое вхождение двухбуквенного подслова в XY появилось на границе X и Y. Но по условию это двухбуквенное слово состоит из разных букв и, значит, учтется при суммировании. Следовательно,

$$\sum_{|z|=2} |P_z(XY/z)| = \sum_{|z|=2} |P_z(X/z)| + \sum_{|z|=2} |P_z(Y/z)| + 2$$

и поэтому $\text{деф}(XY) = \text{деф}(X) + \text{деф}(Y)$.

(г) Докажем, что $\mathrm{def}(X) \leq 3n$ для любого несократимого X из Δ_n^*. Пусть X – несократимое слово из Δ_n^* и

$$X = Z_1 a_{i_1}^{l_1} Z_2 a_{i_2}^{l_2} \ldots Z_{N+1} \quad (**),$$

где ни одно из слов Z_j $(1 \leq j \leq N+1)$ не содержит подслов вида a_i^3, все $l_k \geq 3$ $(1 \leq k \leq N)$ и взяты максимально возможными, т.е. a_i^k не является ни последней буквой слова $Z_1 a_{i_1}^{l_1} \ldots Z_k$ ни первой буквой слова $Z_{k+1} a_{i_{k+1}}^{l_{k+1}} \ldots Z_{N+1}$.

В силу п(а) $\mathrm{def}(Z_j) = 0$ $(1 \leq j \leq N+1)$. Отсюда и из п(в) вытекает, что

$$\mathrm{def}(X) = \sum_{k=1}^{N} \left[\frac{l_k - 1}{2} \right] \quad (***)$$

Будем говорить, что представление (**) в ы д е л я е т подслова $a_{i_k}^{l_k}$. Так как X несократимое, то в представлении (**) не может быть 3 или более выделенных вхождений подслов, содержащих букву a_i, т.к. иначе будет выполнено условие (*) из теоремы 2 с a_i^3 в качестве y. Следовательно, для всякой буквы a_i^3 из Δ_n имеет место один из трех случаев.

(г I) Среди выделенных подслов нет ни одного, содержащего a_i^3. В этом случае буква a_i не дает никакого "вклада" в величину $\mathrm{def}(X)$, вычисляемую по (***).

(г 2) Есть ровно одно выделенное подслово вида $a_i^k (k \geq 3)$. Тогда, по условию (*) из теоремы 2 $\mathrm{def}(a_i^k) \leq 3$. Значит в (***) есть только одно слагаемое, обусловленное вычислением дефицита слова вида a_i^k, и оно не превосходит 3.

(г 3) Есть два выделенных подслова: a_i^k и a_i^p. Возможны только следующие значения k и p:

$k=3$, $p=3$;	$k=4$, $p=3$;
$k=3$, $p=4$;	$k=4$, $p=4$;
$k=5$, $p=4$;	$k=4$, $p=5$;
$k=3$, $p=5$;	$k=5$, $p=3$;

поскольку для наборов κ и ρ, хотя бы по одному из параметров больших, чем приведенные в таблице, слово X будет сократимым, т.к. выполнится условие (*) из теоремы 2 с a_i^3 в качестве y Вычислив значения $\partial e\phi\,(a_i^{\kappa}) + \partial e\phi\,(a_i^{\rho})$ для κ и ρ, перечисленных в таблице, нетрудно убедиться, что "вклад" буквы a_i в величину $\partial e\phi(X)$ при вычислении ее по (***), также не превосходит 3.

Таким образом "вклад" всякой буквы $a_i\,(1 \le i \le n)$ в величину $\partial e\phi\,(X)$ не превосходит 3 и, значит, $\partial e\phi\,(X) \le 3n$. По условию (*) из теоремы 2 для всякого подслова Z при $|Z|=2$ выполняется неравенство $|P_r(X/Z)| \le 8$. Значит,

$$\sum_{|Z|=2} |P_r(X/Z)| \le 8n^2,$$

и поэтому $|X| \le 4n^2 + 3n + 1$. Теорема доказана.

6. Докажем, что оценка, полученная в теореме 4, точна. Определим отношение строгой принадлежности слова к алфавиту: $X \Subset \Delta_n^*$ означает, что $X \in \Delta_n^*$ и всякая буква из Δ_n входит в X.

Теорема 5. Для любого положительного n можно построить такое несократимое слово X, строго принадлежащее к Δ_n^*, что $|X| = 4n^2 + 3n + 1$.

Доказательство. Требуемое слово, обозначаемое ниже посредством X_n, является результатом индуктивного построения, определяемого следующими (графическими) равенствами:

$$X_1 = a_1^8,$$

$$X_{i+1} = a_{i+1} X_i a_{i+1}^8 (a_i a_{i+1})^3 (a_1 a_{i+1})^4 \ldots (a_{i-1} a_{i+1})^4.$$

Индукцией нетрудно показать, что X_n имеет длину, указанную в формулировке теоремы. Проследив за процессом построения X_n, можно убедиться, что всякое двухбуквенное подслово X_n имеет только 4 непересекающихся вхождения в X_n и, следовательно, при

$|y|=2$ имеем $|P_\tau(X/y)| \leqslant 8$. Аналогично можно убедиться в том, что для любого слова $y \in \Delta_n^*$ такого, что $|y| \geqslant 3$ выполняется неравенство

$$P_\tau(X_n/y) \leqslant \begin{cases} 6 & \text{при } |y|=3 \\ 8 & \text{при } |y|=4 \\ |y| & \text{при } |y|>4 \end{cases}$$

Из полученных оценок для длины проекции X_n на различные слова и теоремы 2 вытекает, что X_n несократимо. Теорема доказана.

7. Следующая теорема дает достижимую нижнюю оценку для $\tau(X)$

Теорема 6. Для всякого слова X

$$\tau(X) \geqslant 3\log_2|X|-2.$$

Для всякого i существует слово X такое, что $|X|>i$ и $\tau(X)=3\log_2|X|-2$.

Доказательство. Пусть G – оптимальная для X грамматика, имеющая приведенный вид, S – ее аксиома, $\{(A_1,\psi_1),\ldots,(A_\ell,\psi_\ell)\}$ – множество схем подстановок. Обращаясь к доказательству теоремы 1, нетрудно видеть, что вывод можно проводить так: применить к аксиоме S везде, где только возможно, пару (A_1,ψ_1), затем к результату применить везде, где это возможно, пару (A_2,ψ_2) и т.д. После применения последней пары (A_ℓ,ψ_ℓ) результатом будет X. Следовательно,

$$|S|\cdot|\psi_1|\ldots|\psi_\ell| \geqslant X.$$

Для всех натуральных a выполняется неравенство $a+2 \geqslant 3\log_2 a$. Следовательно,

$$\tau(X)=|S|+\sum_{i=1}^{\ell}(|\psi_i|+2) \geqslant 3\log_2|S| -$$

$$-2+\sum_{i=1}^{\ell}3\log_2|\psi_i| = 3\log_2(|S|\cdot|\psi_1|\ldots|\psi_\ell|) -$$

$$-2 \geqslant 3\log_2|X|-2.$$

Для доказательства второго утверждения теоремы рассмотрим слова X_i, определяемые равенством $X_i = a^{(4^i)}$. Построим грамматику, языком которой является X_i.

$$V_T = \{a\}. \quad V_N = \{a_0, a_1, \ldots, a_{i-1}\}$$

$$S = a_0 a_0 a_0 a_0$$

$$P = \{(a_0, a_1^4); (a_1, a_2^4); \ldots; (a_{i-1}, a^4)\}.$$

Нетрудно видеть, что

$$\tau(X_i) = |S| + |\bigcup P| = 4 + 6\cdot(i-1) = 3\cdot 2i - 2,$$

что и требовалось доказать.

Автор благодарен Ю.В.Матиясевичу и А.О.Слисенко за обсуждение работы и полезные советы.

ЛИТЕРАТУРА

I. Solomonoff R.I., A Formal Theory of Inductive Inference, "Information and Control", I964, 7, №I, I-22; 224-254.

2. Бонгард М.М., Проблема узнавания. М., I967.

3. Загоруйко Н.Г., Самохвалов К.Ф. Природа проблемы распознавания образов. В сб. "Вычислительные системы". I966 г. вып. 36, 3-I9.