Московский государственный университет
имени М.В. Ломоносова
Механико-математический факультет

На правах рукописи
УДК 519.766

Холоденко Александр Борисович

ОБ АВТОМАТНОЙ АППРОКСИМАЦИИ РЕАЛЬНЫХ ЯЗЫКОВ

01.01.09 - дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

МОСКВА - 2008

Работа выполнена на кафедре Математической теории интеллектуальных систем Механико-математического факультета Московского государственного университета имени М.В. Ломоносова.

Научный руководитель: доктор физико-математических наук, профессор Бабин Дмитрий Николаевич.

Официальные оппоненты: доктор физико-математических наук, профессор Чуличков Алексей Иванович; кандидат физико-математических наук доцент Карташов Сергей Иванович.

Ведущая организация: Московский энергетический институт (технический университет).

Защита диссертации состоится 6 июня 2008 года в 16 часов 40 минут на заседании диссертационного совета Д.501.001.84 при Московском государственном университете им. М.В. Ломоносова по адресу 119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, Московский государственный университет им. М. В. Ломоносова, Механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в библиотеке Механико-математического факультета МГУ им. М.В. Ломоносова (Главное здание, 14 этаж).

Автореферат разослан 6 мая 2008 года.

Ученый секретарь
диссертационного совета
Д.501.001.84 при МГУ
доктор физико-математических наук,
профессор А.О. Иванов

Общая характеристика работы

Актуальность темы

Естественный язык - это основной инструмент коммуникации людей, поэтому неудивительно, что изучается он с древнейших времён. Однако с появлением и развитием вычислительной техники встала задача «обучить» компьютер естественному языку, то есть построить чёткие и эффективные модели, способные решать различные задачи, требующие умения обрабатывать тексты на естественном языке.

Среди современных задач, связанных с работой с текстами на естественном языке, можно выделить поиск текстовой информации в больших и сверхбольших базах данных и знаний; автоматическое рубрицирование текстов; построение интеллектуальных вопрос-ответных систем, способных, например, отвечать на наиболее типичные вопросы пользоватлей; автоматический перевод текстов с одного языка на другой; генерация текстов на заданную тему (например, всевозможных отчётов); аннотирование и реферирование текстов; распознавание речи; оптическое распознавание печатных и рукописных символов; создание человеко-машинных интерфейсов с использованием естественного языка 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

[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. и 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]Daniel Sleator and Davy Temperley. 1991. 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.

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, однако в области распознавания речи наибольшее распространение 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.

Таким образом, к настоящему моменту в науке 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.

Представленная работа также рассматривает вопросы моделирования естественного языка в задачах, связанных с обработкой и анализом различной информации. Основная часть диссертации посвящена вопросам построения языковых моделей, предназначенных для использования в системах распознавания русской речи, а также изучению их математических свойств.

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

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

представления русского языка в системе распознавания. 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. В представленной работе предпринята попытка не только построить адекватную и практически применимую модель для русского языка, но и изучить её математические свойства, получить новый инструментарий для разработки и анализа вероятностых языковых моделей для систем распознавания речи.

Цель работы

Цели настоящей работы:

- решить задачу исправления ошибок в формальных языках;
- изучить частотные свойства естественных языков;
- определить и изучить классы формальных языков, близких к естественным;
- обобщить эвристические свойства n-граммных моделей на формальные языки.

Методы исследования

В диссертации использованы методы теории автоматов, теории формальных языков, комбинаторики, теории графов и математического анализа.

Научная новизна и ценность работы

Работа носит теоретический характер. Основные результаты диссертации являются новыми и состоят в следующем:

1. Решена задача исправления ошибок в формальных языках. Построен алгоритм, решающий задачу принадлежности слова формальному языку при наличии в нем ошибок разных типов. Найдены полиномиальные относительно длины слова алгоритмы проверки принадлежности слова регулярному или контекстно-свободному языку при условии искажения анализируемых слов.

2. n-Граммная модель успешно модернизирована для работы с русским языком.

3. Введено понятие регулярных марковских языков и марковских автоматов и изучены их свойства.

4. Предложен алгоритм аппроксимации марковского автомата при помощи простейших автоматов, доля которых среди всех автоматов с тем же числом состояний стремится к нулю.

Указанные результаты могут быть полезны специалистам, занимающимся теорией автоматов и теорией формальных языков.

Помимо чисто теоретической значимости, полученные в диссертации результаты также имеют следующие важные практические приложения, которые могут быть полезны специалистам, занимающимся работой с естественными языками:

1. Введено понятие обобщенного слова как ориентированного информационного графа без (ориентированных) циклов с отметками в виде букв и весов на ребрах. При потенциально экспоненциальной мощности множества слов, представляемых обобщенным словом, построены полиномиальные алгоритмы нахождения пути в обобщенных словах из регулярных ($O(n^2)$) и контекстно-свободных языков ($O(n^3)$), где n является длиной обобщенного слова и не зависит от используемой грамматики. Данные алгоритмы находят применение в задачах коррекции результатов распознавания и в задачах проверки правописания.

2. Построена вероятностная языковая модель для русского языка, пригодная к использованию в составе системы распознавания слитной речи.

Апробация работы

Результаты диссертации докладывались на следующих семинарах механико-математического факультета МГУ им. М.В.Ломоносова и научных конференциях:

1. Семинар «Теория автоматов» под руководством академика, профессора, д.ф-м.н. В.Б. Кудрявцева (2002 – 2008гг.).

2. Семинар «Теория дискретных функций и приложения» под руководством профессора, д.ф-м.н. Д.Н. Бабина (1999 – 2008 гг.).

3. Международная конференция «Информационные технологии в инновационных проектах» (Ижевск, 1999).

4. Международный семинар «Диалог'99» по компьютерной лингвистике и ее приложениям (Таруса, 1999).

5. IV Всероссийская конференция «Нейрокомпьютеры и их применение» (Москва, 2000).

6. V International Congress on mathematical modeling, (Dubna, 2002).

7. IX Международный семинар «Дискретная математика и её приложения» (Москва, 2007).

Публикации

Основное содержание диссертации опубликовано в 7 работах автора, список которых приведен в конце автореферата [1-7].

Структура и объем диссертации

Диссертация состоит из введения, пяти глав и списка литературы. Текст диссертации изложен на 99 страницах. Список литературы содержит 45 наименований.

Содержание работы

Во **введении** даётся общая характеристика работы, а также приводятся основные понятия и результаты, изложенные в диссертации.

В **первой главе** приводится обзор существующих в настоящее время языковых моделей, описываются их сильные и слабые стороны, изучается их пригодность к использованию в лингвистическом обеспечении систем распознавания слитной речи. В этой главе рассматриваются как дискретные, так и вероятностные подходы, демонстрируются преимущества и недостатки каждого из них.

Известно, что подавляющее 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. Поэтому важность создания хорошей модели языка для систем распознавания речи трудно переоценить.

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 этой главе рассматриваются:

- **дискретные модели:**
  - *регулярные языки;*
  - *контекстно-свободные языки;*
  - *системы, основанные на использовании лингвистических экспертных систем и систем понимания и учёта смысла;*
- **вероятностные модели:**
  - *n-граммы;*
  - *системы, основанные на деревьях решений;*
  - *вероятностные обобщения контекстно-свободных грамматик.*

Здесь показаны преимущества и недостатки каждого из этих подходов, а также приведена общепринятая на сегодняшний день оценка качества модели, так называемый *коэффициент неопределённости (в англоязычной литературе используется термин «perplexity coefficient»*[9]*).* 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 случае невозможности или нежелательности проводить полный эксперимент с использованием системы распознавания, коэффициент неопределённости позволяет сравнивать между собой две модели, а также, в случае использования одинаковых моделей, сравнивать относительную сложность языков.

[9]L. R. Bahl, J. K. Baker, F. Jelinek, and R.L . Mercer. 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.

Во **второй главе** более подробно рассматривается вопрос использования дискретных моделей, в первую очередь – регулярных и контекстно-свободных языков. Для обоих этих случаев предложены алгоритмы анализа, пригодные для работы в составе систем распознавания речи, которые характеризуются необходимостью анализа не одного, а большого множества вариантов входного слова. В работе это моделируется введением т.н. обобщенного слова, которое представляет дерево вариантов результатов распознавания. Для всех алгоритмов вычислены оценки их сложности, а также приведены примеры их использования в различных задачах.

Определение. ***Обобщенным словом*** *над алфавитом A назовем связный ориентированный граф без ориентированных циклов, каждому ребру которого приписана некоторая буква из алфавита A и неотрицательное вещественное число.*

В каждом таком графе есть источник (вершина без входящих рёбер) и сток (вершина без исходящих рёбер). Они называются начальной и конечной вершиной обобщенного слова соответственно.

Задача формулируется следующим образом.

Пусть дано обобщённое слово α и некоторая грамматика Γ. *Требуется найти в обобщённом слове путь, ведущий из начальной вершины обобщенного слова в конечную и обладающий двумя свойствами:*

1. *слово, полученное конкатенацией букв, записанных вдоль этого пути, должно быть допустимым словом в грамматике* Γ*;*

2. *сумма весов, стоящих на рёбрах пути, должна быть минимальна.*

В работе показано, что в случае, если грамматика Γ задаётся конечным автоматом, данная задача может быть решена за время $O(n^2)$, где n - количество вершин в графе обобщённого слова. В том случае, если грамматика Γ является контекстно-свободной, то данная задача может быть решена за время $O(n^3)$, где n - количество вершин в графе обобщённого слова.

Описанные во второй главе работы алгоритмы также позволяют несколько обобщить задачу, а именно - перечислить *все* пути, удовлетворяющие первому условию, либо найти такой путь, который может быть представлен в виде конкатенации произвольного числа слов, допустимых в грамматике Γ.

Кроме того, в данной главе приведены два примера применения построенной техники: в задаче коррекции результатов оптического распознавания символов и в задаче поиска минимального исправления ошибочно написанного слова (так называемая задача spellchecker'a).

Последняя задача решается при помощи моделирования операций пропуска, вставки и замены буквы в обобщённом слове и применения получен-

ных алгоритмов для нахождения оптимального пути, то есть пути с наименьшим суммарным штрафом.

В **третьей главе** рассматривается статистический подход к построению языковых моделей. 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[10]. В этой главе предложен механизм адаптации стандартного n-граммного подхода, который позволяет использовать механизм n-грамм для построения русского распознавателя.

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 $\alpha = a_{i_1}a_{i_2}\ldots a_{i_s}$ является допустмым словом языка, то в n-граммной модели делается допущение о том, что

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

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

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

может быть оценена через произведение вероятностей вида $P(a_{i_j}|a_{i_{j-n+1}}a_{i_{j-n+2}}\ldots a_{i_{j-1}})$. 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.

Как уже было отмечено, для моделирования русского языка такие модели подходят плохо. В этой главе предложено 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.

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

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 модели, основанной на леммах, составил около 230; для моделей на английском языке этот коэффициент обычно оказывается в районе 100). Коэффициент неопределённости в категорной модели (построенной исключительно на морфологической информации) оказался около 20.

Таким образом, в этой главе показано, что существующие n-граммные подходы могут быть адаптированы для работы с русским языком. Поэтому исследование свойств подобных вероятностных моделей является важной задачей и с точки зрения создания полноценной системы распознавания слитной русской речи.

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 следующие две главы посвящены более детальному анализу свойств n-грамм. Для этого используется аппарат теории автоматов и регулярных языков.

В **четвёртой главе** вводится обобщение понятия n-граммной модели на бесконечные формальные языки. А именно, вводится частота встречаемости слова w на s-ом месте, а затем рассматривается предельная частота встречаемости слова w как предел при $s \to \infty$.

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:

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

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

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

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 $\mathfrak{L}$ для пары (w, a).

Определение. *Язык $\mathfrak{L}$ 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 $\mathfrak{M}(n)$. Через $\mathfrak{M}$ обозначим класс марковских языков, то есть языков, являющихся марковскими при любом порядке n:

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

Показано, 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 4.4 (Оценка числа марковских языков)
Для достаточно больших 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 **Теорема 4.1** и **Теорема 4.2**.

Теорема 4.1.
Если язык является марковским порядка n, то он также является марковским порядка $k < n$.

Теорема 4.2.
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 ситуация становится обратной. А именно, справедлива

Теорема 4.3.
Пусть язык $\mathfrak{L} = \mathfrak{L}_{\mathfrak{A}}$ задан автоматом $\mathfrak{A} = \{A, Q, \varphi, Q_F, q_0\}$. Тогда из $\mathfrak{L} \in \mathcal{M}(2^{|Q|})$ следует, что $\mathfrak{L} \in \mathcal{M}$.

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 4.5.
Если активная матрица автомата $\mathfrak{A}$ *имеет единственное максимальное по модулю собственное значение, то задаваемый этим автоматом язык является марковским.*

Доказательство теоремы 4.5 содержит способ нахождения произвольной n-граммы 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 4.4** для установления принадлежности языка $\mathfrak{L}$ к классу марковских языков достаточно проверить только существование n-грамм при $n = |Q|$, 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.

В **пятой главе** изучаются вопросы моделирования одних марковских языков другими. В частности, показано, что для любого марковского языка можно построить автомат, обладающий набором биграмм, сколь угодно близким к набору биграмм исходного автомата.

Показано, что класс марковских языков не замкнут относительно основных теоретико-языковых операций: объединения, пересечения и дополнения, поэтому необходимо рассматривать более узкие классы языков. Примером такого класса является, например, 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 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. *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 5.1

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. Это утверждение сформулировно в работе в виде теоремы 5.2:

Теорема 5.2

Для любого марковского автомата $\mathfrak{A}$ и любого $\varepsilon > 0$ найдётся автомат $\mathfrak{A}'$ такой, что $\rho(\pi_{\mathfrak{A}}, \pi_{\mathfrak{A}'}) < \varepsilon$, и этот автомат может быть получен из «простейших» автоматов – отрезков и циклов путём применения к ним операции склейки автоматов.

При этом под расстоянием между двумя матрицами мы понимаем максимум поэлементного модуля разности, то есть для двух биграммных матриц

$\pi = \{G_{ab}\}$ и $\pi' = \{G'_{ab}\}$

$$\rho(\pi, \pi') = \max_{a,b \in A} |G'_{ab} - G_{ab}|.$$

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 5.1 даёт конструктивный способ построения такого автомата.

Я выражаю глубокую и искреннюю благодарность своему научному руководителю – доктору физико-математических наук, профессору Дмитрию Николаевичу Бабину за постановку задач, постоянную поддержку и внимание к работе.

Я благодарю научного сотрудника лаборатории Проблем теоретической кибернетики, кандидата физико-математических наук Ивана Леонидовича Мазуренко за ценные обсуждения.

Я выражаю глубокую признательность заведующему кафедрой Математической теории интеллектуальных систем, академику, профессору Валерию Борисовичу Кудрявцеву за постоянное внимание к работе и поддержку, а также всем сотрудникам кафедры Математической теории интеллектуальных систем и лаборатории Проблем теоретической кибернетики за творческую атмосферу, способствующую научной работе.

Публикации автора по теме диссертации

1. **Холоденко А.Б.** *Использование лексических и синтаксических анализаторов в задачах распознавания для естественных языков.* // Интеллектуальные системы. Т.4, вып. 1-2, 1999, с.185-193.

2. **Холоденко А.Б.** *О построении статистических языковых моделей для систем распознавания русской речи.* // Интеллектуальные системы. Т.6, вып. 1-4, 2002, с.381-394.

3. **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.

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

5. **Холоденко А.Б.** *Лексический анализатор в распознавании последовательных образов.* // Информационные технологии в инновационных проектах: Материалы докладов. Международная конференция, 20-22 апреля 1999 г. -Ижевск: ИжГТУ, 1999, с. 43-44.

6. **Холоденко А.Б.** *Исправление ошибок в формальных языках.* // Нейрокомпьютеры и их применение: Сборник докладов. IV Всероссийская конференция, Москва. 16-18 февраля 2000 г. -М.: Издательское предприятие редакции журнала "Радиотехника", 2000, с. 627-630.

7. **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.