О свойствах марковских регулярных языков

А. Б. Холоденко

В работе рассматривается обобщение n-граммной вероятностной языковой модели, широко применяемой в современных системах распознавания слитной речи, на случай бесконечных (регурярных) языков. Вводится понятие марковского языка, устанавливаются основные свойства таких языков.

Пусть $\mathfrak{A} = (A, Q, \varphi, Q_F, q_0)$ — конечный детерминированный автомат, A — входной алфавит, S — множество состояний, $Q_F \subseteq Q$ — множество финальных состояний, $\varphi : A \times Q \to Q$ — функция переходов, q_0 — начальное состояние автомата.

Введем несколько важных обозначений.

Через $\mathfrak{L}_{\mathfrak{A}} = \{\alpha \in A^* \,|\, \varphi(q_0, \alpha) \in Q_F\}$ обозначим язык, порождаемый автоматом $\mathfrak{A}$. В тех случаях, когда не возникает разночтений, индекс $\mathfrak{A}$ мы будем опускать.

Для натурального числа $s \in \mathbb{N}$ обозначим через $\mathfrak{L}(s)$ множество слов языка $\mathfrak{L}$ длины s:

$$\mathfrak{L}(s) = \{\alpha \in \mathfrak{L} : |\alpha| = s\}.$$

Через $\mathcal{P}\mathfrak{L}$ обозначим множество префиксов слов языка $\mathfrak{L}$, включая сами слова:

$$\mathcal{P}\mathfrak{L} = \{\alpha \in A^* \,|\, \exists \beta \in A^*, \alpha\beta \in \mathfrak{L}\},\ \mathfrak{L} \subseteq \mathcal{P}\mathfrak{L}.$$

Через $\mathfrak{L}_\gamma$ обозначим множество слов языка $\mathfrak{L}$, оканчивающихся на γ, то есть

$$\mathfrak{L}_\gamma = \{\alpha \in A^* \in \mathfrak{L} \,|\, \exists \beta \in A^*,\ \alpha = \beta\gamma\}.$$

Пусть $|w| = n$. Обозначим через $l_w(s)$ число слов языка $\mathfrak{L}$, имеющих с $(s-n+1)$-ой по s-ую букву подслово w, то есть

$$l_w(s) = |\mathcal{P}\mathfrak{L}_w(s)|.$$

В самом деле (см. рис. 1),

$$\alpha \in \mathcal{P}\mathfrak{L}_w(s) \Longleftrightarrow \alpha = \beta w \gamma \in \mathfrak{L},\ \beta, \gamma \in A^*,\ |\beta w| = s.$$

Рис. 1. $\alpha \in \mathcal{P}\mathfrak{L}_w(s)$.

Введём $G_w(s)$ — частоту встречаемости слова w на s-ом месте как

$$G_w(s) = \frac{l_w(s)}{\sum\limits_{|w'|=|w|} l_{w'}(s)}.$$

Через $G_w = \lim\limits_{s\to\infty} G_w(s)$ обозначим предельную частоту встречаемости слова w среди слов той же длины.

Пусть $w \in A^*$ — слово и $a \in A$ — буква, $|wa| = n$.

Введём величину $\Gamma_{w,a}(s)$ как

$$\Gamma_{w,a}(s) = \frac{l_{wa}(s)}{\sum\limits_{|w'|=|w|} l_{w'}(s)}.$$

Эта формула может быть переписана как

$$\Gamma_{w,a}(s) = \frac{G_{wa}(s)}{\sum\limits_{w'} G_{w'a}(s)} = \frac{G_{wa}(s)}{G_a(s)},$$

в случае, если не возникает деления на ноль.

В самом деле, множество слов длины s, кончающихся на u в языке $\mathcal{PL}$ в точности состоит из непересекающихся множеств слов, кончающихся на vu, при всех возможных v.

$$\mathcal{PL}_u(s) = \bigcup_{v,|vu|=n} \mathcal{PL}_{vu}(s).$$

Определение. Величину

$$\Gamma_{w,a}(s) = \lim_{s\to\infty} \Gamma_{w,a}(s),$$

если она существует, назовём n-граммой языка $\mathcal{L}$.

В случае $n = 1$ получаем униграмму $\Gamma_{\Lambda,a} = G_a$ в случае $n = 2$ получаем биграмму $\Gamma_{b,a}$ — встречаемость буквы b перед буквой a.

Определение. Язык $\mathcal{L}$ назовём марковским языком порядка n, если существуют все n-граммы $\Gamma_{w,a}$, где $|wa| = n$ и существуют все частоты G_v, где $|v| = n$.

Множество марковских языков порядка n обозначим через $\mathcal{M}(n)$. Через $\mathcal{M}$ обозначим класс марковских языков, то есть языков, являющихся марковскими при любом порядке n:

$$\mathcal{M} = \bigcap_{n=1}^{\infty} \mathcal{M}(n).$$

Легко увидеть, что, например, язык A^* является марковским. Покажем, что не всякий язык является марковским. В частности, справедливо следующее утверждение:

Утверждение 1. *Язык, порождённый автоматом $\mathfrak{A}$ (рис. 2) не является марковским языком порядка 1.*

Доказательство. В самом деле:

$$\mathcal{PL}_2(s) = (\{00, 01, 10, 11\}^* 2)(s).$$

$$l_2(s) = \begin{cases} 2^{s-1} & \text{если } s \text{ нечетно} \\ 0 & \text{если } s \text{ четно} \end{cases}$$

$$\mathcal{PL}_3(s) = (\{00, 01, 10, 11\}^* \{0, 1\} 3)(s).$$

$$l_3(s) = \begin{cases} 0 & \text{если } s \text{ нечетно} \\ 2^{s-1} & \text{если } s \text{ четно} \end{cases}$$

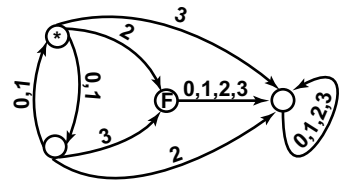

Рис. 2. Пример языка, не являющегося марковским.

$$\mathcal{PL}_0(s) = (\{0,1\}^*0)(s).$$
$$\mathcal{PL}_1(s) = (\{0,1\}^*1)(s).$$
$$l_0(s) = l_1(s) = 2^{s-1}.$$
$$\Gamma_{\Lambda,2}(s) = \begin{cases} \frac{1}{3} & \text{если } s \text{ нечетно} \\ 0 & \text{если } s \text{ четно} \end{cases}$$
$$\Gamma_{\Lambda,3}(s) = \begin{cases} 0 & \text{если } s \text{ нечетно} \\ \frac{1}{3} & \text{если } s \text{ четно} \end{cases}$$

Следовательно, $\Gamma_{\Lambda,2}$ и $\Gamma_{\Lambda,3}$ не существуют. Утверждение доказано.

1. Свойства введённых n-грамм

Установим теперь простейшие свойства n-грамм.

Утверждение 2 (аддитивность частот слева). *Если* $\mathcal{L} \in \mathcal{M}(n)$, *то для* $w_2 \in A^*$, $|w_2| = k < n$ *выполнено:*

$$\sum_{w_1,|w_1|=n-k} G_{w_1w_2} = G_{w_2}.$$

Доказательство. Заметим, что при $s > n$

$$\sum_{v,|vu|=n} l_{vu}(s) = l_u(s).$$

В самом деле, множество слов длины s, кончающихся на u в языке $\mathcal{PL}$ в точности состоит из непересекающихся множеств слов, кончающихся на vu, при всех возможных v.

$$\mathcal{PL}_u(s) = \bigcup_{v,|vu|=n} \mathcal{PL}_{vu}(s).$$

Если $\alpha \in \mathcal{PL}_u(s)$, то $\alpha = \alpha_1 vu$, где $|vu| = n$, поэтому $\alpha \in \mathcal{PL}_{vu}(s)$.

Если $\alpha \in \mathcal{PL}_{vu}(s)$, то $\alpha \in \mathcal{PL}_u(s)$. $\mathcal{PL}_{v_1 u}(s) \cap \mathcal{PL}_{v_2 u}(s) = \varnothing$. Получаем, что

$$\sum_{v,|v|=n-k} G_{vw_2}(s) = \sum_{v,|v|=n-k} \frac{l_{vw_2}(s)}{\sum\limits_{x,|x|=n} l_x(s)} = \frac{l_{w_2}(s)}{\sum\limits_{y,|y|=n-k} \left(\sum\limits_{z,|z|=k} l_y z(s) \right)} =$$

$$= \frac{l_{w_2}(s)}{\sum\limits_{z,|z|=k} \left(\sum\limits_{y,|y|=n-k} l_y z(s) \right)} = \frac{l_{w_2}(s)}{\sum\limits_{z,|x|=k} l_z(s)} = G_{w_2}(s).$$

Утверждение доказано.

Утверждение 3 (аддитивность n-грамм слева).

$$\Gamma_{v,a} = \sum_{u,|uv|=n} \Gamma_{uv,a}.$$

В самом деле, по предыдущему утверждению 2 для $|v| < n$ имеем:

$$G_v = \sum_{u,|uv|=n} G_{uv}.$$

Следовательно,

$$\Gamma_{v,a}(s) = \frac{l_{va}(s)}{\sum\limits_{v'} l_{v'a}(s)} = \frac{\sum\limits_{u} l_{uva}(s)}{\sum\limits_{u} \sum\limits_{v'} l_{xv'a}(s)} = \frac{\sum\limits_{u} l_{uva}(s)}{\sum\limits_{uv'} l_{uv'a}(s)} = \sum_{u} \Gamma_{uv,a}(s).$$

Что и требовалось показать.

Замечание. Аналогичные формулы для суммирования справа не верны.

В самом деле, заметим, что формула

$$\sum_{|w_2|=k} G_{w_1 w_2} = G_{w_1}$$

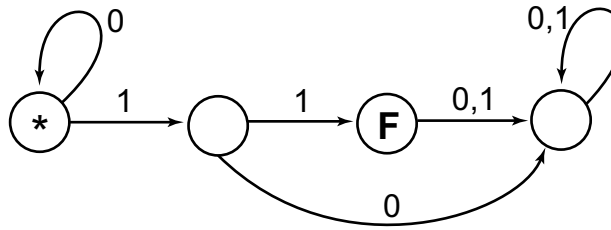

Рис. 3. Отсутствие аддитивности справа.

может не выполняться, так как слово αw_1 не всегда может быть продолжено до слова $\alpha w_1 w_2 \beta \in \mathfrak{L}$, $|w_1 w_2| = n$. Здесь исключены пустые буквы.

Например, автомат (рис. 3) порождает язык $\{0^n 11 | n \in \mathbb{N}\}$.

Здесь $G_0 = \frac{1}{3}$, $G_1 = \frac{2}{3}$, $G_{00} = \frac{1}{3}$, $G_{01} = \frac{1}{3}$, $G_{10} = 0$, $G_{11} = \frac{1}{3}$.

Таким образом, $G_{10} + G_{00} = G_0$ и $G_{01} + G_{11} = G_1$, однако $G_{01} + G_{00} \neq G_0$ и $G_{10} + G_{11} \neq G_1$.

2. Свойства марковских языков

Покажем несколько важных свойств марковских языков.

Теорема 1. *Если язык является марковским порядка n, то он также является марковским порядка $k < n$.*

Доказательство этого факта напрямую следует из утверждений 2 и 3, поскольку, как следует из этих утверждений, частоты и n-граммы языков более низкого порядка получаются суммированием слева соответственно частот и n-граммов языка более высокого порядка. Таким образом, все они существуют.

Замечание. Следует отметить, что обратное неверно. Например, существует язык, являющийся марковским первого порядка и не являющийся марковским второго порядка.

Доказательство. Рассмотрим автомат (рис. 4).

Пусть $X_i(t)$ — множество, а $x_i(t)$ — число слов ведущих из начального состояния 0 в состояния $i = 0, 1, 2, 3$. Выполнены следующие

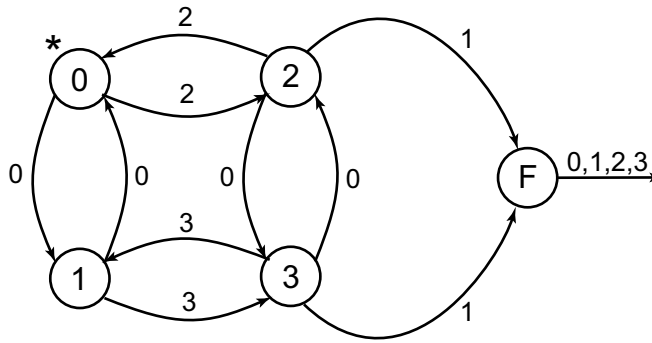

Рис. 4. Контрпример к обратному утверждению Теоремы 1.

рекуррентные соотношения.

$$\begin{aligned} x_0(t+1) &= x_1(t) + x_2(t); \\ x_1(t+1) &= x_0(t) + x_3(t); \\ x_2(t+1) &= x_0(t) + x_3(t); \\ x_3(t+1) &= x_1(t) + x_2(t), \end{aligned}$$

или, другими словами,

$$\begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}(t+1) = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}(t).$$

Пусть

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \text{ тогда } A^2 = \begin{pmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{pmatrix}$$

и $A^{2k+1} = 2^{2k}A$, $A^{2k} = 2^{2k-2}A^2$.

Получаем:

$$x_0(t) = x_3(t) \begin{cases} 0 & \text{если } t \text{ нечетно} \\ 2^{t-1} & \text{если } t \text{ четно} \end{cases}$$

$$x_1(t) = x_2(t) = \begin{cases} 2^{t-1} & \text{если } t \text{ нечетно} \\ 0 & \text{если } t \text{ четно} \end{cases}$$

$$l_0(t) = x_0(t) + x_1(t) + x_2(t) + x_3(t) = 2^t;$$

$$l_1(t) = x_2(t) + x_3(t) = 2^{t-1};$$

$$l_2(t) = x_0(t) + x_2(t) = 2^{t-1};$$

$$l_3(t) = x_1(t) + x_3(t) = 2^{t-1};$$

$$\Gamma_{\Lambda,0}(t) = \frac{2^t}{2^t + 3 * 2^{t-1}} = \frac{2}{2+3} = \frac{2}{5};$$

$$\Gamma_{\Lambda,1}(t) = \Gamma_{\Lambda,1}(t) = \Gamma_{\Lambda,1}(t) = \frac{2^{t-1}}{2^t + 3 * 2^{t-1}} = \frac{1}{2+3} = \frac{1}{5};$$

Таким образом, автомат на рисунке 4 порождает марковский язык порядка 1. Тем не менее,

$$\Gamma_{2,1}(t) = \frac{G_{21}(t)}{G_1(t)} = \frac{x_0(t)}{x_2(t) + x_3(t)} = \begin{cases} 0 & \text{если } t \text{ нечетно} \\ 1 & \text{если } t \text{ четно} \end{cases}$$

Следовательно, $\Gamma_{2,1}$ не существует. Пример на рис. 4 может быть обобщён на случай произвольного n.

В результате получаем следующее утверждение:

Теорема 2. *Для любого $n \in \mathbb{N}$ существует язык $\mathfrak{L}$, такой, что $\mathfrak{L} \in \mathfrak{M}_{n-1}$, но при этом $\mathfrak{L} \notin \mathfrak{M}_n$.*

Тем не менее, для любого фиксированного языка $\mathfrak{L}$ можно указать такой номер N, что из утверждения $\mathfrak{L} \in \mathfrak{M}(N)$ будет следовать, что $\mathfrak{L} \in \mathfrak{M}$.

Более точно, справедлива следующая теорема:

Теорема 3. *Если язык $\mathfrak{L} \in \mathfrak{M}(2^{|Q|})$, то $\mathfrak{L} \in \mathfrak{M}$, где Q — множество состояний автомата, задающего язык $\mathfrak{L}$.*

3. Число марковских языков

Утверждение 4. *Всякий сильносвязный автомат A порождает марковский язык с равномерными частотами встречаемости слов.*

Доказательство. Поскольку в сильносвязном автомате для любой пары состояний q_1 и q_2 существует слово β, переводящее автомат из состояния q_1 в состояние q_2, то есть такое, что $\varphi(q_1, \beta) = q_2$, получаем, что $\forall w \in A^*$, $a \in A$, $|wa| = n$

$$\mathcal{PL}_{wa}(s) = \{\alpha wa | \alpha \in A^{s-n}\}.$$

В самом деле: пусть слово αwa таково, что $\alpha \in A^{s-n}$. Тогда $\varphi(q_0, \alpha wa) = q_1$, и по только что написанному свойству это слово может быть продолжено до слова из $\mathfrak{L}$ некоторым словом β, поскольку для любого финального состояния $q_F \in Q$ существует такое слово β, что $\varphi(q_1, \beta) = q_F \in Q$. А это как раз и означает, что $\alpha wa\beta \in \mathfrak{L}$.

Отсюда следует, что $l_{wa}(s) = 2^{s-n}$, для любых w и a, поэтому

$$\Gamma_{w,a} = \frac{2^{s-n}}{|A|^{n-1}2^{s-n}} = \frac{1}{|A|^{n-1}},$$
$$G_v = \frac{l_v(s)}{\sum\limits_{|v'|=n} l_{v'}(s)} = \frac{2^{s-n}}{|A|^n 2^{s-n}} = \frac{1}{|A|^n}.$$

Таким образом, утверждение доказано.

В качестве следствия этого утверждения можно получить оценку числа марковских языков. Обозначим через $\mathcal{M}_N$ класс марковских языков, задаваемых автоматами не более чем с N состояниями. Через $\mathcal{R}_N$ обозначим класс всех регулярных языков, задаваемых автоматами не более чем с N состояниями. Тогда справедлива следующая теорема:

Теорема 4 (Оценка числа марковских языков). *Для достаточно больших N*

$$\frac{\mathcal{M}_N}{\mathcal{R}_N} > \left(1 - \frac{1}{e}\right).$$

4. Достаточное условие марковости

Пусть дан язык $\mathfrak{L}$ и задающий его автомат $\mathfrak{A}$.

Определение. Активным графом автомата $\mathfrak{A}$ назовём подмножество ребёр диаграммы переходов автомата $\mathfrak{A}$, которые входят хотя

бы в один путь из начальной вершины в одну из финальных вершин, а также инцидентные им вершины.

Определение. Активной матрицей автомата $\mathfrak{A}$ назовём матрицу инцедентности его активного графа.

Справедлива следующая теорема:

Теорема 5. *Если активная матрица автомата $\mathfrak{A}$ имеет единственное максимальное по модулю собственное значение, то задаваемый этим автоматом язык является марковским.*

Замечание. Следует отметить, что это достаточное условие марковости, но не необходимое.

В самом деле, рассмотрим произвольный автомат $\mathfrak{A}$, удовлетворяющий условиям теоремы. Рассмотрим теперь автомат $\mathfrak{A}'$, представляющий собой параллельное соединение двух копий автомата $\mathfrak{A}$. Очевидно, что языки, задаваемые автоматами $\mathfrak{A}$ и $\mathfrak{A}'$, совпадают. Однако у автомата $\mathfrak{A}'$ кратность всех собственных значений удвоилась по сравнению с исходным автоматом $\mathfrak{A}$. Так что у него существует два максимальных по модулю собственных числа, однако задаваемый им регулярный язык является марковским.

Список литературы

[1] Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. М.: Наука, 1985.

[2] Ахо А., Ульман Д. Теория синтаксического анализа, перевода и компиляции. Т. 1, 2. М.: Мир, 1978.

[3] Холоденко А. Б. О построении статистических языковых моделей для систем распознавания русской речи // Интеллектуальные системы. Т. 6. Вып. 1–4. 2002. С. 381–394.