Об автоматной аппроксимации естественных языков

Д. Н. Бабин, А. Б. Холоденко

В статье даётся краткий обзор математических моделей естественных языков, в том числе и работ авторов в этом направлении. Во второй части статьи определены автоматные языки, имеющие предельные частотные свойства и сформулирован ряд теорем про них.

Введение

Обработка текстов на естественном языке включает поиск текстовой информации в больших и сверхбольших базах данных и знаний; автоматическое рубрицирование текстов; построение интеллектуальных вопрос-ответных систем, способных отвечать на наиболее типичные вопросы пользоватлей; автоматический перевод текстов с одного языка на другой; генерацию текстов на заданную тему; аннотирование и реферирование текстов; распознавание речи; оптическое распознавание печатных и рукописных символов; создание человеко-машинных интерфейсов и так далее. Все эти области требуют специализированных лингвистических и математических моделей, позволяющих представлять синтаксис и семантику текста в удобном для автоматической обработке виде.

Исторически, одной из первых задач, потребовавших построения довольно сложной модели естественного языка, была задача автоматического перевода, впервые сформулированная американцами А. Бутом и У. Уивером в 1946 году. Работы над системами автоматического перевода стимулировали развитие ряда языковых моделей,

которые в дальнейшем нашли своё место во многих областях вычислительной техники. К примерам таких моделей можно отнести иерархию грамматик Хомского [1] (в первую очередь — теорию автоматов [2] и теорию контекстно-свободных грамматик [3]); грамматики Вудса [4]; вероятностные автоматы [5]; грамматики зависимостей [6]; модели «Смысл-текст» [7] и многие другие.

Изучение подобных моделей привело к созданию развитой теории формальных языков, в рамках которой были сформулированы чисто математические задачи, такие как проблема принадлежности слова языку, заданному формальной грамматикой; проблема нахождения пересечения двух языков; проблема сложности описания языка и так далее.

Параллельно с развитием систем автоматического перевода и общения на естественном языке в начале 60-х годов двадцатого века начались исследования по созданию систем речевого общения, включающих в себя, помимо остальных блоков, также блок распознавания речи. До тех пор, пока объёмы словарей подобных систем не превосходили порога в несколько сотен слов, эти системы могли строиться без учёта каких бы то ни было моделей языка. В том случае, если система обладала большим словарём, удовлетворительной работы без учёта особенностей языка добиваться уже не удавалось. Это также подстегнуло исследования по моделированию естественного языка.

К настоящему моменту накоплено значительное количество различных подходов к формализации естественного языка, сформулированных в виде математических конструкций. Значительная доля этих конструкций хорошо изучена, однако «универсальной» модели, которая могла бы очень точно аппроксимировать реальный язык и оказалась бы идеально адаптированной к различным задачам, до сих пор создать не удалось. Также остаётся открытым вопрос о построении модели, оценивающей «естественность» текста. Подобная модель нашла бы очень широкое применение в различных системах, начиная с оптимизации работы поисковых систем в сети интернет и заканчивая улучшением качества работы систем распознавания речи.

В области распознавания речи наибольшее распростронение получили вероятностные языковые модели. Самая простая из них — так

называемая n-граммная модель [8] до настоящего времени используется в большинстве современных коммерческих систем распознавания речи.

Для значительной части романских и германских языков, а также для ряда азиатских языков (например, китайского и японского) в настоящее время уже разработаны коммерческие системы распознавания речи. Удачной коммерческой системы для русского языка до сих пор не существует. Одной из основных причин этого является отсутствие эффективной модели для представления русского языка. Поэтому любые исследования в этой области являются актуальными. Кроме того, большинство работ, посвященных вопросам построения языковых моделей для систем распознавания речи, имеют ярко выраженную «инженерную» направленность.

Превосходство человека над компьютерами в точности распознавания речи в первую очередь обусловлено именно учётом человеком контекста высказывания (в том числе и смысла) и умением отличить правильно построенное предложение от неправильного.

Поскольку в задачах распознавания речи обычно приходится иметь дело не с линейной последовательностью букв в слове, а с деревом вариантов распознавания, используемые здесь модели должны обладать некоторыми особенностями. А именно, они должны иметь достаточно высокую скорость работы, чтобы справляться с экспоненциальным взрывом количества вариантов распознавания, а также допускать возможность работы «слева направо», то есть позволять на ранних этапах отсекать те варинаты распознавания, которые не могут быть продолжены до правильного предложения естественного языка.

В настоящее время в промышленных и экспериментальных системах распознавания используются как дискретные, так и вероятностные модели. К наиболее часто используемым дискретным моделям относятся регулярные языки, контекстно-свободные языки, а также системы, основанные на использовании лингвистических экспертных систем и систем понимания и учёта смысла. Вероятностные модели включают в себя n-граммы, системы, основанные на деревьях решений, и вероятностные обобщения контекстно-свободных грамматик.

Каждый из вышеупомянутых подходов имеет свои преимущества и недостатки. Более подробно они рассмотрены в работах [9] и [10]. Общепринятой на сегодняшний день оценкой качества модели является так называемый *коэффициент неопределённости* (в англоязычной литературе используется термин *«perplexity coefficient»* [11]). В случае n-граммных моделей, которые и будут преимущественно изучаться в дальнейшем в рамках данной работы, коэффициент неопределённости показывает среднюю степень ветвления в модели, то есть сколькими способами в среднем может быть продолжено фиксированное начало предложения.

В настоящее время в задачах распознавания слитной речи чаще всего используются вероятностные модели, построенные на принципе независимости от «далёкой» истории, так называемые n-граммные модели.

Если в общем виде вероятностная модель позволяет вычислить вероятность того, что слово $\alpha = a_{i_1}a_{i_2}\dots a_{i_s}$ является допустимым словом языка, то в n-граммной модели делается допущение о том, что

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

Это приводит к тому, что вероятность $P(a_{i_1}a_{i_2}\dots a_{i_s})$, расписанная в виде произведения условных вероятностей

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

может быть оценена 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.

Известно, что прямой перенос моделей, пригодных для построения систем распознавания английской речи, на случай русского распознавателя невозможен, поскольку приводит к слишком громоздким конструкциям, которые невозможно использовать в рамках существующих на сегодняшний день компьютерных технологий [12].

В работе [13] авторами предложен механизм адаптации стандартного n-граммного подхода, который позволяет использовать формализм n-грамм для построения русского распознавателя.

Как уже было отмечено, для моделирования русского языка такие модели подходят плохо. В работе [14] подробно описано обобщение n-граммной модели, которое позволило распространить аппарат n-граммных моделей и на русский язык. Главными трудностями в русском языке на пути создания таких моделей являются большое количество словоформ (что приводит к серьёзному увеличению словарей системы) и относительно свободный порядок слов. Основной особенностью предложенного подхода является декомпозиция общей n-граммной модели в декартово произведение двух моделей: модели, построенной на леммах слов, и модели, построенной на морфологической информации.

Модель, построенная на морфологических классах, содержит словарь порядка 550 «слов» и позволяет решать задачи, связанные с моделированием морфологического строения предложения, в том числе, например, снятие омонимии. В то же время модель, построенная на леммах, отвечает за представление смысловой составляющей языка и по своим качественным характеристикам примерно соответствует аналогичным моделям для английского языка. Словарь системы составляет приблизительно 130 тыс. лемм.

Каждая из этих моделей была построена и обучена на материале российской периодики. При этом было показано, что полученные характеристики языковых моделей примерно соответствуют среднестатистическим характеристикам моделей для английского языка (коэффициент неопределённости в триграммной модели, основанной на леммах, составил около 230; для моделей на английском языке этот коэффициент обычно оказывается в районе 100). Коэффициент неопределённости в категорной триграммной модели (построенной исключительно на морфологической информации) оказался равен 20.

Результаты приведённых в работе [14] исследований показывают, что существующие n-граммные подходы могут быть адаптированы для работы с русским языком. Поэтому исследование свойств подоб-

ных вероятностных моделей является важной задачей и с точки зрения создания полноценной системы распознавания слитной русской речи.

Регулярные языки с предельными частотными свойствами

К сожалению, несмотря на то, что, как уже было отмечено выше, наибольшее распространение в мире получили именно n-граммные модели, их формальные математические свойства исследованы довольно мало. Поэтому авторами была предпринята попытка провести более детальный анализ свойств n-грамм (см., например, [15]). Для этого используется аппарат теории автоматов и регулярных языков.

Мы построим обобщение понятия n-граммной модели на бесконечные формальные языки. Для этого в начале вводится частота встречаемости слова w на s-ом месте, а затем рассматривается предельная частота встречаемости слова w как предел при s стремящемся к бесконечности.

Более точно:

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:

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

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

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

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

$$l_w(s) = |\mathcal{PL}_w(s)|.$$

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

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

Определение. Величину

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

если она существует, назовём n-граммой языка $\mathcal{L}$ для пары (w, 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:

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

Нетрудно показать, что в классе регулярных языков существуют языки, не являющиеся марковскими, поэтому выделение и изучение подкласса марковских регулярных языков оказывается оправданным. Тем не менее, число марковских регулярных языков достаточно велико.

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 1 (Оценка числа марковских языков). *Для всякого натурального числа N*

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

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 для любого $k < n$.*

Теорема 3. *Для любого $n \in \mathbb{N}$ существует язык $\mathfrak{L}$, такой, что $\mathfrak{L} \in \mathcal{M}(n-1)$, но при этом $\mathfrak{L} \notin \mathcal{M}(n)$.*

Таким образом, марковские языки образуют строго сужающуюся последовательность:

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

С другой стороны, если язык $\mathfrak{L}$ фиксирован, то цепочка вложений для него обрывается на конечном шаге и становится возможным установить его принадлежность к классу марковских языков за конечное число шагов, а именно, справедлива

Теорема 4. *Если язык $\mathfrak{L}_{\mathfrak{A}}$ задан автоматом $\mathfrak{A} = \{A, Q, \varphi, Q_F, q_0\}$, то из $\mathfrak{L}_{\mathfrak{A}} \in \mathcal{M}(2^{|Q|})$ следует, что $\mathfrak{L}_{\mathfrak{A}} \in \mathcal{M}$.*

Любая n-грамма может быть вычислена по диаграмме переходов автомата, однако это требует умения находить собственные числа для матриц большой размерности.

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 инцедентные им вершины.

Определение. Активной матрицей автомата $\mathfrak{A}$ назовём матрицу инцедентности его активного графа (взятую с учетом кратностей ребер).

Справедлива следующая теорема, дающая пример достаточного условия марковости:

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

Теорема 5 допускает конструктивное доказательство, то есть предлагает детерминированную процедуру, позволяющую вычислить произвольную n-грамму для языка $\mathfrak{L}$. В соответствии с этой процедурой, искомая n-грамма либо будет вычислена, либо будет доказано, что такой n-граммы не существует. В соответствии с **Теоремой 4** для установления принадлежности языка $\mathfrak{L}$ к классу марковских языков достаточно проверить только существование n-грамм при $n = 2^{|Q|}$, где Q — множество состояний задающего язык $\mathfrak{L}$ автомата. Таким образом, установление принадлежности произвольного языка $\mathfrak{L}$ классу марковских языков может быть получено за конечное число шагов.

Оказывается, что класс марковских языков не замкнут относительно основных теоретико-языковых операций: объединения, пересечения и дополнения, поэтому имеет смысл ввести в рассмотрение более узкие классы языков. Примером такого класса является класс *каскадно-дефинитных языков*.

Понятие дефинитных языков (то есть таких языков, функция переходов которых «забывает» далёкую предысторию) является прямым переносом идеологии марковских языков непосредственно на

теорию автоматов. Однако это свойство оказывается слишком жестким: любой дефинитный язык является марковским языком и все его n-граммы (для любого фиксированного значения n) равны между собой. Поэтому сами по себе дефинитные языки не очень интересны в свете рассмотрения марковских языков. Тем не менее, из них можно получить новый класс языков, с одной стороны небольшой (его доля среди всех языков, задаваемых автоматами с N состояниями, стремится к нулю с ростом N), а с другой стороны — в некотором смысле «всюду плотный» в множестве марковских языков.

Пусть $\mathfrak{A}^1 = (A, Q^1, \varphi^1, Q^1_F, q^1_0)$, $\mathfrak{A}^2 = (A, Q^2, \varphi^2, Q^2_F, q^2_0)$. Пусть также $q^1 \in Q^1$ и $q^2 \in Q^2$. Введём операцию склейки двух автоматов по паре состояний (q^1, q^2).

Определение. Результатом склейки автоматов $\mathfrak{A}^1$ и $\mathfrak{A}^2$ называется автомат

$$\mathfrak{A} = (A, Q^1 \cup Q^2 \setminus \{q^1\}, \varphi, Q^1_F \cup Q^2_F \setminus \{q^1\}, q^1_0), \text{ где}$$

$$\varphi(q,a) = \begin{cases} \varphi^1(q,a) & \text{если } q \in Q^1 \setminus \{q^1\} \text{ и } \varphi^1(q,a) \neq q^1 \\ q^2 & \text{если } q \in Q^1 \setminus \{q^1\} \text{ и } \varphi^1(q,a) = q^1 \\ \varphi^1(q^2,a) & \text{если } q = q^1 \\ \varphi^2(q,a) & \text{если } q \in Q^2. \end{cases}$$

Каскадно-дефинитные языки получаются из класса дефинитных языков рекурсивно путём применения операции склейки двух автоматов по паре состояний.

Оказывается, введённая таким образом операция склейки двух автоматов по паре состояний позволяет получать автоматы с заданными свойствами из набора простейших автоматов: циклов и отрезков.

Сформулируем сначала следующее утверждение.

Утверждение 1. *Множество $\{G_{ab}\}$ является системой биграмм для некоторого регулярного языка $\mathfrak{L}$ тогда и только тогда, когда для любого i, $1 \leqslant i \leqslant |A|$ выполнено:*

$$\Sigma_i \geqslant M_i, \tag{1}$$

где Σ_i — сумма по i-ому столбцу, M_i — максимум по i-ой строке в матрице переходов для автомата, задающего язык $\mathfrak{L}$.

Определение. Условие (1) будем называть условием биграммности множества $\{G_{ab}\}$, а соответствующую ей матрицу π будем называть биграммной матрицей.

Теорема 6. *Для всякой* ***рациональной*** *биграммной матрицы π найдётся автомат $\mathfrak{A}$, матрица биграмм которого $\pi_{\mathfrak{A}}$ будет в точности совпадать с исходной биграммной матрицей π, и который может быть получен из «простейших» автоматов — отрезков и циклов путём применения к ним операции склейки автоматов.*

В том случае, если биграммная матрица автомата $\mathfrak{A}$ содержит иррациональные числа, то для любого $\varepsilon > 0$ её можно приблизить рациональной биграммной матрицей и построить автомат, имеющий в точности заданную рациональную биграммную матрицу.

Таким образом, для произвольной биграммной матрицы может быть построен автомат $\mathfrak{A}'$, имеющий биграммную матрицу, сколь угодно близкую к данной. При этом следует отметить, что Теорема 6 даёт конструктивный способ построения такого автомата.

Список литературы

[1] Хомский Н. Синтаксические структуры // Новое в лингвистике. Вып. II. М., 1962.

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

[3] Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. М.: Мир, 1978.

[4] Вудс В. А. Сетевые грамматики для анализа естественных языков // Кибернетический сборник. Новая серия. Вып. 13. М.: Мир, 1978. С. 120–158.

[5] Бухараев Р. Г. Основы теории вероятностных автоматов. М.: Наука, 1985.

[6] Sleator D., Temperley D. Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU–CS–91–196, October 1991.

[7] Мельчук И. А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М.: Наука, 1974.

[8] EAGLES. «HANDBOOK of Standards and Resources for Spoken Language Systems». Mouton de Gruyter, 1997.

[9] Холоденко А. Б. О языковых моделях для систем распознавания русской речи // Интеллектуальные системы в производстве: Периодический научно-практический журнал. № 1. Ижевск: Изд-во ИжГТУ, 2003. С. 146–155.

[10] Бабин Д. Н., Мазуренко И. Л., Холоденко А. Б. О перспективах создания системы автоматического распознавания слитной устной русской речи // Интеллектуальные системы. Т. 8. Вып. 1–4. 2004. С. 45–70.

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

[12] Kanevsky D., Monkowsky M., Sedivy J. Large Vocabulary Speaker-Independent Continuous Speech Recognition in Russian Language. Proc. SPECOM'96, St.-Petersburg, October 28–31, 1996.

[13] Kholodenko A. To the creating of the language models for Russian // V International Congress on mathematical modeling. September 30 – October 6, 2002. Dubna, Moscow Region. Book of abstracts. V. 2, M.: «Janus-K», 2002. P. 97.

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

[15] Холоденко А. Б. О марковских регулярных языках // Материалы IX Международного семинара «Дискретная математика и её приложения». 18–23 июня 2007 года. М.: Изд-во механико-математического факультета МГУ, 2007. С. 358–361.