На правах рукописи

КОРНЫШОВ АЛЕКСАНДР НИКОЛАЕВИЧ

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРОЦЕССА АНАЛИЗА БЛИЗОСТИ ПРЕДИКАТОВ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ

Специальность 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Великий Новгород - 2007

Работа выполнена на кафедре Программного обеспечения вычислительной техники и автоматизированных систем Государственного образовательного учреждения “Новгородский государственный университет им. Ярослава Мудрого”

Научный руководитель: доктор технических наук, профессор Емельянов Геннадий Мартинович

Официальные оппоненты: доктор технических наук, профессор Немирко Анатолий Павлович

кандидат технических наук, доцент Макаров Владимир Алексеевич

Ведущая организация: Государственное учреждение «Научно-исследовательский институт прикладной математики и кибернетики Нижегородского государственного университета им. Н. И. Лобачевского Министерства образования Российской Федерации».

Защита состоится “07” сентября 2007 г. в 16 часов на заседании диссертационного совета Д 212.168.04 в Новгородском государственном университете им. Ярослава Мудрого (173003, Россия, г. Великий Новгород, ул. Б. Санкт-Петербургская, 41)

С диссертацией можно ознакомиться в библиотеке университета.

Автореферат разослан “__” июля 2007 г.

Ученый секретарь диссертационного совета Д 212.168.04
Доктор физико-математических наук, профессор Эминов С.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Обработка Естественного Языка (ЕЯ) является одним из приоритетных направлений в области развития Искусственного Интеллекта. Настоящая диссертационная работа посвящена решению проблемы использования существующих в ЕЯ закономерностей для автоматизации накопления и систематизации Семантических Знаний в Интеллектуальных Системах (ИС).

Актуальность работы. Следует выделить две группы затруднений технического характера, которые мешают применению в информационных системах интеллектуально-коммуникативных технологий, предполагающих взаимодействие человека и ЭВМ на ЕЯ.

Первая группа затруднений состоит в необходимости предварительного ввода в ЭВМ всей полноты Знаний о ЕЯ. Объем Семантических Знаний велик, и поэтому их ввод требует огромных затрат труда множества квалифицированных специалистов, в том числе для предварительной формализации и систематизации этих Знаний. При традиционном подходе не учитывается возможность неполноты и противоречивости самих Знаний о ЕЯ, относительность представления как о ЕЯ, так и ситуациях его использования, а также потребность в постоянном изменении введённых Семантических Знаний. Если для некоторых областей применения на коротком временном промежутке функционирования ИС изменениями в ЕЯ можно пренебречь, то в целом изменчивость – глубинное и универсальное свойство как естественных, так и искусственных языков, и его необходимо учитывать.

Причина чрезмерной трудоёмкости – отсутствие на настоящий момент методов предметно-адаптивной формализации и автоматической систематизации Знаний о ЕЯ. Как следствие этого, при использовании существующих подходов к формализации Знаний о ЕЯ для каждого языка и тематического подмножества требуется производить заново как формализацию, так и ввод Семантических Знаний, что увеличивает затраты труда на разработку ИС. Разработка математической модели процесса ввода и систематизации информации о ЕЯ с помощью автоматического выявления и применения машиной закономерностей каждого ЕЯ позволила бы решить задачу автоматического накопления и систематизации ИС Знаний об используемом ЕЯ. Здесь можно выделить задачу интеллектуализации процесса пополнения Семантических Знаний, которая заключается в автоматическом построении машиной части модели ЕЯ, и задачу интеллектуализации самого процесса ввода – человеко-машинного общения, который, чтобы избавить от рутинной деятельности человека-оператора, необходимо осуществлять на ЕЯ. В данном случае предмет общения как раз ЕЯ, и логично и удобно было бы организовать работу оператора по вводу с использованием самого ЕЯ. Оператору, которому в этом случае уже не нужно быть экспертом-лингвистом, а только носителем языка, достаточно ввести в ЭВМ обучающее множество прецедентов Смысловой Эквивалентности (СЭ) высказываний на ЕЯ.

Вторая группа затруднений состоит в том, что в реализованных на практике методах Обработки ЕЯ присутствует противоречие между скоростью

обработки, которая достигается при применении простых правил преобразований, и универсальностью представления ЕЯ-преобразований. Универсальность позволяет более полно описывать ЕЯ с помощью сочетания сложных правил. Как следствие уменьшается трудоёмкость описания ЕЯ, но увеличивается вычислительная сложность алгоритмов анализа, которые воспроизводят полноту и непротиворечивость языкового описания путём согласования множества правил. При автоматическом накоплении и систематизации ИС Знаний о ЕЯ неполнота и противоречивость Знаний, вводимых оператором в простой, но универсальной форме, устраняются на этапе обучения ИС, и их не требуется восполнять вычислениями на этапе анализа и преобразований ЕЯ-высказываний, причём каждый раз делать это заново.

Объект исследований – Предикаты Семантических Отношений (СО) Конструкций ЕЯ, с помощью которых в модели ЕЯ представляются Семантические Знания. Систематизация Семантических Знаний на основе присутствующих в них закономерностей является процессом обобщения Предикатов СО.

Предметом исследований является процесс анализа близости Предикатов СО. Определение меры близости Семантических Знаний с помощью сравнения Предикатов СО использует закономерности, которые выявлены в ЕЯ в процессе обобщения Предикатов. Возможности как процесса анализа близости, так и процесса обобщения Предикатов ограничены уровнем автоматического выявления закономерностей в ЕЯ алгоритмом обобщения. Поэтому за исключением предельного случая – автоматического построения машиной полной модели ЕЯ, для процесса обобщения Предикатов возможна лишь частичная автоматизация, тогда как процесс анализа близости Предикатов СО, используя готовые результаты процесса обобщения Предикатов, допускает полностью автоматическое выполнение.

Актуальность исследований обоснована отсутствием в настоящий момент научно обоснованных методов обобщения Предикатов применительно к описанию СО в ЕЯ, а также методов анализа близости Предикатов СО в той мере, в какой возможно автоматическое обобщение Предикатов.

Цель и задачи работы. Целью настоящей диссертационной работы является разработка методов количественной оценки близости Предикатов в процессе их обобщения в ходе машинного обучения распознаванию СО Конструкций ЕЯ. Для достижения поставленной цели в работе решаются следующие задачи:

1. Разработка концептуальной модели процесса обобщения Предикатов в ходе машинного обучения распознаванию СО Конструкций ЕЯ;
2. Построение и исследование свойств формального аппарата математического моделирования процесса обобщения Предикатов СО с использованием количественной оценки меры близости Предикатов в ходе машинного обучения распознаванию СО Конструкций ЕЯ;
3. Исследование свойств сложных систем Предикатов СО и моделирование с их помощью различных видов синонимических

преобразований, известных из лингвистики, в том числе проблемных, а также решение практической задачи морфологической, сортовой, родовидовой классификации лексики;

4. Разработка алгоритмов нахождения количественной оценки близости Предикатов в общем виде с учетом возможных методов оценивания и оптимизация этих алгоритмов для сравнения сложных, иерархизированных систем Предикатов СО. Использование количественной оценки близости для трансформации систем Предикатов СО в процессе обобщения Предикатов.

Методы исследований. При проведении исследований в работе использовались методы математической логики и теории множеств; основные положения теоретической и когнитивной лингвистики, системной типологии языков и когнитологии, теории формальных языков, а также прикладных методов анализа данных и знаний.

Научная новизна. В ходе решения поставленных задач получены следующие результаты, являющиеся новыми в данной области исследований:

1. Предложен комплексный подход к решению ряда задач компьютерной Обработки ЕЯ. Показано, что задача распознавания СЭ ЕЯ-высказываний сводится к задаче сравнения систем Предикатов СО и нахождения количественной оценки их близости. Последняя задача решается теми же методами, что и задача обобщения систем Предикатов для предварительного машинного обучения ИС распознаванию СО Конструкций ЕЯ.

2. Разработана модель ЕЯ, которая позволяет универсальным образом представить СО в ЕЯ с помощью Наборов Правил Преобразований (НПП) и модели ситуаций ЕЯ-употребления. На основе модели можно проводить машинное обучение распознаванию произвольных СО Конструкций ЕЯ.

3. Предложен алгоритм распознавания СО ЕЯ-высказываний с помощью процесса обобщения систем Предикатов СО, который объединяет в один процессы анализа Конструкций ЕЯ и сравнения Смыслов ЕЯ-высказываний.

4. Доказаны теоремы о том, что вычислительная сложность процесса обобщения Предикатов СО линейно зависит от количества Предикатов в системе НПП, которая является результатом обобщения исходной, а вычислительная сложность процесса распознавания СЭ ЕЯ-высказываний не экспоненциально, как в существующих алгоритмах синонимического перифразирования, а линейно зависит от количества уровней синонимии в иерархизированной системе.

Практическая значимость и внедрение. Областью непосредственного практического применения теоретических результатов настоящей работы является автоматизация обучения, автоматический контроль знаний с помощью тестирования на ЕЯ путем машинного анализа СЭ между ответами учащихся и эталонами, заданными педагогом, поскольку при данном применении знания учащихся, фиксируемые в текстах ЕЯ, постоянно измеряются/оцениваются экспертом-педагогом в приложении к стабильному ситуационному контексту вопросов и предмета.

Разработанные в диссертации методы решения задач Обработки ЕЯ доведены до реализации. Разработанные в диссертации методы и алгоритмы

количественной оценки близости Предикатов СО, таксономии Конструкций ЕЯ, логического вывода усложнением вариантов нашли практическое воплощение в программном комплексе, который в дальнейшем планируется использовать для решения задач автоматизации составления тезаурусов по дисциплинам специальности "Программное обеспечение вычислительной техники и автоматизированных систем" в учебном процессе Новгородского государственного университета (имеются акты о внедрении).

Результаты проведенных исследований использовались в работе по гранту РФФИ № 06-01-00028.

Достоверность и эффективность. Достоверность полученных теоретических результатов подтверждается корректностью доказательств теорем об алгоритмической разрешимости и вычислительной сложности процесса распознавания СО Конструкциях ЕЯ и процесса обобщения систем Предикатов СО Конструкций ЕЯ. Также достоверность подтверждается соответствием модели Семантики Конструкций ЕЯ формальным критериям, сформулированным при постановке задачи.

Эффективность предложенных алгоритмов распознавания СЭ ЕЯ-высказываний в сравнении с известными на сегодняшний день алгоритмами синонимического перифразирования подтверждается теоремами; иллюстрацией работы алгоритмов являются примеры, приведенные в приложении. Также показано, что система НПП позволяет промоделировать проблемные и сложные синонимические преобразования, известные из лингвистики.

Апробация работы и публикации. Основные положения и полученные результаты диссертационной работы апробированы в докладах на конференциях: XVIII международной научно-методической конференции “Математика в вузе” (Великий Новгород, 2005), 6-й международной научной конференции «Интеллектуализация обработки информации» (ИОИ-2006) (Крым, Алушта, 2006), в докладах на научных конференциях и семинарах в рамках Дней Науки в Новгородском государственном университете имени Ярослава Мудрого (Великий Новгород, 2005-2007) и опубликованы в 6 работах, список которых приводится в конце автореферата.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы, включающего 91 наименование, а также двух приложений. Основная часть работы изложена на 126 страницах, содержит 8 рисунков и 1 таблицу.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность выбранной темы; формулируется цель и содержание поставленных задач, а также объект и предмет исследования; сообщается теоретическая значимость и прикладная ценность полученных результатов.

В **первой главе** диссертационной работы формулируются требования к моделированию процесса обобщения Предикатов СО с использованием количественной оценки близости Предикатов в ходе машинного обучения

распознаванию СО Конструкций ЕЯ, а также выбираются теоретический подход к моделированию ЕЯ, модель представления Знаний о СО Конструкций ЕЯ и методы, подходящие для оценивания близости Семантических Знаний в этой модели.

В **параграфе 1.1** на основе объединения формального и функционального подходов дано определение ЕЯ как сложной Знаковой Системы (ЗС), основной функцией которой является использование в качестве средства общения, абстрактная модель которой задаётся формальным механизмом порождения всех возможных высказываний в этой ЗС, а также формальным механизмом установления соответствия между Смыслами высказываний. Перечислены конкретизирующие условия постановки и решения задачи установления СЭ высказываний на ЕЯ: *внешние условия*, называемые Семантическими Знаниями (какая информация из текста высказывания или из тезауруса учитывается, какие подсмыслы включаются, какие Знания используются, на каком уровне представления ЕЯ-высказываний производится сравнение), *операция* сравнения (любую операцию сравнения Смыслов можно свести к СЭ путем добавления функции к Смыслам), *тип значения результата* сравнения (логический, непрерывный числовой или какой-то еще, например, ранг оценки близости).

Под Предикатом понимается функция, которая устанавливает СЭ для определенных Конструкций ЕЯ с учетом внешних условий, которые задаются Семантическими Знаниями, типа операции, который задается Смысловой функцией, а также типа значения результата операции сравнения. Семантические Знания могут быть заданы как набор Предикатов. Семантические Знания учитывают особый вид отношения СЭ – совокупность СЭ, которая обозначается обобщающим термином СО. Приблизительность результата и зависимость его от тезауруса заставляют переформулировать задачу установления СЭ высказываний на ЕЯ как более общую задачу распознавания СО высказываний на ЕЯ.

Распознавание требует наличия предварительно сформированных Семантических Знаний. Задача машинного обучения автоматическому распознаванию СО Конструкций ЕЯ с помощью количественной оценки близости Предикатов СО, задающих в совокупности Знания о СО Конструкций ЕЯ, состоит в выработке по прецедентам СЭ Конструкций ЕЯ таких Знаний о СО, которые позволяют распознавать как можно большее число СО Конструкций ЕЯ как можно точнее.

В **параграфе 1.2** описаны наиболее общие функциональные и структурные свойства ЕЯ, которым должна удовлетворять модель Семантики Конструкций ЕЯ. Сформулированы требования, которые можно использовать в качестве критериев адекватности при оценке разработанных в работе моделей.

Рассмотрены общие функциональные свойства ЕЯ: *гибкость* ЕЯ – возможность выражения одной и той же информации множеством различных способов в зависимости от наличествующих возможностей, потребностей и обстоятельств речевого акта; *нестрогость* ЕЯ – возможность включения и исключения подсмыслов, что меняет СЭ для одних и тех же Конструкций ЕЯ;

устойчивость ЕЯ – возможность изменения Семантических Знаний во времени с постепенной коррекцией ошибок и снятием возникающих противоречий.

Наиболее общие структурные свойства Конструкций ЕЯ для всех известных ЕЯ: *парадигматика* – способность некоторого признака принимать одно из множества альтернативных значений по некоторой шкале; *синтагматика* – возможность позиционного или ролевого комбинирования нескольких альтернативных признаков каким-либо зависимым образом; *предикация* – зависимость как интерпретации, так и генерации отдельных высказываний от языкового и ситуационного контекста; *разделимость* – возможность независимого оперирования частями высказывания на ЕЯ, то есть установления СЭ ЕЯ-высказываний по частям.

В **параграфе 1.3** проанализированы существующие модели Семантики Конструкций ЕЯ и методы моделирования Знаний в применении к Семантике, выявлены их достоинства и недостатки с целью поиска сочетания модели и методов, которые наиболее подходят для машинного обучения распознаванию СО Конструкций ЕЯ. Такая модель Семантики Конструкций ЕЯ должна:

- Учитывать многозначность как отдельных слов, так и групп слов;
- Устанавливать Смысл и для фраз с пропущенными словами;
- Учитывать синтаксические связи в предложениях и структурные связи в тексте;
- Быть универсальна для представления любых ЕЯ-преобразований;
- Позволять обучение с учетом регулярности синонимии в ЕЯ;
- Допускать сравнение Смыслов фраз без порождения всех возможных инвариантных по Смыслу фраз.

Для моделирования Знаний о многозначности слов и Смысла фраз с пропущенными словами подходит применение стандартных, статистических методов и нейросетей; эти методы предполагают возможность обобщения вероятностных данных. Для моделирования сложных структур, синонимии по частям и регулярности подходят используемые в общем случае методы анализа закономерностей в данных и знаниях, для которых необходимо адаптировать модель Семантических Знаний. Методы анализа закономерностей в данных и знаниях выбраны в качестве основных в дальнейших исследованиях.

В **параграфе 1.4** в качестве модели Семантики Конструкций ЕЯ, которая наиболее приемлема для целей моделирования, выбрана предложенная автором настоящей диссертации модель ситуаций употребления ЕЯ, дополненная механизмом построения Концептуальных Знаний о ЕЯ путем анализа ЕЯ-закономерностей. Такая дополненная модель, названная концептуально-ситуационной, удовлетворяет всем вышеперечисленным требованиям.

Ситуации языкового употребления, представляются с помощью тройки :

$$S = (O, P, V), \text{ где} \qquad (1)$$

S – ситуация, фиксирующая однозначный ЕЯ-контекст; O – множество подразумеваемых в этой ситуации объектов, которые соответствуют в конкретной ситуации ЕЯ-употребления денотатам, а в прототипе ситуации становятся концептами; P – множество реляционных отношений между объектами O. Заданные отношения и используемые объекты соответствуют

всевозможным событиям, связям и логическим условиям существования данной ситуации. Отношения P носят вероятностный характер, а выражаемые через них логические условия могут быть противоречивы. V – множество альтернативных форм поверхностного выражения ситуации S, то есть фиксации Смысла данной ситуации S с помощью Конструкций ЕЯ. Выбор тех или иных форм из V для поверхностного выражения Смысла ситуации S также носит вероятностный характер, причем коррелируется их использование в разных ситуациях, что учитывается с помощью G – отдельного механизма согласования Семантических Знаний о ситуациях ЕЯ-употребления.

Модель (1) достаточно универсальна для представления любых Семантических Знаний о ЕЯ, поскольку отношения P могут быть любого вида. Это позволяет моделировать с ее помощью любые ЕЯ-преобразования. Также не определен конкретный вид Конструкций ЕЯ для альтернативных форм поверхностного V выражения ситуации ЕЯ-употребления S, он может быть заимствован из различных моделей Семантики Конструкций ЕЯ, например, это могут быть шаблоны или помеченные деревья.

В **параграфе 1.5** разработана концептуальная модель процесса обобщения Предикатов в ходе машинного обучения распознаванию СО Конструкций ЕЯ, включающая в качестве объектов моделирования:

- Представление ЕЯ как системы ситуаций ЕЯ-употребления;
- Модель представления Знаний о Семантике Конструкций ЕЯ с помощью множества прецедентов и соответствующих им Семантических Знаний в форме Предикатов СО;
- Процесс машинного обучения распознаванию СО Конструкций ЕЯ;
- Методы количественной оценки близости Семантических Знаний в форме Предикатов СО;
- Механизм установления ЕЯ-закономерностей путем обобщения Предикатов СО в ходе Концептуального Анализа Семантических Знаний.

Моделирование установления соответствия между Смыслами высказываний на ЕЯ основывается не на формальном механизме порождении всех возможных высказываний в ЗС ЕЯ, а на основной функции ЕЯ, которая заключается в использовании ЕЯ в качестве средства общения. Именно потому, что мощность множества Смыслов превосходит мощность множества правил синонимических преобразований, можно выявлять лишь закономерности в использовании правил для выражения Смыслов. ЕЯ L представляет собой множество $\{S\}_L$ ситуаций ЕЯ-употребления. Механизм установления соответствия между Смыслами высказываний на L можно представить с помощью Предикатов СО E_S, которые определяют в заданной ситуации $S \in \{S\}_L$ степень СЭ любой пары высказываний из D_L как отображение $E_S : D_L \to E$. Здесь V_L – множество возможных высказываний на некотором ЕЯ L; $D_L = V_L \times V_L$; E – множество вещественных чисел от 0 до 1. В случае строгости ЕЯ L: $E = \{0,1\}$ и СЭ заданы однозначно: $E_S = D_S$, где $D_S \subseteq D_L$.

Модель Семантики Конструкций ЕЯ L как системы ситуаций ЕЯ-употребления:

$$\{S\}_L = (\{E_S\}_L, G_L, N_L), \text{ где} \qquad (2)$$

$E_L = \{E_S\}_L$ – система Семантических Знаний в форме Предикатов СО; через $N_L : V_L \to E$ заданы нормы ЕЯ; механизм $G_L : \{(O, P)\}_L \leftrightarrow V_L$ используется для согласования Семантических Знаний о ситуациях ЕЯ-употребления (здесь $\{(O, P)\}_L$ – множество возможных концепций ЕЯ L).

Механизм формирования Семантических Знаний по прецедентам ***K***, называемый Концептуальным Анализом, сопоставляет любому произвольному множеству текущих Семантических Знаний $E' : D' \to E$ ($D' \subset D_L$) множество экстраполированных на все высказывания на ЕЯ L Концептуальных Знаний $K' : D_L \to E$. Обозначив множество всех возможных отображений типа E' через $E^* = \{E'\}$, а типа K' через $K^* = \{K'\}$, получим:

$$\boldsymbol{K} : E^* \to K^* \tag{3}$$

Концептуальные Знания $K_1 = \boldsymbol{K}(E_1)$ составлены на основе множества уже известных ИС фактов СЭ $E_1 : D_1 \to E$ ($D_1 \subset D_L$, $E_1 \subset E_L$), Точность ε Знаний K_1 относительно расширяющего E_1 множества E_2 еще неизвестных ($E_1 \subset E_2 \subset E_L$ и при этом $E_2 : D_2 \to E$, $D_1 \subset D_2 \subset D_L$) можно определить как:

$$\varepsilon = 1 - \sum_{i=1}^{|D_2|} \frac{|E_2(d_i) - K_1(d_i)|}{|D_2|} \tag{4}$$

где $d_i \in D_2$, а $K_1(d_i)$ – прогноз $E_2(d_i)$, полученный путем экстраполирования закономерностей, выявленных Концептуальным Анализом на множестве E_1. В качестве критерия адекватности модели ***K*** (при соблюдении условий случайности и статистической значимости выборок D_1 и D_2) можно потребовать достаточно высокую точность прогноза при значительном, экспоненциальном соотношении размеров множеств D_1 и D_2:

$$\frac{\ln |D_2|}{\ln |D_1|} >> 2 \tag{5}$$

Во **второй главе** формализован механизм Концептуального Анализа и построен формальный аппарат математического моделирования процесса обобщения Предикатов СО с использованием количественной оценки меры близости Предикатов.

В **параграфе 2.1** исследованы принципы таксономии прецедентов СО Конструкций ЕЯ механизмом Концептуального Анализа и установлено, что такая таксономия представима с помощью Предикатов в форме НПП над множествами структурированных значений аргументов СО.

В **параграфе 2.2** с учетом требований к модели Семантических Знаний, сформулированных в первой главе, формализовано представление Предикатов СО Конструкций ЕЯ в форме НПП, которая наиболее подходит для организации процесса обобщения Предикатов. Каждый НПП p :

$$F_p : V_1 \Leftrightarrow V_2 \Leftrightarrow \ldots \Leftrightarrow V_n \text{ , где} \tag{6}$$

F_p – Смысл преобразований над множеством структурированных значений аргументов Предиката СО (V_1, V_2, …, V_n), который задается с помощью данного НПП; V_i – одна из форм поверхностного выражения Смысла F_p :

$$V_i = (X_{i1}, X_{i2}, \ldots X_{im}), \text{ где} \quad (7)$$

X_{ij} – либо переменная, у которой значение может изменяться (в этом случае $X_{ij} \in X_p$ – множеству переменных данного НПП p), либо константа для данного F_p ($X_{ij} = c_r$, где символ $c_r \in C$, то есть алфавиту ЗС ЕЯ). На изменяемых переменных может быть установлено отношение E_X Смыслового равенства вида: $X_{ij} \Leftrightarrow X_{kl}$, которое можно представить как отображение:

$$E_X : X_p \times X_p \to \{0,1\} \quad (8)$$

причем $E_X = 1$ для тех пар переменных, для которых задано равенство, и 0 в противном случае. В общем случае V_i и F_p – предикаты на множестве всех возможных векторов-констант $\mathbf{C}^* = \{(c_1, c_2, \ldots c_r, \ldots c_q)\}$, где $c_r \in C$ – любой допустимый в ЗС C данного ЕЯ символ (набор знаков или их заменитель):

$$V_i : \mathbf{C}^* \to \{0,1\}, F_p : \mathbf{C}^* \to \{0,1\} \quad (9)$$

Более сложные НПП могут быть частично или полностью выражаться через более простые. Под сложностью Предиката СО F_p понимаются размеры $|\mathbf{C_p}|$ подмножества $\mathbf{C_p} \subset \mathbf{C}^*$, на котором данный Предикат всегда истинен: $F_p(\mathbf{C_p}) \equiv 1$. В случае статистической истинности (учитывая ошибки и то, что закономерности могут носить вероятностный характер) более простого Предиката СО F_{p1} на подмножестве переменных более сложного Предиката F_{p2} ($|\mathbf{C_{p1}}| < |\mathbf{C_{p2}}|$, $\mathbf{C_{p2}} \subset \mathbf{C_{p1}} \cdot \mathbf{C}^*$) можно говорить о статистической выражаемости F_{p2} через F_{p1}: $F_{p2} \Rightarrow F_{p1}$ (считаем, что соотносимые в Предикатах переменные располагаются на одних и тех же местах в векторах-константах). Таким образом система Предикатов НПП может быть иерархизирована по сложности: более сложные Предикаты СО будут представляться с помощью более простых НПП, которые будут располагаться на более низких уровнях иерархии. Сочетание произвольной синонимии с регулярными преобразованиями позволяет задавать сложные комбинации СЭ с учетом возможных исключений.

В процессе обобщения Предикатов СО используется два вида гипотез: 1) о выражении Предиката через комбинацию простых ($F_{сложн} \Rightarrow F_{прост}$ при том, что $|\mathbf{C_{прост}}| < |\mathbf{C_{сложн}}|$) и 2) о существовании Предиката, обобщающего группу близких НПП. Выражение через комбинацию $F_{сложн} \Rightarrow F_1 \times F_2 \times \ldots \times F_{макс}$ позволяет экстраполировать, то есть дополнить левую часть до правой с учетом исключений и статистической значимости такой замены на комбинацию НПП: $(F_1 \times F_2 \times \ldots \times F_{прост}) \Rightarrow F_{сложн}$. Тем самым приходим ко второму случаю, когда для обобщающего Предиката $F_{обоб}$ верно, что каждый Предикат из группы $\{F_{близк}\}$ близких НПП: $F_{близк} \Rightarrow F_{обоб}$ (при том, что $|\mathbf{C_{близк}}| < |\mathbf{C_{обоб}}|$).

Пусть $\mathbf{F_{исх}} = \{F_{исх}\}$ – произвольное подмножество, $\mathbf{F_{исх}} \subset \mathbf{F}^*$ – множества всех возможных систем Предикатов СО. Процесс обобщения Предикатов:

$$\Pi_{ОБОБ}: \mathbf{F}^* \to \mathbf{F}^* \quad (10)$$

есть поиск на пространстве $\mathbf{F}^*$ для исходной системы НПП $\mathbf{F_{исх}}$ такого его компактного представления $\mathbf{F_{min}} \subset \mathbf{F}^*$, для которого для $\forall F_{исх} \in \mathbf{F_{исх}}$ всегда $\exists \mathbf{F} \subset \mathbf{F_{min}}$, такое что: $F_{исх} \Rightarrow F$ (произведение $F = F_1 \times F_2 \times \ldots \times F_i \times \ldots F_q$, где любой $F_i \in \mathbf{F}$), при том, что $\mathbf{C}_{исх} \subseteq \mathbf{C}_F$. Размер искомого множества минимален

$|F_1| + \ldots + |F_i| + \ldots + |F_q| = \mathbf{min}$, ($F_i \in \mathbf{F_{min}}$). При этом пространство поиска $\mathbf{F}^*$ ограничено пределами моделируемой части ЕЯ так, что всегда $\mathbf{F_{min}} \Rightarrow D_S$ (в противном случае происходило бы уже обобщение самого D_S до одного единственного Предиката, тождественного 1).

Теорема 2.1. Процесс обобщения Предикатов СО $\Pi_{\text{ОБОБ}} : \mathbf{F}^* \rightarrow \mathbf{F}^*$ (10) алгоритмически разрешим. Доказательство: процесс конечен, так как конечно множество всех возможных систем Предикатов СО $\mathbf{F_{исх}}^* \subset \mathbf{F}^*$, которые можно построить из элементов $\mathbf{F_{исх}}$ так, чтобы сумма размеров Предикатов построенной системы НПП не превышала суммы размеров Предикатов $\mathbf{F_{исх}}$. За конечное число шагов поиска на пространстве $\mathbf{F_{исх}}^*$ может быть найдено компактное представление $\mathbf{F_{min}}$ с минимальным размером среди множеств с допустимой точностью $\varepsilon \le \varepsilon_{max}$ представления Семантических Знаний D_S.

В **параграфе 2.3** выявлено, что процесс распознавания СЭ линейных структур ЕЯ-высказываний является частным случаем процесса сравнения систем Предикатов СО, результатом которого является вычисление количественной оценки меры близости систем Предикатов. Процесс сравнения систем Предикатов СО может быть осуществлен теми же методами, что и процесс обобщения Предикатов в ходе обучения распознаванию СО.

С использованием построенного аппарата описан процесс распознавания СО Конструкций ЕЯ, и установлено, что в ходе данного процесса порождение Конструкций ЕЯ, инвариантных по Смыслу, не приводит к экспоненциальному росту времени перебора с увеличением числа уровней иерархии и количества Предикатов за счет снятия неоднозначности, возникающей на начальных уровнях, при преобразовании на последующих, а также объединения в один процессов анализа Конструкций ЕЯ и сравнения Смыслов ЕЯ-высказываний.

Теорема 2.2. Процесс распознавания СЭ ЕЯ-высказываний $v : A \Leftrightarrow B$ в заданной системе НПП **F** алгоритмически разрешим. Это следует из конечности множества истинности $\mathbf{C}_v(\mathbf{F})$, которое построено для Предикатов и их комбинаций в системе **F** при условии, что размер векторов-констант из $\mathbf{C}_v(\mathbf{F})$ не превышает размера *v*. Если *v* принадлежит $\mathbf{C}_v(\mathbf{F})$, установлена точная СЭ *A* и *B*, в противном случае ищется комбинация векторов-констант из $\mathbf{C}_v(\mathbf{F})$, наиболее приближающая *v*. Если оценка достоверности такого приблизительного представления превышает допустимый предел, то установлена приблизительная СЭ *A* и *B*.

В **параграфе 2.4** разработаны и исследованы различные способы и методы вычисления количественной оценки близости Предикатов СО. Сравнение методов позволило установить, что методы дают тем более схожие результаты, чем больше насыщенность Концептуальных Знаний, поэтому необходимо нормировать оценку близости Предикатов, полученную с помощью выбранных методов, на точность Концептуальных Знаний, вычисленную теми же методами.

Формула точности $\varepsilon = 1 - R(K_2, K_1)$ Концептуальных Знаний:

$$1 - \varepsilon = R(K_2, K_1) = \sum_{i=1}^{|K_2|} \frac{|K_2(d_i) - K_1(d_i)|}{|K_2|} \qquad (11)$$

K_1 относительно базового множества Знаний K_2. Под $R(K_2, K_1)$ понимается расстояние между Знаниями $K_2 \cap K_1$ и K_2, нормированное на размер базового множества Знаний $|K_2|$. В качестве базовых Знаний может быть выбран один из сравниваемых Предикатов $П_1$ и $П_2$, тогда размер общего множества истинности $|\mathbf{C}(П_1 \cap П_2)| = |\mathbf{C_1} \cap \mathbf{C_2}| = |\mathbf{C}(П_1 \cap П_2)| = R(П_1, П_2) \cdot |\mathbf{C_1}| = R(П_2, П_1) \cdot |\mathbf{C_2}|$. Оценить *информативность* Концепции $П_1$ при добавлении ее на очередном шаге обучения к текущим Концептуальным Знаниям $П_2$ можно как $R(П_2, П_1)$, так как часто предполагается, что $|\mathbf{C_2}| >> |\mathbf{C_1}|$, то есть Концептуальные Знания значительно информативнее Концепции. А вот $R(П_1, П_2)$ будет количественной оценкой *известности* для ИС новой Концепции $П_1$. Если $R(П_1, П_2) = 1$, то Концепция $П_1$ может быть полностью распознана на основе текущих Концептуальных Знаний $П_2$. Если оценка известности только близка к единице, то предполагается приблизительная Смысловая выразимость преобразований $П_1$ через $П_2$: $П_1 \Rightarrow П_2$.

В качестве базовых Знаний может использоваться обобщающий Предикат $П_{1\cup 2} = П_1 \cup П_2$. В этом случае $|\mathbf{C}(П_1 \cap П_2)| = Б(П_2, П_1) \cdot |\mathbf{C}(П_1 \cup П_2)|$. Нормированная количественная оценка меры близости Предикатов $П_1$ и $П_2$: $Б(П_2, П_1) = R(П_{1\cup 2}, П_1) + R(П_{1\cup 2}, П_2) - 1$ выражает количественно отношение множества истинности пересечения Смыслов (информации) Предикатов $П_1$ и $П_2$ к множеству истинности их общей Смысловой информативности. Получить Предикат $П_{1\cup 2}$ можно просто логическим сложением множеств истинности объединяемых Предикатов: $|\mathbf{C}(П_1 \cup П_2)| = |\mathbf{C_1} \cup \mathbf{C_2}|$. При втором подходе $П_{1\cup 2}$ рассматривается как результат процесса обобщения Предикатов $П_1$ и $П_2$.

Мера близости $\boldsymbol{\varepsilon}$ Концептуальных Знаний K_1 и K_2 относительно моделируемых Семантических Знаний E_L может быть выражена через их точности $\varepsilon_1 = 1 - R(E_L, K_1)$, $\varepsilon_2 = 1 - R(E_L, K_2)$ и взаимную меру близости $Б(K_2, K_1)$:

$$\boldsymbol{\varepsilon} = \frac{\varepsilon_1 + \varepsilon_2}{1 + Б(K_2, K_1)} \qquad (12)$$

В **третьей главе** диссертационной работы показано, что построенный во второй главе аппарат Предикатов СО в форме НПП в полной мере удовлетворяет требованиям к моделированию Семантики Конструкций ЕЯ, сформулированным в первой главе.

В **параграфе 3.1** исследуются свойства сложных систем Предикатов СО. Выяснено, что иерархизация упрощает построение сложных систем Предикатов СО: процесс распознавания СО и процесс обобщения Предикатов в ходе обучения ИС распознаванию СО Конструкций ЕЯ. Упорядоченность последовательности применения НПП от простого к сложному уровням синонимии в процедуре порождения вариантов перифразирования приводит к тому, что вычислительная сложность данной процедуры линейно зависит только от количества уровней синонимии и даже в маловероятном для реального ЕЯ случае неоднозначности распознавания ЕЯ-высказывания на всех уровнях синонимии оказывается ограничена сверху количеством НПП в системе Предикатов СО.

Теорема 3.1. Вычислительная сложность процесса обобщения Предикатов СО исходной системы НПП $\mathbf{F_{исх}}$ линейно зависит от количества k Предикатов в конечной системе НПП $\mathbf{F_{кон}} = П_{ОБОБ}(\mathbf{F_{исх}})$, которая является результатом обобщения исходной. В доказательстве рассматривается $\mathbf{C}(\mathbf{F_{исх}})$ с заданным отношением следования векторов-констант. Для установления статистической истинности нужна группа прецедентов размером $r_{гр}$. Количество подвыражений не может быть больше числа переменных в F_i, то есть ограничено $r_{пред}$. Верхний предел вычислительной сложности: $k \cdot r_{гр} \cdot r_{пред}$, где $r_{гр}$ и $r_{пред}$ – максимальные.

Теорема 3.2. Вычислительная сложность процесса распознавания СЭ ЕЯ-высказываний $v : A \Leftrightarrow B$ в заданной системе НПП $\mathbf{F}$ линейно зависит от количества уровней синонимии $k_{ур}$ в этой системе и в худшем случае от количества НПП k в этой системе. В доказательстве сравниваются части v с выражениями Предикатов некоторого уровня, начиная с первичного, доказывается их выводимость. Верхний предел вычислительной сложности $k_{ур} \cdot r_{неод} \cdot r_{пред}$, причем $k_{ур} \cdot r_{неод} \le k$, где $r_{неод}$ – максимальное количество неоднозначностей или вариантов возникающих на одном уровне.

В **параграфе 3.2** показано, что с помощью аппарата Предикатов СО в форме НПП возможно промоделировать различные виды синонимических преобразований, известные из лингвистики. Преимущества в том, что использование системы НПП позволяет классифицировать виды замен синонимов и представить их в регулярной форме, а также учитывать изменения Смысла зависимых лексем – фразем.

В **параграфе 3.3** показано, как предлагаемый аппарат моделирования разрешает затруднения моделей Семантики Конструкций ЕЯ, которые основаны на перифразировании (проблему распределения адъюнктов расщепляемого элемента и затрудненность конверсивных преобразований при наличии во фразе синкатегорематических обстоятельств, вложенных или пресуппозиционных Конструкций), а также позволяет моделировать преобразования, которые учитывают различные уровни синонимий (благодаря замене строгого порядка в иерархии пребразований на упорядоченность, когда более простые преобразования могут быть подвыражениями преобразований сколь угодно сложных уровней), и сложные виды синонимических преобразований: выходящие за пределы простого предложения и за рамки регулярной Лексико-Функциональной синонимии.

В **параграфе 3.4** на основе анализа пересечения выделенных классов стандартными методами Концептуального Анализа, которыми реализуется процесс обобщения системы Предикатов СО, решаются задачи морфологической, сортовой, родовидовой классификации лексики, задача установления базовых Семантических Значений Моделей Управления слов ЕЯ.

В **четвертой главе** описаны алгоритмы нахождения количественной оценки близости для систем Предикатов СО в общем виде с учетом возможных методов оценивания: алгоритм распознавания СЭ ЕЯ-высказываний в системе НПП и алгоритм сравнения систем НПП, которые задают Предикаты СО. Алгоритмы оптимизированы для сравнения сложных, иерархизированных систем НПП с помощью механизма логического вывода путем постепенного

усложнения вариантов, более эффективного, чем методы стандартного логического вывода перебором пространства состояний в глубину и в ширину.

В **параграфе 4.1** приводятся алгоритмы с пояснениями к ним. В алгоритме нахождения количественной оценки меры близости систем Предикатов СО используется процедура $\cap(F_1, F_2)$ сравнения двух НПП F_1 и F_2 из разных систем НПП, которая в результате формирует НПП пересечения множеств истинности этих Предикатов $F_1 \Leftrightarrow F_2$ и два остаточных НПП: $F_{1ост} = F_1 - F_1 \Leftrightarrow F_2$ и $F_{2ост} = F_2 - F_1 \Leftrightarrow F_2$. Связи этих НПП в иерархиях систем сохраняются с учетом того, что старые НПП (F_1 и F_2) являются суммами пересечения ($F_1 \Leftrightarrow F_2$) и одного из остатков ($F_{1ост}$ или $F_{2ост}$). $\cap(F_1, F_2) = (F_1 \Leftrightarrow F_2, F_{1ост}, F_{2ост})$. Функция $g(F_i)$ дает оценку размера множества истинности Предиката F_i за вычетом размеров множеств истинности его подвыражений. Последнее необходимо во избежание повторов при подсчете.

АЛГОРИТМ СРАВНЕНИЯ СИСТЕМ ПРЕДИКАТОВ СО В ФОРМЕ НПП

Вход: $\mathbf{F_1}$, $\mathbf{F_2}$ – сравниваемые системы НПП, задающих Предикаты СО.

Выход: G – количественная оценка меры близости $\mathbf{F_1}$ и $\mathbf{F_2}$.

НАЧАЛО

G = 0; nG = 0; {F_i} = $\mathbf{F_1} \cup \mathbf{F_2}$; // Логическое объединение множеств НПП

$\mathbf{F_{тек}} = \subset\{F_i\}$ | *подвыражения*(F_i)=$\varnothing$; // НПП первичного уровня из{F_i}

ЦИКЛ ПОКА ($\mathbf{F_{тек}} \neq \varnothing$)

НЦ

ЦИКЛ ДЛЯ КАЖДОЙ пары НПП $F_1 \in \mathbf{F_1}$ и $F_2 \in \mathbf{F_2}$ | $F_1 \in \mathbf{F_{тек}}$ *ИЛИ* $F_2 \in \mathbf{F_{тек}}$

Вычислить $\cap(F_1, F_2) = (F_1 \Leftrightarrow F_2$=Обобщение($F_1$, F_2, U), $F_{1ост}$, $F_{2ост})$;

// Допустимый уровень обобщения подвыражений ограничен U

// При U=0 никакого обобщения F_1 и F_2 не производится

$\mathbf{F_{тек}} = \subset\{F_i\}$ | *подвыражения*(F_i)$\subset \mathbf{F_{тек}}$; // возможные НПП след. уровня

КЦ

ЦИКЛ ДЛЯ КАЖДОГО F_i *ИЗ* {F_i}:

ЕСЛИ F_i имеет вид $F_i \Leftrightarrow F_j$ *ТО* G = G+$g(F_i)$ *ИНАЧЕ* nG = nG+$g(F_i)$;

// $g(F_i)$ вычисляется с учетом g(*подвыражений*(F_i))

G = G/(G+nG); // нормировка оценки в диапазон от 0 до 1.

КОНЕЦ.

В **параграфе 4.2** рассмотрено применение оценки близости систем Предикатов СО для трансформации систем НПП в процессе обобщения Предикатов в ходе машинного обучения ИС распознаванию СО Конструкций ЕЯ. Разработанные алгоритмы могут совместно применяться для поиска вариантов выражений сложных Предикатов через комбинацию простых и для обобщения близких Предикатов СО или групп прецедентов. В этом случае формирование сложных систем НПП путем динамической иерархической таксономии Знаний требует организации такого обучения с участием эксперта, координирующего процессы обучения. Алгоритмы можно модифицировать, разрешив согласование фрагментов и вычисление пересечения множеств истинности Предикатов выполнять не только для структурированных значений

аргументов СО целиком, но и для частей аргументов. Работа эксперта в этом случае сведется к корректировке процесса обучения.

Алгоритмы позволяют оценить известность и информативность для ИС Концепций СЭ ЕЯ-высказываний и принять на основе такой оценки решение об останове процесса обобщения Предикатов СО. Установив в алгоритмах соответствующий уровень допустимости обобщения подвыражений, можно получить количественную оценку структурных изменений в Концептуальных Знаниях.

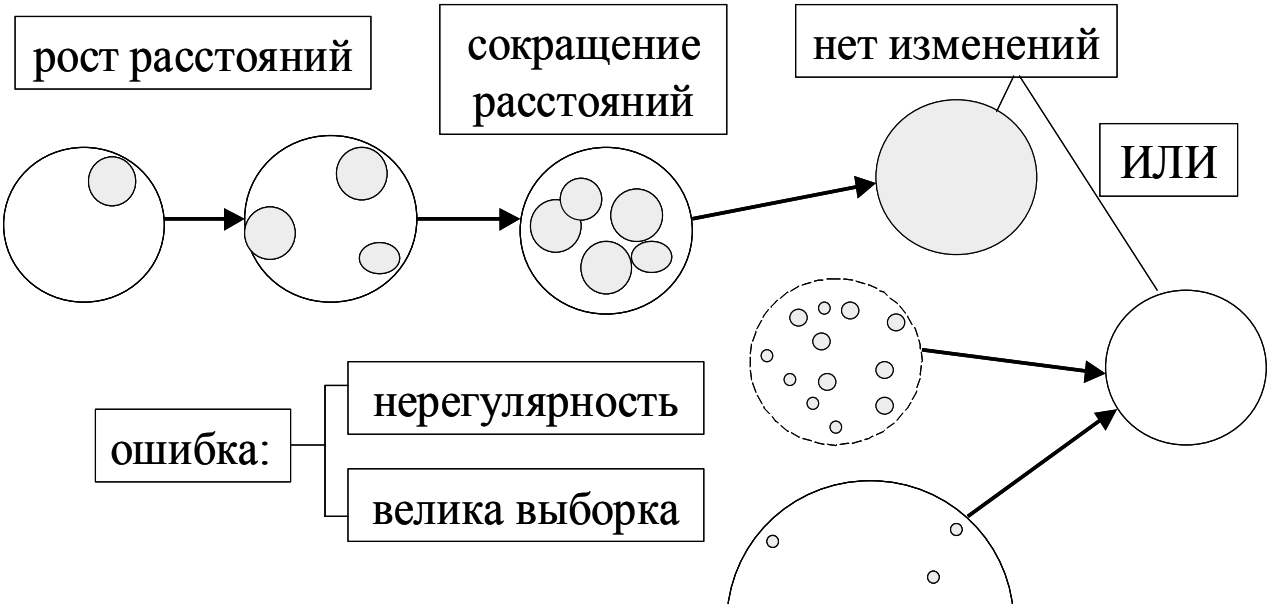

Рисунок 1. Стадии и варианты процесса обучения распознаванию СО.

В **параграфе 4.3** даны практические рекомендации по использованию разработанных методов и алгоритмов. Выделение СО может производится по текстам ЕЯ без формирования обучающей выборки прецедентов СЭ ЕЯ-высказываний. Порождение множества прецедентов СЭ ЕЯ-высказываний, которое отвечает условиям гарантированного формирования структуры Концептуальных Знаний, возможно при организации обучения ИС по схеме «Ученики – Учителя». Использование методов динамической и иерархической таксономии в Концептуальном Анализе позволяет естественным образом разделить этапы и уровни обучения, распараллелить процесс машинного обучения распознаванию СО Конструкций ЕЯ. Алгоритмы обучения распознаванию СО Конструкций ЕЯ могут быть значительно оптимизированы для использования при такой организации машинного обучения.

В **заключении** сформулированы основные научные и практические результаты диссертации и предложены возможные направления дальнейших исследований, связанные с разработкой общеязыковой модели ЕЯ.

Основные результаты диссертационной работы

1. Выявлены наиболее общие функциональные и структурные свойства ЕЯ, на основе чего предложена модель ЕЯ как множества ситуаций ЕЯ-употребления с механизмом проверки ЕЯ-высказываний на допустимость в речи и механизмом согласования употребления форм поверхностного выражения Смысла в разных ситуациях ЕЯ-употребления;
2. Разработана формальная концептуальная модель обучения распознаванию СО Конструкций ЕЯ с помощью обобщения Семантических Знаний, которые представлены Предикатами СО, механизмом Концептуального Анализа, который выявляет закономерности в ЕЯ-употреблении;
3. Установлено, что таксономия множества прецедентов СО механизмом Концептуального Анализа представима с помощью Предикатов в форме НПП над множествами структурированных значений аргументов СО, на основе чего построен формальный аппарат математического моделирования процесса обобщения Предикатов СО в форме НПП;
4. С использованием построенного аппарата решена задача распознавания СО Конструкций ЕЯ путем сравнения систем Предикатов СО в форме НПП, результатом которого является вычисление количественной оценки близости систем Предикатов СО;
5. Доказаны теоремы о алгоритмической разрешимости процесса обобщения Предикатов СО и Процесс распознавания СЭ ЕЯ-высказываний. Также доказаны: теорема о том, что вычислительная сложность процесса обобщения Предикатов СО исходной системы НПП линейно зависит от количества Предикатов в конечной системе НПП, которая является результатом обобщения исходной, и теорема о том, что вычислительная сложность процесса распознавания СЭ ЕЯ-высказываний заданной системе НПП линейно зависит от количества уровней синонимии в этой системе и в худшем случае от количества НПП в этой системе;
6. Показано, что с помощью количественной оценки близости систем Предикатов СО может быть проведена иерархизация сложных систем Предикатов СО; иерархизация позволяет реализовать алгоритм распознавания СО Конструкций ЕЯ, вычислительная сложность которого не экспоненциально, а линейно растет с увеличением числа уровней иерархии синонимии и количества Предикатов;Промоделировано множество синонимических преобразований, известных из лингвистики, включая учитывающие различные уровни синонимий, и сложные виды синонимических преобразований: выходящие за пределы простого предложения и за рамки регулярной Лексико-Функциональной синонимии; при этом разрешены проблемы моделей Семантики Конструкций ЕЯ, основанных на перифразировании; решена задача морфологической, сортовой, родовидовой классификации лексики;
8. Разработаны алгоритмы нахождения количественной оценки близости Предикатов в общем виде с учетом возможных методов оценивания,

оптимизированные для сравнения сложных, иерархизированных систем Предикатов СО; показано применение этих алгоритмов для трансформации систем НПП в процессе обобщения Предикатов СО.

Список опубликованных работ по теме диссертации

1. Емельянов Г. М., Корнышов А. Н., Михайлов Д. В. Концептуально-ситуационное моделирование процесса перифразирования высказываний естественного языка как обучение на основе прецедентов // Искусственный интеллект, №2. – Донецк, 2006. – С.72-75.
2. Емельянов Г. М., Корнышов А. Н., Михайлов Д. В. Концептуально-ситуационное моделирование процесса перифразирования высказываний естественного языка как обучение на основе прецедентов // Интеллектуализация обработки информации : Тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины. – Симферополь, 2006. – С. 78-79
3. Корнышов А. Н. Обучение на основе прецедентов в задаче распознавания смысловой эквивалентности // Сборник тезисов докладов аспирантов, соискателей, студентов XIII научной конференции преподавателей, аспирантов и студентов НовГУ 3-8 апреля 2006 г. – Великий Новгород, 2006. – С. 136.
4. Корнышов А. Н., Михайлов Д. В. Концептуальный уровень и его использование в задаче моделирования синонимических преобразований высказываний естественного языка // Материалы XVIII международной научно-методической конференции “Математика в вузе”. – Великий Новгород, 2005. – С. 118-120.
5. Корнышов А. Н., Михайлов Д. В. Предикаты семантических отношений в задаче моделирования системы концептуальных зависимостей в тезаурусе предметной области // Тезисы докладов аспирантов, соискателей, студентов XIV научной конференции преподавателей, аспирантов и студентов НовГУ. Великий Новгород, 2-7 апреля 2007 г. (В печати).

В издании, рекомендуемом ВАК РФ :

6. Корнышов А. Н., Михайлов Д. В. Концептуально-ситуационное моделирование высказываний естественного языка в задаче анализа их смысловой эквивалентности // Вестник Новгородского государственного университета имени Ярослава Мудрого, серия “Технические науки”, №34. – Великий Новгород, 2005. – С. 76-80.

Автор /Корнышов А.Н./

Используемые в автореферате сокращения

ЕЯ – Естественный Язык
ИС – Интеллектуальные Системы
СЭ – Смысловая Эквивалентность
ЗС – Знаковая Система
СО – Семантические Отношения
НПП – Набор Правил Преобразований