Частотные регулярные языки

Д. Н. Бабин

Естественные языки обладают свойством постоянной частоты встречаемости букв и пар букв. В статье изучены регулярные языки с этим свойством.

Ключевые слова: естественный язык, регулярный язык, цепь Маркова, марковский язык.

Введение

Обработка текстов на естественном языке, распознавание образов заданных последовательностями букв, распознавание речи, оптическое распознавание печатных и рукописных символов, создание человеко-машинных интерфейсов и так далее требуют специализированных лингвистических и математических моделей, простейшими из которых являются специальные регулярные языки.

Еще в начале 20 века выдающимся русским ученым А. А. Марковым был создан аппарат цепей, впоследствии названных цепями Маркова, и опробован [1] на вычислении переходных вероятностей между соседними буквами (биграммами) в поэме А. С. Пушкина «Евгений Онегин». В дальнейшем этот аппарат получил широкое применение для распознавания и статистического моделирования естественных языков. Обобщением открытых Марковым зависимостей является n-граммная модель [2], которая до настоящего времени используется в большинстве современных коммерческих систем распознавания речи. Модель позволяет вычислить вероятность того, что слово $\alpha = a_{i_1}a_{i_2}\ldots a_{i_s}$ является допустимым словом языка. В модели делается допущение о том, что лишь ограниченной длины n предистория влияет на следующую букву, а именно

$$P(a_{i_j}|a_{i_1}a_{i_2}\ldots a_{i_{j-1}}) \approx P(a_{i_j}|a_{i_{j-n+1}}a_{i_{j-n+2}}\ldots a_{i_{j-1}}).$$

Это приводит к тому, что вероятность, расписанная в виде произведения условных вероятностей

$$P(a_{i_1}a_{i_2}\dots a_{i_s}) = P(a_{i_1}) \times P(a_{i_2}|a_{i_1}) \times \dots \times P(a_s|a_{i_1}a_{i_2}\dots a_{i_{s-1}}),$$

выражается через произведение вероятностей вида

$$P(a_{i_j}|a_{i_{j-n+1}}a_{i_{j-n+2}}\dots a_{i_{j-1}}).$$

Очевидно, что в таком случае модель сводится к конечному множеству вероятностей, каждую их которых можно оценить на этапе обучения системы, вычислив частоту встречаемости соответствующих слов в обучающей выборке. Это свойство естественных языков иметь в пределе фиксированные вероятности вида $P(a_{i_j}|a_{i_{j-n+1}}a_{i_{j-n+2}}\dots a_{i_{j-1}})$ может быть обобщено на регулярные языки. Получается описанный ниже класс регулярных языков с предельными частотными свойствами. Другим подходом в анализе языков является вложение изучаемого множества слов в язык с фиксированными частотами биграмм (2-грамм). Этот приём используется для проверки текстового множества на принадлежность естественному языку: а именно, частота пар букв (или ключевых слов) в естественном языке является фиксированной. Изучаемое множество слов может быть вложено в *биграммный язык* — множество *всех* слов с фиксированными частотами пар соседних букв. Оказывается, биграммные языки обладают уникальными свойствами. По матрице частот можно явно определить непустоту, мощность, конечность или бесконечность биграммного языка. Свойства биграммных языков связаны с эйлеровыми свойствами ориентированных гиперграфов, порожденными матрицами биграмм.

1. Регулярные языки с предельными частотными свойствами

Какие же свойства имеют регулярные языки, имеющие частотные свойства естественных языков. Перенесем понятия n-граммной модели на регулярные языки. Для этого определим частоту встречаемости слова w на s-ом месте, а затем рассмотрим предел этой частоты при s стремящемся к бесконечности.

Пусть $M = (A, Q, \varphi, Q_F, q_0)$ — конечный детерминированный автомат [3], A — входной алфавит, Q — множество состояний, $Q_F \subseteq Q$ — множество финальных состояний, $\varphi : A \times Q \to Q$ — функция переходов, q_0 — начальное состояние автомата. Через $L_M = \{\alpha \in A^* | \varphi(q_0, \alpha) \in Q_F\}$ обозначим язык, порождаемый автоматом M. Для натурального числа $s \in \mathbb{N}$ обозначим через $L(s)$ множество слов языка L длины s:

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

Через PL обозначим множество префиксов слов языка L, включая сами слова:

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

Через L_γ обозначим множество слов языка L, оканчивающихся на γ, то есть

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

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

$$l_w(s) = |PL_w(s)|.$$

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

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

Величину

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

если она существует, назовём n-граммой языка L для пары (w, a).

Язык 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 $ML(n)$. Через ML обозначим класс марковских языков, то есть языков, являющихся марковскими при любом порядке n:

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

Нетрудно показать, что в классе регулярных языков существуют языки, не являющиеся марковскими, и число марковских регулярных языков достаточно велико. Обозначим через ML_N класс марковских языков, задаваемых автоматами с не более чем N состояниями; через R_N обозначим класс всех регулярных языков, задаваемых автоматами с не более чем N состояниями. Справедливы теоремы 1–4 [4].

Теорема 1. *Для достаточно больших N*

$$\frac{ML_N}{R_N} > \left(1 - \frac{1}{e}\right).$$

Оказывается, что классы марковских языков строго вкладываются друг в друга. Это показывают теоремы 2 и 3.

Теорема 2. *Если язык является марковским порядка n, то он также является марковским порядка $k < n$.*

Теорема 3. *Для любого $n \in \mathbb{N}$ существует язык L, такой, что $L \in ML_{n-1}$, но при этом $L \notin ML_n$.*

Таким образом, марковские языки образуют строго сужающуюся последовательность:

$$ML(1) \supset ML(2) \supset ML(3) \supset \ldots \supset ML(n) \supset \ldots$$

С другой стороны, если язык L фиксирован, то ситуация становится обратной. А именно, справедлива

Теорема 4. *Пусть язык $L = L_M$ задан автоматом $M = \{A, Q, \varphi, Q_F, q_0\}$. Тогда из $L \in ML(2^{|Q|})$ следует, что $L \in ML$.*

2. Биграммные языки

Биграммой в алфавите A называется двухбуквенное слово $ab \in A^*$, $a, b \in A$. Обозначим через $\theta_\beta(\alpha)$ количество подслов β в слове α. Значение $\theta_\beta(\alpha)$ при данных β и α назовем кратностью β в слове α. По слову $\alpha \in A^*$ можно построить квадратную матрицу биграмм $(\Theta(\alpha))_{i,j=1}^{|A|}$ размера $|A| \times |A|$ такую, что на месте (i, j) матрицы будет стоять значение $\theta_{a_i a_j}(\alpha)$.

Обозначим через Ξ множество квадратных матриц размера $|A| \times |A|$, каждый элемент которых является целым неотрицательным числом.

Назовем языком $L(\Theta)$, порожденным матрицей $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу биграмм Θ, то есть $L(\Theta) = \{\beta | \Theta(\beta) = \Theta\}$. Построим по матрице $\Theta \in \Xi$ ориентированный гиперграф G_Θ, вершинами которого будут буквы из алфавита A, при этом ребра будут соответствовать биграммам с учетом их кратностей, то есть θ_{ab} будет порождать θ_{ab} ориентированных ребер $a \to b$, а θ_{cc} будет порождать θ_{cc} петель $c \to c$.

Связный ориентированный гиперграф называется эйлеровым [5], если у всех вершин количество входящих ребер равно количеству исходящих ребер. Граф называется почти эйлеровым, если у всех вершин, кроме двух, количество входящих ребер равно количеству исходящих ребер, а у оставшихся двух вершин разность количества входящих ребер и количества исходящих ребер равна $+1$ и -1 соответственно.

Мы можем рассматривать матрицу биграмм как матрицу пропорций. Получается язык, F_Θ, в котором сохраняются отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)$ $\forall a, b, c, d \in A$, $\theta_{cd}(\alpha) > 0$ для любого слова α из этого языка.

$$F_\Theta = \bigcup_{k=1}^{\infty} L(k\Theta),$$

Имеют место теоремы 5–6 [6].

Теорема 5. *Для алфавита $A = \{0, 1\}$ и матрицы биграмм $\Theta \in \Xi$, задающей эйлеров или полуэйлеров граф число слов N_Θ с матрицей биграмм Θ равно*
1. $C_{\theta_{11}+\theta_{10}}^{\theta_{11}} C_{\theta_{00}+\theta_{10}}^{\theta_{00}}$; при $\theta_{01} > \theta_{10}$

2. $C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}}$*; при* $\theta_{01} < \theta_{10}$
3. $C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}}\right)$*; при* $\theta_{01} = \theta_{10}$
(здесь под C_n^k *понимается число сочетаний из* n *по* k*, то есть* $C_n^k = \frac{n!}{k!(n-k)!}$*).*

Теорема 6. *Если* G_Θ *эйлеровый граф, то* F_Θ *счетно, если* G_Θ *полуэйлеровый граф, то* F_Θ *конечно, в остальных случаях* F_Θ *пусто.*

Список литературы

[1] Марков А. А. Исчисление вероятностей. — СПб.: Типография Императорской Академии Наук, 1913.

[2] Bahl L. R., Baker J. K., Jelinek F., Mercer R. L. Perplexity a measure of the difficulty of speech recognition tasks. / Program of the 94$^{\text{th}}$ Meeting of the Acoustical Society of America // J. Acoust. Soc. Am. — 1977. Suppl. no. 1. Vol. 62. — P. S63.

[3] Кудрявцев В. Б., Алёшин С. В., Подколзин А. С. Введение в теорию автоматов. — М.: Наука, 1985.

[4] Бабин Д. Н., Холоденко А. Б. Об автоматной аппроксимации естественных языков // Интеллектуальные системы. — Т. 12, вып. 3–4. 2008. — С. 125–136.

[5] Оре О. Теория графов. — М.: Наука, 1980.

[6] Петюшко А. А. Частотные языки // Интеллектуальные системы в производстве. — 2012. № 1. — С. 192–201.