На правах рукописи

МИХАЙЛОВ Дмитрий Владимирович

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ, МЕТОДЫ И АЛГОРИТМЫ ФОРМИРОВАНИЯ ЗНАНИЙ О СИНОНИМИИ ДЛЯ ЗАДАЧ АНАЛИЗА И СЖАТИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора физико-математических наук

Великий Новгород – 2012

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования “Новгородский государственный университет имени Ярослава Мудрого” на кафедре информационных технологий и систем.

Научный консультант -
доктор технических наук, профессор **Емельянов Геннадий Мартинович**

Официальные оппоненты:

Немирко Анатолий Павлович, доктор технических наук, профессор, ФГБОУ ВПО “Санкт-Петербургский государственный электротехнический университет “ЛЭТИ” им. В. И. Ульянова (Ленина)”, профессор кафедры биотехнических систем;

Минаков Игорь Александрович, доктор технических наук, Учреждение Российской академии наук Институт проблем управления сложными системами РАН, старший научный сотрудник лаборатории анализа и моделирования сложных систем;

Чернов Владимир Михайлович, доктор физико-математических наук, ФГБОУ ВПО “Самарский государственный аэрокосмический университет имени академика С.П.Королева (национальный исследовательский университет)”, профессор кафедры геоинформатики и информационной безопасности.

Ведущая организация: Научно-исследовательский институт прикладной математики и кибернетики ФГБОУ ВПО “Нижегородский государственный университет им. Н.И. Лобачевского”.

Защита состоится “15” февраля 2013 г. в 10 часов на заседании диссертационного совета Д 212.215.07, созданного на базе ФГБОУ ВПО “Самарский государственный аэрокосмический университет имени академика С.П.Королева (национальный исследовательский университет)” (СГАУ), по адресу: 443086, Самара, Московское шоссе, 34.

С диссертацией можно ознакомиться в библиотеке СГАУ.

Автореферат разослан “___” _____ 2012 г.

Ученый секретарь
диссертационного совета Белоконов И.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Важнейшей составляющей компьютерного анализа смысла текста является выделение класса семантической эквивалентности (СЭ). Для поисковых и вопросно-ответных систем это позволяет сократить время поиска информации и упростить семантических анализ запроса путём разделение знаний о языке на уровни. В системах машинного перевода иерархия классов СЭ уменьшает число необходимых трансформационных правил и повышает адекватность варианта перевода исходному тексту. В программах обучения языку классы СЭ есть основа знаний о формах выражения нужной мысли в изучаемом языке. В системах тестирования знаний интерпретация ответа на тестовое задание открытой формы (ТЗОФ) есть анализ принадлежности классу СЭ правильного ответа, задаваемого разработчиком теста.

Тем не менее, серьёзных попыток смоделировать на ЭВМ формирование знаний о синонимии в естественном языке (ЕЯ) во взаимосвязи с процессом накопления знаний о языке в целом и об окружающем мире не предпринималось, несмотря на многочисленные публикации, посвященные:

– синтаксису, его связи с семантикой и лексическими средствами языка, реализующими механизм синонимического перефразирования – Мельчук И.А., Жолковский А.К., Гладкий А.В., Апресян Ю.Д., Кибрик А.Е., Тестелец Я.Г., Солганик Г.Я., Тузов В.А. и др.;

– компьютерным словарям, тезаурусу и машинному фонду русского языка – Караулов Ю.Н., Нариньяни А.С., Рубашкин В.Ш., Попов Э.В., Леонтьева Н.Н., Демьянков В.З. и др.;

– системам тестирования знаний – Аванесов В.С., Красильникова В.А., Майоров А.Н., Челышкова М.Б., Останин К.С., Якимов В.Н. и др.;

– информационному поиску – Леонтьева Н.Н., Осипов Г.С., Попов Э.В., Рубашкин В.Ш., Фомичёв В.А., Соснин П.И., Тихомиров И.А., Журавлёв Ю.И., Гуревич И.Б., Кузнецов С.О., Райгородский А.М., Мучник И.Б. и др.

Современные поисковые системы, анализируя ЕЯ-запрос, используют статистику встречаемости слов запроса в различных контекстах с учётом возможных синонимов с целью поиска документа, максимально релевантного запросу. Аналогичный принцип используется и в статистическом переводе. Данный подход полностью оправдывает себя в задаче информационного поиска, но он не позволяет воссоздать целостный образ самой ситуации использования естественного языка для описания фрагмента действительности. Сказанное значимо, в частности, при подготовке ТЗОФ, когда известен фрагмент реальности и разработчику теста требуется выделить все возможные формы описания этого фрагмента в заданном естественном языке.

В связи с этим задача разработки эффективных средств машинного представления знаний о СЭ в совокупности с реализацией механизма взаимодействия знаний о естественном языке и предметной области (ПО) является чрезвычайно актуальной.

Г.М. Емельяновым, Т.В. Кречетовой и Е.П. Курашовой в 1998–2000 гг. была предпринята попытка решить эту задачу с привлечением уровня глубинного синтаксиса ЕЯ в рамках модели СЭ на основе грамматик деревьев (Δ-грамматик). Указанный математический аппарат, предложенный А.В. Гладким и И.А.Мельчуком и расширенный разделением преобразований узлов и ветвей, позволяет формализовать синонимические преобразования ЕЯ-высказываний на уровне универсальной (абстрактной) лексики без существенного ограничения входного ЕЯ и ПО решаемых задач. Но и данному подходу присущи серьёзные недостатки, а именно:

– на уровне глубинного синтаксиса текст представлен фразами, каждая из них соответствует простому распространенному предложению. При этом нельзя говорить о необходимых и достаточных признаках синонимии текстов исключительно по результатам анализа применимости правил синонимических преобразований к деревьям глубинного синтаксиса отдельных фраз и, как следствие, делать выводы о целесообразности трансформаций того или иного типа;

– словарная подсистема предполагается закрытой ввиду существенной сложности описываемой словарём информации;

– отсутствует формализация компонент, отождествляемых с условиями применимости правил синонимических преобразований. По оценке И.А. Мельчука, в теоретическом плане сами правила не претендуют на полноту и возможно их расширение по результатам соответствующих исследований.

Диссертация посвящена разработке методов и алгоритмов формирования знаний о синонимии в естественном языке на основе ситуаций его употребления для описания фрагментов действительности. В данной работе впервые предложено одновременное формирование предметных и языковых знаний непосредственно по текстам, вводимым пользователем без специальной подготовки в области языкознания.

Объект исследования настоящей диссертационной работы – программные средства распознавания, анализа и сжатия текста на естественном языке.

Предметом исследования являются методы и алгоритмы формирования знаний о синонимии.

Цель диссертации заключается в разработке и теоретическом обосновании структуры знаний о синонимии, а также методов и алгоритмов их формирования и использования для совокупности задач оценки семантической схожести текстов предметно-ограниченного естественного языка, автоматизации пополнения и компрессии баз языковых и предметных знаний.

Для достижения поставленной цели в работе решаются следующие *задачи*:

- анализ существующих методов формализации семантики конструкций ЕЯ и определение общих требований, предъявляемых к механизму сравнения смыслов на функциональном уровне;
- разработка и исследование методов анализа СЭ на уровне варьирования абстрактной лексикой;
- разработка методов автоматизированного формирования и кластеризации знаний о семантике конструкций предметно-ограниченного естественного языка с учётом взаимосвязи языковых уровней;
- исследование и алгоритмизация механизма использования морфологии и синтаксиса ЕЯ для задач кластеризации, разделения и сжатия баз предметных и языковых знаний;
- разработка и исследование методов численной оценки семантической схожести текстов предметно-ограниченного естественного языка;
- разработка архитектуры программной системы, реализующей предложенные принципы, методы и алгоритмы.

Методы исследования. Для решения поставленных в работе задач были использованы методы формальной теории языков, математической логики и теории множеств, теории решеток и анализа формальных понятий, системной типологии языков и когнитологии, основные положения теоретической и когнитивной лингвистики, а также прикладные методы анализа данных и знаний.

Научная новизна. В диссертации разработаны теоретические основы автоматизированного формирования знаний о синонимии и их использования для сокращения объёмов баз предметных и языковых знаний в задачах анализа текстов. В частности, новыми являются следующие результаты:

- методика автоматизированного формирования и экспериментальной оценки знаний выделением классов семантической эквивалентности текстов, учитывающая целостный образ ситуации употребления предметно-ограниченного подмножества естественного языка для описания факта действительности;
- подход к решению задачи распознавания сверхфразовых единств в текстах на уровне глубинного синтаксиса. При этом динамическая информационная модель совокупности правил Δ-грамматики сводит поиск последовательности преобразований с заданными свойствами к известным задачам сетей Петри;
- принцип выделения и кластеризации семантических отношений как теоретическая основа формирования смыслового эталона на множестве эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка;
- метод и алгоритмы автоматизированного формирования смыслового эталона на множестве СЭ-фраз в виде решётки формальных понятий, а также метод компрессии текстовой базы знаний на основе выделенных эталонов;
- метод численной оценки семантической схожести текстов предметно-ограниченного ЕЯ с учётом разделения языковых и предметных знаний;
- типовая архитектура программной системы контроля знаний, реализующая предложенные в работе принципы, методы и алгоритмы.

Теоретическая и практическая значимость. Диссертационная работа носит теоретико-прикладной характер. Полученные в ней результаты, разработанные методы и реализующие их программы могут быть использованы для решения широкого класса задач обработки текстов, а также сжатия информации без потери полезной смысловой составляющей. Наряду с ЕЯ-текстами, выделение смысловых эталонов предлагаемыми в работе методами актуально для задач распознавания и анализа семантики любых сложных информационных объектов, в том числе изображений, при формировании баз данных и знаний. Результаты диссертационной работы реализованы в рамках следующих НИР:

1. Грант РФФИ № 03-01-00055-а “Разработка математического аппарата для распознавания сверхфразовых единств в текстах”, рук. Емельянов Г. М., отв. исп. Михайлов Д.В.
2. Грант РФФИ № 06-01-00028-а “Разработка методов автоматизированного пополнения тезауруса для задач распознавания смысловой эквивалентности текстов”, рук. Емельянов Г. М., отв. исп. Михайлов Д.В.
3. Грант РФФИ № 10-01-00146-а “Разработка методов автоматизированного накопления и систематизации знаний о морфологии и синтаксисе естественного языка для задач семантической кластеризации текстов”, рук. Емельянов Г. М., отв. исп. Михайлов Д.В., гос. рег. № 0120.1 164263, 2010-2012 г.
4. Грант № ТОО-3.3-408 Минобразования РФ, отв. исп. Михайлов Д.В.
5. Контракт № И 0675 ФЦП “Интеграция”, отв. исп. Михайлов Д.В., гос. рег. № 0120.0 300918.
6. ГБ НИР “Разработка и исследование математических моделей многопараметрических систем”, рук. Емельянов Г.М., отв. исп. Михайлов Д.В., по заданию Минобрнауки РФ, гос. рег. № 0120.0 704719, 2007-2011 г.

Достоверность теоретических результатов обеспечивается применением апробированного математического аппарата, корректностью изложения основных теоретических положений работы с формулировкой необходимых утверждений, лемм и теорем, строгостью математических доказательств, согласованностью с ранее полученными результатами других авторов. Теоретические положения иллюстрируются примерами реализации компонент программной системы тестирования знаний и решения возникающих при этом инженерных задач.

Личный вклад автора. В диссертационной работе обобщены результаты, полученные лично автором. Постановка и решение задачи распознавания сверхфразовых единств в текстах на уровне глубинного синтаксиса принадлежит автору. Решение задач формирования и кластеризации знаний на основе синтаксического контекста существительного предложено автором как обобщение результатов, полученных совместно с Н.А. Степановой. Теоретические основы формирования знаний о языке на основе ситуаций его употребления развиты автором совместно с А.Н. Корнышовым. Метод оценки семантической схожести текстов предметно-ограниченного ЕЯ, а также метод и алгоритмы выделения смыслового эталона на множестве эквивалентных по смыслу ЕЯ-фраз, метод компрессии текстовой базы знаний и подход к интерпретации ответа испытуемого на тестовое задание открытой формы (включая архитектуру программной системы контроля знаний) разработаны лично автором. Эксперименты на ЭВМ подготовлены и выполнены автором в рамках выпускных квалификационных работ студентов специальностей “Прикладная математика и информатика” и “Программное обеспечение вычислительной техники и автоматизированных систем”.

Апробация работы. Результаты работы представлялись на 35 конференциях, семинарах и конгрессах, в том числе проводимых РАН: 10-й, 12-й, 13-й, 14-й, 15-й Всероссийских конференциях “Математические методы распознавания образов”, 2001, 2005, 2007, 2009, 2011; 6-й, 7-й, 8-й, 9-й, 10-й Международных конференциях “Распознавание образов и анализ изображений: новые информационные технологии”, 2002, 2004, 2007, 2008, 2010; проводимых РАН совместно с Национальными академиями наук Украины и Беларуси 4-й, 5-й, 6-й, 7-й, 8-й Международных конференциях “Интеллектуализация обработки информации”, 2002, 2004, 2006, 2008, 2010.

Публикации. Всего по теме диссертации опубликовано 75 работ, среди них одна монография, 18 статей в журналах, входящих в перечень, рекомендованный ВАК для публикации основных результатов докторских диссертаций. Имеется свидетельство о регистрации программы для ЭВМ. В трудах международных конференций представлено 28 работ, в трудах всероссийских – 7 работ.

Структура и объем диссертации. Диссертация состоит из введения, шести глав, заключения, списка литературы и двух приложений. Общий объем диссертации составляет 333 страницы машинописного текста. Основная часть работы изложена на 237 страницах и содержит 78 рисунков и 15 таблиц. Список литературы включает 188 наименований.

На защиту выносятся следующие основные положения:

1. Методика автоматизированного формирования и экспериментальной оценки знаний, основанная на концепции ситуации употребления естественного языка как единицы формализованного описания его семантики.
2. Подход к нахождению системы целевых выводов в Δ-грамматике как основа выделения сверхфразовых единств в текстах на уровне глубинного синтаксиса.
3. Принцип формирования и кластеризации семантических отношений как основы классов СЭ.

4. Метод и алгоритмы выделения смыслового эталона на множестве эквивалентных по смыслу фраз предметно-ограниченного естественного языка.
5. Численная оценка семантической схожести текстов предметно-ограниченного естественного языка относительно ситуаций его употребления.
6. Метод компрессии текстовой базы знаний с применением смысловых эталонов.

Диссертация включает исследование процессов накопления знаний о синонимии в естественном языке; создание и исследование информационной модели указанного явления; разработку принципов и методов извлечения знаний, а также средств автоматизации построения концептуальной модели предметной области на основе классов СЭ для текстов предметно-ограниченного ЕЯ, что полностью соответствует паспорту специальности 05.13.17 – "Теоретические основы информатики".

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность темы работы, дан краткий обзор современного состояния проблематики и литературы по теме исследования, сформулированы цели и задачи, определена структура диссертации.

Первая глава посвящена общей постановке задачи автоматизированного накопления знаний о синонимии как основы кластеризации предметных и языковых знаний. Вводится понятие ситуации языкового употребления (СЯУ), рассматриваемой в качестве единицы формализованного описания семантики ЕЯ:

$$S = (O, R, Ts), \tag{1.1}$$

где O – множество символов, отождествляемых с некоторыми понятиями; Ts – множество альтернативных форм описания ситуации в некоторой знаковой системе; $R \subset O^n$, где $n \in 1,\ldots,|O|$. Отношения из множества R, как и формы из Ts, могут быть произвольными. В качестве элементов Ts в работе рассматриваются совокупности символьных цепочек (содержательно – семантически эквивалентные ЕЯ-фразы), причём для $\forall Ts_i \in Ts \;\; \exists Tr_i : \; Ts_i = Synt(Tr_i)$, где Tr_i есть ориентированное помеченное дерево, а $Synt$ – сюръективная функция, определяемая правилами синтаксиса языка. Тогда $O = M \cup V$, $M \cap V \neq \varnothing$, где для $\forall o_j \in M$ найдётся $o_k \in V$ такое, что понятию o_j соответствует дочерний узел с пометкой w_j, а понятию o_k – родительский узел с пометкой w_k в Tr_i. Далее будем говорить, что слово, соответствующее символьной цепочке w_j, подчинено (синтаксически) слову, отождествляемому с w_k.

Сама задача СЭ формулируется следующим образом.

Задача 1.1. Дано множество ЕЯ-текстов *G*. Требуется: по результатам синтаксического разбора каждого $g_i \in G$ выявить множества $V(g_i)$ и $M(g_i)$, а также тернарное отношение $I \subseteq G \times M \times V$: $M = \bigcup_i M(g_i)$, $V = \bigcup_i V(g_i)$. Далее на основе I необходимо сформировать множество R и выделить группы текстов по сходству встречаемости понятий в одних и тех же $r_j \in R$.

Задача 1.1 наиболее естественно решается методами анализа формальных понятий (АФП). При этом для $A \subseteq G$ и $B \subseteq M \times V$ вводится пара отображений: $A' = \{(m,v) : m \in M, v \in V \mid \forall g \in A : m(g) = v\}$, $B' = \{g \in G \mid \forall (m,v) \in B : m(g) = v\}$. Па-

ра (A,B), где $A'=B$ и $B'=A$, есть формальное понятие (ФП) с объемом A и содержанием B. Классам СЭ здесь будут соответствовать классы формальных понятий в решётке, а задача накопления знаний о синонимии сводится к совокупности подзадач, решаемых далее в главах:

- формирование прецедентов синонимии для уровня абстрактной лексики;
- кластеризация отношений из множества R в составе тройки (1.1);
- численная оценка схожести СЯУ.

Вторая глава посвящена решению проблемы полноты представления смысла при формировании прецедентов ситуаций синонимии для уровня абстрактной лексики. При этом содержательную основу сжатия смысловой информации составляют сверхфразовые единства на уровне глубинного синтаксиса.

Для теоретического обоснования алгоритмической разрешимости построения последних вводится динамическая информационная модель (в терминологии работ Г.М. Емельянова и Е.И. Смирновой) совокупности правил Δ-грамматики на основе аппарата ограниченных сетей Петри. Рассматриваемые Δ-грамматики задаются четвёрками $\Gamma=(W_R,V_R,\varphi,\Pi)$, где V_R – конечное множество пометок на ветвях дерева: $V_R=\{a_1,a_2,\ldots,a_k\}$; W_R – конечное множество пометок на узлах; φ – матрица ограничений на характер размещения на ветвях дерева пометок из V_R: для $\forall i=1,\ldots,k$ из любого узла дерева выходит не более $\varphi(a_i)=n_i$ ветвей с пометкой a_i; Π – конечное множество правил преобразований деревьев, причём для $\forall rule_j\in\Pi$ задаётся множество Rap условий его применимости. Содержательно $\forall rap_l\in Rap$ выступает в роли прецедента, с которым отождествляется класс СЭ на уровне абстрактной лексики.

Определение 2.1. Лексической синонимической конструкцией (ЛСК) будем далее называть комплекс лексических единиц $wr_k\in W_R$ и связей $vr_j\in V_R$ между ними, замена которого описывается некоторым $rule_i\in\Pi$. Каждой ЛСК соответствует свое ключевое слово C_0, при этом в общем случае произвольная wr_k в составе ЛСК есть значение некоторой лексической функции от C_0.

Представим вход правила $rule_j\in\Pi$ как описание поддерева, заменяемого правилом. Тогда определение возможности применения преобразований из Π к заданному дереву есть анализ применимости каждого $rule_j\in\Pi$, с выделением ключевого слова ЛСК и представлением результата в виде списка пар:

$$\left\{\left(rule_j,C_0(j)\right):\ j=1,\ldots,|\Pi|\right\}. \qquad (2.1)$$

В работе некоторого $rule_j\in\Pi$ в общем случае следует выделить два состояния: соответствующее заменяемому дереву Tio_1 и соответствующее заменяющему дереву Tio_2, где $Tio_k=\langle Wio_k,Vio_k\rangle$, Wio_k – множество узлов, Vio_k – множество ветвей. Условие $rap_l\in Rap$ представляет собой формальное описание допустимости перехода из состояния Tio_1 в Tio_2. Правило $rule_j$ может быть применено к дереву Tio_1, если $\vee_{l=1}^{m} rap_l=true$, где $m=|Rap|$. Обозначим $\vee_{l=1}^{m} rap_l$ далее как r_{12}. При этом r_{12} следует интерпретировать как “определение события, разрешающего переход от Tio_1 к Tio_2”. Применение правила $rule_j\in\Pi$ сводится к выполнению перехода:

$$rule_j(r_{12}): Tio_1 \xrightarrow{rule_j(r_{12})} Tio_2. \qquad (2.7)$$

Отдельному правилу соответствует элементарная сеть Петри вида

$$N = \{P, T, F, H, M_0\}. \qquad (2.8)$$

При этом множество состояний правила есть множество позиций сети $P = \{p_1, p_2\}$, где $p_1 \Leftrightarrow Tio_1$, а $p_2 \Leftrightarrow Tio_2$. Множество возможных переходов T представлено единственным переходом из Tio_1 в Tio_2: $t = rule_j(r_{12}): p_1 \xrightarrow{t} p_2$. Компоненты F и H есть отображения $F: P \times T \to \{0,1\}$ и $H: T \times P \to \{0,1\}$, соответственно. Для сети вида (2.8) $F(p_1, t) = 1$, $F(p_2, t) = 0$, $H(t, p_1) = 0$, $H(t, p_2) = 1$, а число допустимых маркировок (разметок) сети равно двум. Начальной маркировке соответствует вектор $M_0 = (1,0)$, второй из допустимых маркировок – вектор $M = (0,1)$.

Множество правил $rule_j \in \Pi$, представленных сетями (2.8), есть множество объектов-примитивов для построения информационной модели системы правил некоторого подмножества Π с определением структурных взаимосвязей между примитивами. При этом сама система формируется следующим образом: для каждой пары $\{rule_1, rule_2\} \subset \Pi$, $rule_1 \neq rule_2$, в системе либо вход у $rule_2$ является выходом для $rule_1$, либо наоборот, вход у $rule_1$ есть выход для $rule_2$.

Пусть $N_i = \{P_i, T_i, F_i, H_i, M_{0i}\}$ – сеть, построенная из примитивов (2.8).

Теорема 2.1. Сеть N_i является безопасной в течение всего времени функционирования моделируемой системы правил.

Последовательность применяемых правил соответствует последовательности $\tau = (t_{1i}, t_{2i}, \ldots, t_{ki})$ срабатываний переходов:

$$Tio_1 \xrightarrow{rule_1(r_{12})} Tio_2 \xrightarrow{rule_2(r_{23})} Tio_3 \to \ldots \to Tio_k \xrightarrow{rule_k(r_{k,\,k+1})} Tio_{k+1}, \qquad (2.9)$$

где $t_{1i} \Leftrightarrow rule_1(r_{12})$, $t_{2i} \Leftrightarrow rule_2(r_{23})$, … , $t_{ki} \Leftrightarrow rule_k(r_{k,\,k+1})$. При этом происходит последовательная смена разметок:

$$M_{0i} \xrightarrow{t_{1i}} M_{1i} \xrightarrow{t_{2i}} M_{2i} \to \ldots \to M_{k-1,\,i} \xrightarrow{t_{ki}} M_{ki}, \qquad (2.10)$$

где $M_{0i} \Leftrightarrow Tio_1$, $M_{1i} \Leftrightarrow Tio_2$, … , $M_{k-1,\,i} \Leftrightarrow Tio_k$, $M_{ki} \Leftrightarrow Tio_{k+1}$.

Множество разметок, достижимых из начальной разметки M_{0i} и образующих множество достижимости сети N_i, находится в зависимости от задания M_{0i}. Функционирование системы описывается в терминах последовательностей срабатываний переходов $t_{1i}, t_{2i}, \ldots, t_{k-1,\,i}, t_{ki}$, каждая из которых есть слово τ в языке $L(N_i)$, называемом свободным языком сети N_i.

Задача приведения деревьев Tio_1 и Tio_{k+1} к виду с одинаковой ЛСК фактически включает в себя три задачи:

1) определение достижимости разметки M_{ki} из начальной разметки M_{0i}. Данная задача есть поиск слова $\tau \in T_i^* \,|\, M_{0i} \xrightarrow{\tau} M_{ki}$, где T_i^* – множество всех слов в алфавите T_i;

2) задача обратимости слова τ: если $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_{ki} \right.$, то существует ли слово $\tau' = \left(t'_{ki}, t'_{k-1,i}, \ldots, t'_{2i}, t'_{1i}\right)$:

$$M_{0i} \xleftarrow{t'_{1i}} M_{1i} \xleftarrow{t'_{2i}} M_{2i} \leftarrow \ldots \leftarrow M_{k-1,i} \xleftarrow{t'_{ki}} M_{ki}, \qquad (2.11)$$

где $M_{0i} \Leftrightarrow Tio_1$, $M_{1i} \Leftrightarrow Tio_2$, … , $M_{ki} \Leftrightarrow Tio_{k+1}$;

3) задача определения оптимального слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_{ki} \right.$. Суть: если существуют $\tau_1, \tau_2, \ldots, \tau_l$: $M_{0i} \xrightarrow{\tau_1} M_{ki}$, $M_{0i} \xrightarrow{\tau_2} M_{ki}$, …, $M_{0i} \xrightarrow{\tau_l} M_{ki}$, то в качестве оптимального берется слово наименьшей длины, причём предпочтение всегда отдаётся обратимому слову.

Для решения указанных задач проводится исследование языка $L(N_i)$.

Лемма 2.2. Проблема достижимости заданной разметки M_{ki} из начальной M_{0i} в сети N_i разрешима.

Обозначим множество всех слов в алфавите T_i как T_i^*.

Теорема 2.3. Проблема определения обратимости слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_{ki} \right.$ языка $L(N_i)$ разрешима.

Теорема 2.4. Проблема поиска оптимального слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_{ki} \right.$ в языке $L(N_i)$ является разрешимой.

Таким образом, во второй главе предложены теоретические основы сжатия информации для прецедентов классов СЭ уровня абстрактной лексики. При этом динамическая информационная модель системы правил Δ-грамматики сводит поиск последовательности преобразований с заданными свойствами к классическим задачам теории сетей Петри.

В третьей главе решается задача формирования и классификации отношений из множества R в составе тройки (1.1). Базовым здесь является прецедент класса СЭ, представляемый условием r_{ij} в (2.7) и (2.9). За основу его формализации берётся введённое Б.Х. Парти и В.Б. Борщевым описание семантики символьной цепочки, соответствующей ЕЯ-слову и обозначающей некоторое $o_i \in O$, совокупностью λ-выражений, каждое из которых описывает некоторое свойство понятия o_i. Назовём далее указанную совокупность теорией лексического значения (ЛЗ) слова. Сама теория ЛЗ слова w_i, заменяемого некоторым $rule_j \in \Pi$, определяется рекурсивно посредством упорядоченной совокупности троек и пар (3.2)–(3.4), связывающих обозначаемое словом w_i понятие $o_i \in O$ с другими понятиями множества O через отношения из множества R:

$$Lm(w_i) = (w_i, LM), \qquad (3.1)$$

при этом отдельный элемент Mp списка LM может представлять либо бинарное отношение между парой понятий $\{o_1, o_2\} \subset O$:

$$Mp = (r_2, o_1, o_2), \qquad (3.2)$$

либо рекурсивно определяемое отношение произвольной арности:

$$Mp = (r_n, o, LM_r), \qquad (3.3)$$

либо

$$Mp = (r_c, LM_r), \tag{3.4}$$

где $r_c \in \{\vee, \&, \neg\}$; LM_r определяется по аналогии с LM; r_2 и r_n – символы (либо символьные цепочки), обозначающие соответствующие отношения.

Для автоматизации получения знаний, представляемых формулами вида (3.1)–(3.4), в **разделе 3.5** решается задача формирования множества R на основе множеств СЭ-фраз предметно-ограниченного ЕЯ. При этом отношения в рамках троек и пар (3.2)–(3.4) будут составлять подмножество множества R.

Рассмотрим $Ts_i \in Ts$ с точки зрения составляющих её символов. У каждой Ts_i выделяется неизменная часть Tc_i, общая для всех $Ts_i \in Ts$, и флективная часть Tf_i. На множестве Tf_i выражаются синтагматические зависимости, которые задаются синтаксическими отношениями и определяют возможность сосуществования словоформ в линейном ряду. Аналогично для слова w_{ij} имеем $W_{ij} = Wc_{ij} \bullet Wf_{ij}$, где W_{ij} – последовательность его символов, $Wc_{ij} \subset Tc_i$ составляют символы неизменной части, именуемой далее основой, $Wf_{ij} \subset Tf_i$ – символы флективной части (флексии), а символом "$\bullet$" обозначается конкатенация символьных последовательностей. Для формирования множества R попарным сравнением W_{ij} различных Ts_i требуется найти:

1) Wc_{ij} и Wf_{ij} каждого W_{ij} при $|Wc_{ij}| \to \max$;

2) отношение R_q, определяющее допустимость сочетания (Wf_{ij}, Wf_{ik}), $k \neq j$.

Введём индексное множество J для неизменных частей всех слов, употребленных во всех $Ts_i \in Ts$. Тогда упорядоченная совокупность индексов $j \in J$ неизменных частей слов, присутствующих в $Ts_i \in Ts$, будет моделью линейной структуры этой фразы (далее обозначается как $Ls(Ts_i)$). Для построения множества R необходимо найти совокупность указанных моделей, отвечающих требованиям проективности.

Пусть $h(j, Ls(Ts_i))$ – позиция индекса j в модели $Ls(Ts_i)$. Тогда множество связей для $Ls(Ts_i)$ определяется как $D: Ts_i \to \{(h(j, Ls(Ts_i)), h(k, Ls(Ts_i))) : j \neq k\}$.

Определение 3.3. Связь $d_{qi} = (h(j, Ls(Ts_i)), h(k, Ls(Ts_i)))$ является *допустимой для* $Ls(Ts_i)$, если $\exists\{Ts_l, Ts_m\} \subset Ts$, $l \neq m$, причем и $Ls(Ts_l)$, и $Ls(Ts_m)$ содержат в качестве подпоследовательности либо $\{j,k\}$, либо $\{k, j\}$. При этом пара (j,k) содержательно соответствует одной синтагме.

Положим, что для $\forall Ts_i \in Ts$, $i = 1, \ldots, |Ts|$, все $d_{qi} \in D(Ts_i)$ удовлетворяют *определению 3.3*.

Определение 3.4. Будем считать модель $Ls(Ts_i)$ проективной относительно множества R в (1.1), если $\sum_{q=1}^{|D(Ts_i)|} \Delta_{qi} \le |Ls(Ts_i)|$, где $\Delta_{qi} = |h(j, Ls(Ts_i)) - h(k, Ls(Ts_i))|$.

На основе $\bigcup_i D(Ts_i)$ формируется граф синтагм (V_J, I_J). Элементами множества вершин V_J являются множества пар (j,k), $\{j,k\} \subset J$, сгруппированных по некоторому индексу k. Множества E_1 и E_2, входящие в V_J, будут соединены ребром из

I_J, если $\exists\{j,k,m\}\subset J:\ (j,k)\in E_1,\ (k,m)\in E_2$ и $j\neq m$. Анализом (V_J, I_J) строится дерево синтаксических связей (V_{JT}, I_{JT}). Формально

$$V_{JT}=J,\ I_{JT}=\{(j,k):\exists E\in V_{JT},(j,k)\in E\}. \qquad (3.11)$$

При этом $k\in V_{JT}$ соответствует корню дерева (3.11), если $\exists E_1\in V_J$, в котором пары индексов сгруппированы по k, $|E_1|>1$, а k не содержится ни в одной паре индексов для $\forall E_2\in V_J:\ E_1\neq E_2$.

Замечание. Число дочерних узлов у корня дерева (3.11) полагается не менее двух, поскольку содержательный интерес для формирования R в (1.1) представляют ситуации действительности с двумя и более участниками.

Рассмотрим построение дерева (3.11) для случая расщепленного предикатного значения (РПЗ) как совокупности вспомогательного предикатного слова-связки и слова, называющего ситуацию. Пусть $Tcnc_i=\{w_{ij}: w_{ij}=\bullet(W_{ij})\}$, где символом "$\bullet$" обозначается конкатенация, последовательно выполняемая над символами из Wf_{ij}. Положим, что $\exists Tp_i\subset Ts_i$ определяющая последовательность $Pcnc_i=\{u_k: u_k=\bullet(Wp_k), \bigcup_k Wp_k=Tp_i\}$, где $Wp_k\in Ts_i$ – последовательность символов слова, для которого не выделены неизменная и флективная часть.

Теорема 3.1. Последовательность $Pcnc_i$ содержит предикатное слово, если $\exists\{j,0,k\}\subset Ls(Ts_i):\ \{w_{ij},u_1,\ldots,u_p,w_{ik}\}\subset Tcnc_i$, где $\{u_1,\ldots,u_p\}=Pcnc_i$, $p=|Pcnc_i|$.

Доказательство следует из определения корня дерева (V_{JT}, I_{JT}) и проективности $Ls(Ts_i)$. Пусть для $Pcnc_i$ выполняется условие *теоремы 3.1*.

Теорема 3.2. Слово $u_k\in Pcnc_i$ принадлежит расщеплённому предикатному значению, если $\exists Ts_j\in Ts:\ Ls(Ts_j)\neq Ls(Ts_i)$, а $u_k\in Pcnc_j$, причём $Pcnc_j$ также отвечает условию *теоремы 3.1*. При этом $\neg\exists Ts_k\in Ts$, где $Pcnc_k\subset Pcnc_i$ и отвечает *теореме 3.1*, а $Ls(Ts_k)\neq Ls(Ts_j)$ и $Ls(Ts_k)\neq Ls(Ts_i)$.

Доказательство следует из доказанной *теоремы 3.1* и определения множества ребер в графе (V_J, I_J).

Замечание. При выполнении условия *теоремы 3.2* u_k может быть в том числе и зависимым словом в составе РПЗ.

Пусть $Pcnc_i'$ – последовательность слов, удовлетворяющих *теореме 3.2*, а $Ts'\subset Ts$, при этом $Ts'=\{Ts_i:|Pcnc_i'|\to\max\}$.

Для $\forall u_k\in\bigcup_i Pcnc_i'$, $Ts_i\in Ts'$, его неизменная и флективная часть выделяются сравнением последовательности Wp_k его символов с аналогичными последовательностями Wp_j для всех $u_j\in\bigcup_l Pcnc_l:\ Ts_l\in(Ts\setminus Ts')$, а $Pcnc_l$ отвечает условию *теоремы 3.1*. При этом необходимо, чтобы $2|Wc_k|>|Wf_k|+|Wf_j|$, где $Wp_k=Wc_k\bullet Wf_k$, а $Wp_j=Wc_k\bullet Wf_j$.

Замечание. Если $Pcnc_i'\cap Pcnc_i\neq\varnothing$, то $\forall u_m\in(Pcnc_i\setminus Pcnc_i')$ представляется вместе со словом слева от него в $Pcnc_i$ (в этом случае u_m рассматривается как предлог).

С учетом $Pcnc_i'$ дерево (3.11) преобразуется следующим образом:

1) корень изменяется с $k=0$ на значение k для $u_k \in Pcnc_i'$ с максимальной встречаемостью в разных $Tcnc_i$ относительно заданной СЯУ;
2) левое поддерево остается без изменений;
3) правое поддерево перевешивается на узел j для $u_j \in Pcnc_i'$ наименьшей встречаемости;
4) в паре $\{u_l, u_m\} \subset Pcnc_i'$ дочерний узел у слова с меньшей встречаемостью.

Далее назовём дерево (3.11), преобразованное согласно указанным правилам, расширенным деревом (3.11). Заметим, что расширенное дерево (3.11) является деревом-прецедентом для множества деревьев $\{Tr_i : Ts_i = Synt(Tr_i)\}$ из определения компонента Ts в составе тройки (1.1).

Таким образом, в третьей главе разработан принцип формирования и кластеризации семантических отношений выделением синтагматических зависимостей. Его программная реализация, представленная в **приложении 1** диссертации фрагментами исходного текста на языке Visual Prolog 5.2, позволяет выделять произвольные отношения в рамках СЯУ за время, оцениваемое сверху как квадрат произведения числа СЭ-фраз и максимального числа слов во фразе.

Четвертая глава посвящена задаче минимизации оптимального слова в языке сети Петри, построенной из примитивов вида (2.8). Основу решения составляет выделение ситуаций синонимических замен на уровне абстрактной лексики (синонимов, конверсивов и расщеплённых предикатных значений) в последовательностях синтаксически соподчинённых слов:

$$Sq_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}, \tag{4.1}$$

где v_1 – предикатное слово; m_{ki} и $\forall v_l \in \{v_2, \ldots, v_{n(k,i)}\}$ – существительные.

Утверждение 4.2. При $R_q(v_1, v_2) = true$ возможно установление указанного отношения между v_1 и $\forall v_l \in \{v_3, \ldots, v_{n(k,i)}, m_{ki}\}$.

Замечание. На основании *утверждения 4.2* справедливо будет утверждать, что $\forall v_l \in \{v_2, \ldots, v_{n(k,i)}\}$ в составе последовательности (4.1) обозначает некоторое понятие, значимое в ситуации v_1, наравне с m_{ki}. Таким образом, если в *задаче 1.1* в качестве множества G рассматривать множество Ts в составе тройки (1.1), то для любой Sq_{ki} $\{v_2, \ldots, v_{n(k,i)}, m_{ki}\} \subset M(Ts_i)$, а $V(Ts_i) = \bigcup_k (Sq_{ki} \setminus \{m_{ki}\})$.

В главе рассматривается концептуальная кластеризация текстов методами АФП на основе последовательностей (4.1). Описываются алгоритмы формирования множеств $M(Ts_i)$, $V(Ts_i)$ и отношения I на основе синтаксического разбора исходных $g_i \in G$ согласно постановке *задачи 1.1*, а также порядок замены конверсивов и расщепленных предикатных значений.

Обозначим функцию, которая ставит в соответствие каждому $v \in V(Ts_i)$ предлог для связи с зависимым словом, как $prep : v \to p_y$; функцию, ставящую в соответствие именному $m \in M(Ts_i)$ символьное обозначение его падежа – как $case : m \to c_y$. Соответст-

вие между словом и его начальной формой зададим функцией $norm$. Пусть $\{Ts_1, Ts_2\}$ – пара анализируемых ЕЯ-фраз. Положим, для Ts_1 выделено множество последовательностей вида (4.1), обозначаемое как $SQ_1 = \{Sq_{k1} : Sq_{k1} \subset Ts_1\}$, $k = 1,\ldots, n(SQ_1)$, аналогично для Ts_2 имеем $SQ_2 = \{Sq_{k2} : Sq_{k2} \subset Ts_2\}$, но при этом либо $k = 1,\ldots, n(SQ_1)$, либо $k = 1,\ldots, n(SQ_1)-1$, где $n(SQ_1) = |SQ_1|$.

Утверждение 4.4. Применительно к паре $\{SQ_1, SQ_2\}$ имеет место конверсив, если для $\forall\, Sq_{k1} \in SQ_1$ найдется $Sq_{j2} \in SQ_2$ такая, что при этом могут иметь место следующие случаи взаимного соответствия Sq_{k1} и Sq_{j2}.

1) $Sq_{k1} = \{v_{11}', v_{k2}, v_{k3}, \ldots, v_{k,idx(k,1)}, m_{k1}\}$, $Sq_{j2} = \{v_{21}', v_{k2}', v_{k3}, \ldots, v_{k,idx(k,1)}, m_{k1}\}$.
   При этом $norm(v_{11}') = norm(v_{21}')$, $norm(v_{k2}) = norm(v_{k2}')$, причем в общем случае $prep(v_{11}') \neq prep(v_{21}')$, а $case(v_{k2}) \neq case(v_{k2}')$. Функция $idx(k,i)$ возвращает максимальное значение второго индекса при v.
2) $Sq_{k1} = \{v_{11}', v_{12}', v_{k2}, v_{k3}, \ldots, v_{k,idx(k,1)}, m_{k1}\}$, $Sq_{j2} = \{v_{21}', v_{k2}', v_{k3}, \ldots, v_{k,idx(k,1)}, m_{k1}\}$.
   Здесь $norm(v_{k2}) = norm(v_{k2}')$, $case(v_{k2}) \neq case(v_{k2}')$ (в общем случае), но при этом для Sq_{j2} $\exists\, Sq_{k1}' \in SQ_1$: $\{Sq_{k1}', Sq_{j2}\}$ соответствует *случаю 1*, $Sq_{k1}' \neq Sq_{k1}$, а для Sq_{k1} $\exists\, Sq_{j2}' \in SQ_2$: $\{Sq_{k1}, Sq_{j2}'\}$ также удовлетворяет требованию *случая 1* настоящего *утверждения и* $Sq_{j2}' \neq Sq_{j2}$.

Таким образом, в четвёртой главе принцип формирования и экспериментальной оценки знаний в виде классов СЭ согласно постановке *задачи 1.1* развит применительно к наличию конверсивов и РПЗ в анализируемых текстах. Критерием выбора возможного варианта замены конверсива либо РПЗ здесь является минимум многозначности при максимальном числе беспредложных смысловых валентностей слова, на которое производится замена. При этом степень многозначности определяется числом СЯУ, в которых фигурирует слово.

Пятая глава посвящена совместному использованию свойств расширенного дерева (3.11) и последовательности вида (4.1) для оценки семантической схожести текстов относительно СЯУ, порождаемых независимо друг от друга.

В **разделе 5.1** индексное множество J, рассмотренное в **разделе 3.5**, определяется для неизменных частей всех слов, употребленных в более чем одной фразе из множества Ts в (1.1), с учетом возможного присутствия слова не во всех фразах указанного множества. При этом удвоенная длина общей неизменной части пары слов всегда больше суммы длин флективных частей.

Пусть LS есть множество моделей линейных структур фраз из Ts на J.

Теорема 5.1. Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам, если $\exists \{Ls(Ts_1), Ls(Ts_2)\} \subseteq LS : Ls(Ts_1) = J_1 \bullet \{j_1\} \bullet J_2$ и $Ls(Ts_2) = J_1 \bullet \{j_2\} \bullet J_2$, где $J_1 \subset J$, $J_2 \subset J$, а "$\bullet$" есть операция типа конкатенации над множеством J.

Пусть PJ – множество пар, отвечающих *теореме 5.1*. Заменим индексы, вошедшие в пары из PJ, на некоторые $j \in (\mathrm{N} \setminus J)$ во всех моделях из LS. Обозначим преобразованное LS как LS', множество заменяемых индексов – как JP, а множество индексов, на которые идёт замена, – как JP', $JP' \cap JP = \varnothing$. Фактически каждая модель в LS' задается на множестве $(J \setminus JP) \cup JP'$.

Пусть JN есть множество индексов с максимальной встречаемостью в разных моделях из LS', $Ls_1(Ts_i) \in LS'$, а $Ls_2(Ts_i)$ – модель линейной структуры Ts_i относительно JN. Обозначим множество моделей второго вида как LJN. Положим также, что имеется $LS'_j \subset LS'$ такое, что для всех $Ls_1(Ts_i) \in LS'_j$ модели $Ls_2(Ts_i)$ одинаковы и соответствуют некоторой $Ls_2(Ts_j) \in LJN$, $Ts_j \in Ts$.

Обозначим множество индексов $j \notin JN$ с максимальной встречаемостью в различных $Ls_1(Ts_i) \in LS'_j$, как JA. Местоположение индекса в расширенном дереве (3.11) и флективные части для слов с индексами из $((J \setminus JP) \cup JP') \setminus (JN \cup JA) \cup \{0\}$ определяются аналогично словам из $Pcnc'_i$ описанным в **разделе 3.5** способом. При этом вместо индексов с ненулевым значением рассматриваются $j \in (JN \cup JA)$.

Для *численной оценки* схожести СЯУ, каждая из которых описывается тройкой (1.1), в **разделе 5.2** вводится представление СЯУ в виде совокупности трёх составляющих, называемой в теории АФП формальным контекстом (ФК):

$$Ks = (Gs, Ms, Is), \tag{5.1}$$

где Gs включает основы слов $w_j \in \bigcup_{i=1}^{|Ts|} Ts_i : \exists \left(w_k \in \bigcup_{i=1}^{|Ts|} Ts_i, Tr_i : \exists Ts_i = Synt(Tr_i) \right)$, при этом w_j соответствует дочернему, а w_k - родительскому узлу в Tr_i (w_k есть синтаксически главное для w_j, w_j - синтаксически зависимое по отношению к w_k в дереве Tr_i); $\forall m_i \in Ms$ есть символьная цепочка, понимаемая как некоторый признак некоторого $g_i \in Gs$, сами признаки могут быть следующих видов, составляющих непересекающиеся подмножества множества Ms и обозначаемых далее посредством соответствующего нижнего индекса:

– указания на основу синтаксически главного слова (индекс 1);

– указания на флексию главного слова (индекс 2);

– связи “основа – флексия” для синтаксически главного слова (индекс 3);

– сочетания флексий зависимого и главного слова (индекс 4). После флексии главного слова через двоеточие при необходимости указывается предлог для связи главного слова с зависимым;

– указания на флексию зависимого слова (индекс 5).

Посредством $Is \subseteq Gs \times Ms$ отношения из множества R в (1.1) разбиваются на классы по сходству основы главного, флексии зависимого слова, а также характеру сочетаний основ и флексий. Для численной оценки схожести СЯУ выполняется редукция ФК (5.1) исключением объектов и признаков РПЗ согласно правилу, очевидным образом вытекающему из *теоремы 5.1* и *утверждения 4.4*.

Пусть $\{m_1, m_2, m_3\} \subset M_1$. Если m_1, m_2 и m_3 взаимно различны, то m_1 соответствует указанию на основу главного, m_2 – зависимого слова РПЗ, а m_3 – на основу однословного эквивалента РПЗ при выполнении трех условий:

1. $\exists g_1 \in Gs : Is(g_1, m_1) = true$, $Is(g_1, m_3) = false$, $m_2 = p_{bs} \bullet g_1$. Здесь p_{bs} есть обозначение символьной константы *"главное – основа:"*.
2. $\exists \{g_2, g_3\} \subset Gs$, при этом объекты g_1, g_2 и g_3 взаимно различаются, а $Is(g_2, m_3) \wedge Is(g_3, m_3) \wedge$
$$\wedge (Is(g_2, m_1) \wedge Is(g_3, m_2) \vee Is(g_2, m_2) \wedge Is(g_3, m_1)) = true.$$
3. Не существует других троек объектов, для которых признак m_3 занимал бы место либо m_1, либо m_2 в вышеуказанных соотношениях.

Помимо редукции формальных контекстов (5.1) отдельных СЯУ, для численной оценки их схожести, представленной далее в **разделе 5.5**, вводится представление тезауруса ПО в виде формального контекста:

$$Kth = (Gth, Mth, Ith), \tag{5.2}$$

где множество Gth состоит из символьных пометок отдельных СЯУ. Множество Mth включает элементы множеств признаков формальных контекстов вида (5.1) всех $gth \in Gth$. Кроме того, в составе Mth выделяются:

– множество указаний на объекты формальных контекстов вида (5.1), генерируемых для элементов Gth (обозначим далее это множество как M_6);

– множество связей "основа – флексия" для зависимого слова (M_7);

– множество сочетаний основ зависимого и главного слова (M_8).

Пусть СЯУ S_1 описывается тройкой вида (1.1) и соответствует заведомо корректному ЕЯ-описанию некоторого факта заданной ПО. Положим также, что S_2 – анализируемая СЯУ. Обозначим ФК вида (5.1): для S_1 – как Ke, а для S_2 – как Kx, где $Ke = (Ge, Me, Ie)$ и $Kx = (Gx, Mx, Ix)$, $Ie \subseteq Ge \times Me$ и $Ix \subseteq Gx \times Mx$, соответственно. Введем обозначения для констант: p_{fl} – для "флексия:", p_b – для "основа:". Результат объединения M_6, M_7, M_8, Me_4, Mx_4, Me_5 и Mx_5, обозначим как M_U.

Определение 5.1. Будем считать, что S_1 и S_2 связаны отношением схожести, если каждому объекту $gx \in Gx$ соответствует такой объект $ge \in Ge$, что выполняется одно из следующих условий:

(1) $gx = ge$ и любой признак $me \in Me$ объекта ge относится и к gx.

(2) $gx = ge$, при этом *условие (1)* не выполняется, но существует $gth \in Gth$, обладающий признаком $mth_1 \in M_6$: $mth_1 = p_b \bullet ge$ при обязательном выполнении следующих условий:

$(\exists me_{fl} \in Me_5 : me_{fl} = p_{fl} \bullet fe) \rightarrow (\exists mth_{17} \in M_7 : mth_{17} = ge \bullet ":" \bullet fe)$,

при этом $(Ie(ge, me_{fl}) \wedge Ix(ge, me_{fl})) \rightarrow Ith(gth, mth_{17})$;

$(\exists me_{bs} \in Me_1 : me_{bs} = p_{bs} \bullet be) \rightarrow (\exists mth_{18} \in M_8 : mth_{18} = ge \bullet ":" \bullet be)$,

при этом $Ie(ge, me_{bs}) \rightarrow Ith(gth, mth_{18})$;

$(\exists\, mx_{bs} \in Mx_1 : mx_{bs} = p_{bs} \bullet bx) \rightarrow (\exists\, mth_{28} \in M_8 : mth_{28} = ge \bullet ":" \bullet bx)$,

при этом $Ix(ge, mx_{bs}) \rightarrow Ith(gth, mth_{28})$.

Кроме того, для $\forall\, mth \in (Mth \setminus M_U)$ истинно:

$$Ith(gth, mth) \rightarrow (Ie(ge, mth) \wedge Ix(ge, mth)). \qquad (5.3)$$

(3) $gx \neq ge$, но существует объект $gth \in Gth$, обладающий признаками $mth_1 \in M_6$: $mth_1 = p_b \bullet ge$ и $mth_2 \in M_6$: $mth_2 = p_b \bullet gx$, при этом для любого признака $mth \in (Mth \setminus M_U)$ справедливо:

$$Ith(gth, mth) \rightarrow (Ie(ge, mth) \wedge Ix(gx, mth)). \qquad (5.4)$$

(4) $gx \neq ge$, но существует объект $gth_1 \in Gth$, обладающий признаком $mth_1 \in M_6$: $mth_1 = p_b \bullet ge$, а для $\forall\, me \in (Me_4 \cup Me_5)$ верно:

$$(Ith(gth_1, mth_1) \wedge Ie(ge, me)) \rightarrow Ith(gth_1, me).$$

При этом существуют признаки $mth_2 \in M_6$: $mth_2 = p_b \bullet gxg$ и $mx \in (Mx_1 \cup Mx_2 \cup Mx_3)$, для которых верно:

$$(Ith(gth_1, mth_2) \wedge Ix(gx, mx)) \rightarrow Ith(gth_1, mx),$$

где $gxg \neq gx$, а пара (gxg, ge) отвечает *условию (3)* при генерации ФК вида (5.1) для объекта gth_1. В то же время существует объект $gth_2 \in Gth$, относительно которого пара (gx, gxg) также будет отвечать *условию (3)* настоящего *определения*. Генерируемый при этом формальный контекст вида (5.1) для gth_2 обозначим как *Kxg*, $Kxg = (Gxg, Mxg, Ixg)$.

Замечание. Оценка схожести ситуаций S_1 и S_2 включает сравнение последовательностей двух и более соподчиненных слов. Выполнимость условий *определения 5.1* анализируется только для главных слов. Последовательности считаются заменяемыми, если возможно их построение по формальному контексту (5.2) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ.

С учётом сопоставления согласно *определению 5.1* объектов формальных контекстов $Ke = (Ge, Me, Ie)$ и $Kx = (Gx, Mx, Ix)$, из которых удалена информация РПЗ, *схожесть* ситуаций S_1 и S_2 *численно оценивается* как

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \qquad (5.5)$$

где $n = |Gx|$, а spc_k есть значение схожести объектов в паре (gx_k, ge). В зависимости от выполнимости условий *определения 5.1* значение spc_k либо равно 1,0, если для пары (gx_k, ge) выполнено *условие (1)*, либо вычисляется по формуле:

$$-\log_2\left(1 - \frac{D_c}{path_C}\right) \times \frac{|BLCS|}{|B_1 \setminus BLCS| + |B_2 \setminus BLCS| + |BLCS|}, \qquad (5.6)$$

если для пары (gx_k, ge) выполнено *условие (2), (3)* либо *(4)*.

Во втором случае имеем гипотетическую решетку ФП (обозначим её как $\Re xe$), в которой объемы объектных ФП (формальных понятий с одним объектом в составе

объема) есть $\{gx_k\}$ и $\{ge\}$ (при выполнении *условия (2)* или *(3)*) либо $\{gx_k\}$, $\{ge\}$ и $\{gxg\}$ (при выполнении *условия (4)*). Значение D_c равно числу сравнимых формальных понятий, составляющих цепочку с вершинным ФП решетки $\Re xe$ в качестве максимального ФП и наименьшим общим суперпонятием (НОСП) для объектных формальных понятий решетки $\Re xe$ – в качестве минимального ФП. Множество $BLCS$ есть содержание (множество признаков всех объектов) этого НОСП, а число $path_C$ равно минимальному числу ФП в цепочке, которой принадлежит вершинное ФП, наименьшее ФП решетки $\Re xe$ и формальное понятие с содержанием $BLCS$.

В случае выполнения любого из *условий (2)*, *(3)* или *(4)* значение $D_c = 2$.

При выполнении *условия (2)* либо *(3)* $path_C = 4$, а в $BLCS$ войдут признаки $mth \in (Mth \setminus M_U)$, для каждого из которых справедливо либо соотношение (5.3) (при выполнении *условия (2)*), либо соотношение (5.4) (при выполнении *условия (3)*). Множества B_1 и B_2 в этом случае определяются следующим образом:

$$B_1 = \{ me : me \in (Me_1 \cup Me_2 \cup Me_3),\ Ie(ge, me) = true\},$$
$$B_2 = \{ mx : mx \in (Mx_1 \cup Mx_2 \cup Mx_3),\ Ix(gx_k, mx) = true\}.$$

Доказательство выполнимости *условия (4)* обычно происходит в несколько итераций. При этом в ходе каждой последующей итерации число признаков, не являющихся общими для gx_k и gxg, всегда меньше, чем в предыдущей. Начальное значение $path_C$, равное 4, в ходе каждой итерации увеличивается на 1, а

$$B_1 = \{mxg : mxg \in (Mxg_1 \cup Mxg_2 \cup Mxg_3),\ Ixg(gxg, mxg) = true\},$$
$$B_2 = \{mx : mx \in (Mxg_1 \cup Mxg_2 \cup Mxg_3),\ Ixg(gx_k, mx) = true\},$$

где $(Mxg_1 \cup Mxg_2 \cup Mxg_3) \subset Mxg$ в соответствии с показанным выше разделением множества признаков формального контекста вида (5.1), а $BLCS = B_1 \cap B_2$.

Далее в **разделе 5.5** приводится пример интерпретации ТЗОФ с вычислением оценок (5.5).

Таблица 1

Сопоставление ответов правильному варианту

ответы	правильный вариант				анализируемый		
вариант	1	2	3	4	1	2	3
основа	флективная часть + предлог						
заниженн	ости	ости	ость	ость	ость	ость	ости
эмпирическ	ого	ого	ого	ого	–	–	–
риск	а	а	а	а	–	–	–
средн	–	–	–	–	ей	ей	ей
ошибк	–	–	–	–	и:на	и:на	и:на
обучающ	–	–	–	–	ей	ей	ей
выборк	–	–	–	–	е	е	е
переобучении	е	–	–	ем	ем	–	е
переподгонк	–	а	ой	–	–	ой	–
связан	–	–	а:с	а:с	а:с	а:с	–
привод	ит:к	ит:к	–	–	–	–	ит:к

Пусть S_1 задана четырьмя вариантами правильного ответа на вопрос о связи переобучения и эмпирического риска. Допустим, имеются три варианта S_2 (см. табл. 1),

связанные отношением схожести с S_1 по *определению 5.1*. Фрагмент тезауруса ПО “Математические методы обучения по прецедентам”, задействованный в доказательстве схожести СЯУ, представлен в табл. 2 ЕЯ-описанием соответствующих фактов.

Таблица 2

Факты предметной области для фрагмента тезауруса

№ п/п	1			2		3		4	
основа	флективная часть + предлог								
заниженн	ость	ость	ости	ости	–	ость	ости	ость	ость
оценк	–	–	–	–	–	и	и	и	и
эмпирическ	ого	–	ого	–	–	–	–	–	–
риск	а	–	а	–	–	–	–	–	–
средн	–	ей	–	ей	–	–	–	–	–
ошибк	–	и:на	–	и:на	–	–	–	и	и
распознавани	–	–	–	–	–	–	–	я	я
обучающ	–	ей	–	ей	–	–	–	–	–
выборк	–	е	–	е	–	–	–	–	–
переусложнени	ем	ем	е	е	–	–	–	–	–
модел	и	и	и	и	–	–	–	–	–
уменьшени	–	–	–	–	е	–	–	–	–
обобщающ	–	–	–	–	ей	ей	ей	–	–
способност	–	–	–	–	и	и	и	–	–
выбор	–	–	–	–	–	–	–	ом	а
решающ	–	–	–	–	его	–	–	его	его
дерев	–	–	–	–	а	–	–	–	–
правил	–	–	–	–	–	–	–	а	а
алгоритм	–	–	–	–	–	а	а	–	–
переподгонк	–	–	–	–	ой	ой	а	–	–
переобучени	–	–	–	–	–	ем	е	–	–
связан	а:с	а:с	–	–	о:с	а:с	–	а:с	–
вызван	а	а	–	–	–	а	–	–	–
обусловлен	а	а	–	–	о	–	–	–	–
привод	–	–	ит:к	ит:к	–	–	ит:к	–	–
завис	–	–	–	–	–	–	–	–	ит:от

Использованные в эксперименте формальные контексты строились по результатам синтаксического разбора фраз, представленных в табл. 1, программой “Cognitive Dwarf”. Как видно из табл. 3, значение схожести будет больше у того варианта S_2, признаки объектов у ФК которого разделяются большим числом объектов формального контекста ситуации S_1 относительно ФК тезауруса.

Таблица 3

Оценка близости ответа правильному варианту

Вариант	$spc(S_1, S_2)$	$\lvert BLCS \rvert$	$\lvert B_1 \setminus BLCS \rvert$	$\lvert B_2 \setminus BLCS \rvert$
1	0,9167	7,7500	0,7500	0,0000
2	0,7917	7,0000	2,0000	0,5000
3	0,8750	7,7500	0,7500	0,7500

Таким образом, в пятой главе предложен метод численной оценки семантической схожести текстов предметно-ограниченного ЕЯ относительно ситуаций его употребления. При этом формальный контекст (5.1) составляет основу выделения классов семантических отношений на базе подхода, изложенного в **разделе 3.5**.

Шестая глава диссертации посвящена разделению и сжатию баз предметных и языковых знаний с применением комплексной методики формирования и кластеризации семантических отношений, изложенной в **разделах 3.5**, **4.1**, **5.2** и **5.3**. Здесь вводится понятие смыслового эталона СЯУ и рассматриваются два приближенных метода его построения с представлением формальным контекстом вида (5.1).

Первый метод основан на подходе к выделению и классификации синтагматических зависимостей, предложенном в **разделе 3.5**.

Пусть $Ke=(Ge,Me,Ie)$ есть искомый формальный контекст эталона. Если $\exists\{j,k\}\subset J:\ (j,k)\in E$ в расширенном дереве (3.11), то для основ b_j и b_k и флексий f_j и f_k соответствующие им элементы множеств Ge и Me, а также элементы отношения Ie, будут сформированы следующим образом.

Случай 1. Индекс k соответствует родительскому узлу, индекс j – дочернему узлу в расширенном дереве (3.11), а линейная структура ЕЯ-фразы не содержит предлог между словами с индексами j и k.

При этом в состав множества признаков Me формального контекста $Ke=(Ge,Me,Ie)$ будут включены признаки $m_1=p_{bs}\bullet b_k$, $m_2=p_{bf}\bullet f_k$, $m_3=p_{fl}\bullet f_j$ и $m_4=f_j\bullet":"\bullet f_k$, основа b_j войдет в множество объектов Ge указанного ФК, а пары (b_j,m_1), (b_j,m_2), (b_j,m_3) и (b_j,m_4) войдут в отношение Ie.

Случай 2. Индекс k соответствует родительскому узлу, индекс j – дочернему узлу в расширенном дереве (3.11), линейная структура ЕЯ-фразы содержит предлог p_y между словами с индексами j и k.

В этом случае признаки m_1 и m_3 формируются аналогично *случаю 1*, $m_2=p_{bf}\bullet f_k\bullet":"\bullet p_y$, $m_4=f_j\bullet":"\bullet f_k\bullet":"\bullet p_y$, пары (b_j,m_1), (b_j,m_2), (b_j,m_3) и (b_j,m_4) включаются в отношение Ie.

Второй метод основан на построении ФК эталона по совокупности ФК вида (5.1) для отдельных СЭ-фраз, задающих СЯУ. При этом формальные контексты указанной совокупности строятся по результатам разбора фраз внешней программой синтаксического анализа. Для отбора объектов и признаков из формальных контекстов фраз вводятся коэффициенты сжатия информации относительно ФК вида (5.1).

Коэффициент сжатия информации по основам равен:

$$ks=\frac{\sum_{i=1}^{nbs}ks_i}{nbs},\qquad(6.4)$$

где $ks_i=\dfrac{\sum_{j=1}^{nbs_i}\sum_{k=1}^{nmf}nas_{ijk}}{nbs_i}$; $nbs=|M_1|$; $nmf=|M_2|$;

$nbs_i=\left|\{g\in Gs:Is(g,m)=true,\ m\in M_1,\ m=p_{bs}\bullet b_i\}\right|$;

$nas_{ijk}=\left|\{m_k\in M_3:Is(g_j,m_k)=true,\ \exists m_{bf}\in M_2,\ m_{bf}=p_{bf}\bullet f_k,\ m_k=b_i\bullet":"\bullet f_k\}\right|$;

p_{bf} соответствует символьной константе *“главное – флексия:”*.

Аналогично определяется коэффициент сжатия информации по флексиям:

$$kf = \frac{\sum_{i=1}^{nfs} kf_i}{nfs}, \qquad (6.5)$$

где $kf_i = \frac{\sum_{j=1}^{nfs_i}\sum_{k=1}^{nmf} naf_{ijk}}{nfs_i}$; $nfs = |M_5|$; $nfs_i = \left|\left\{g \in Gs : Is(g,m) = true,\ m \in M_5,\ m = p_{fl} \bullet f_i\right\}\right|$;

$naf_{ijk} = \left|\left\{m \in M_4 : Is(g_j, m) = true,\ \ \exists\, m_{bf} \in M_2,\ m_{bf} = p_{bf} \bullet f_k,\ m = f_i \bullet ":" \bullet f_k\right\}\right|$.

В **разделе 6.2** представлена пара алгоритмов (*алгоритмы 6.1* и *6.2*), реализующих построение формального контекста эталона. Из них *алгоритм 6.1* выполняет отбор объектов и признаков из формальных контекстов отдельных фраз по максимуму коэффициентов (6.4) и (6.5) результирующего ФК. Признак будет включен в множество признаков ФК эталона, если он входит в пятерку признаков $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \bullet b$, $m_2 = p_{bf} \bullet f_1$, $m_3 = b \bullet ":" \bullet f_1$, $m_4 = p_{fl} \bullet f_2$, $m_5 = f_2 \bullet ":" \bullet f_1$. При этом основе b не должен соответствовать объект ФК, если есть другой объект этого же ФК, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \bullet b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого ФК при том, что признак m относится более чем к одному объекту.

Замечание. Последовательности трех и более соподчиненных слов, встречающиеся более чем в 49% исходных СЭ-фраз, выделяются предварительно на этапе синтаксического разбора. Для каждой такой последовательности строится свой ФК вида (5.1), который будет объединен с ФК эталона. Данный шаг предпринят в целях нежелательного занижения коэффициентов (6.4) и (6.5) при выполнении *алгоритма 6.1*.

Таблица 4

Ситуации языкового употребления

i	Фраза максимальной длины из определяющих СЯУ
1	*Нежелательное переобучение является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*
2	*Тренировочная выборка, на ней проявляется эффект заниженных значений средней ошибки, причиной же является переусложненная модель.*
3	*Контрольная выборка, принятие деревом решения на ней будет с большей вероятностью ошибки именно по причине переподгонки.*
4	*Оценка частоты ошибок на выборке, взятой в качестве контрольной, может для алгоритма оказаться заниженной по причине переподгонки.*
5	*Заниженность оценки ошибки распознавания зависит от выбора правила принятия решений.*
6	*Число закономерностей алгоритмической композиции влияет на частоту ошибок логического классификационного алгоритма на контрольной выборке.*

Качественно процесс формирования смысловых эталонов характеризуется соотношением размеров тезауруса, задаваемого формальным контекстом (5.2), при построении его на основе формальных контекстов вида (5.1) для всех СЭ-фраз каждой СЯУ и на основе эталонов при заданном числе СЯУ в тезаурусе. Пример указанного соотношения приведен на рис. 1 для СЯУ из табл. 4. Часть указанных СЯУ была задействована при построении тезауруса, представленного в табл. 2.

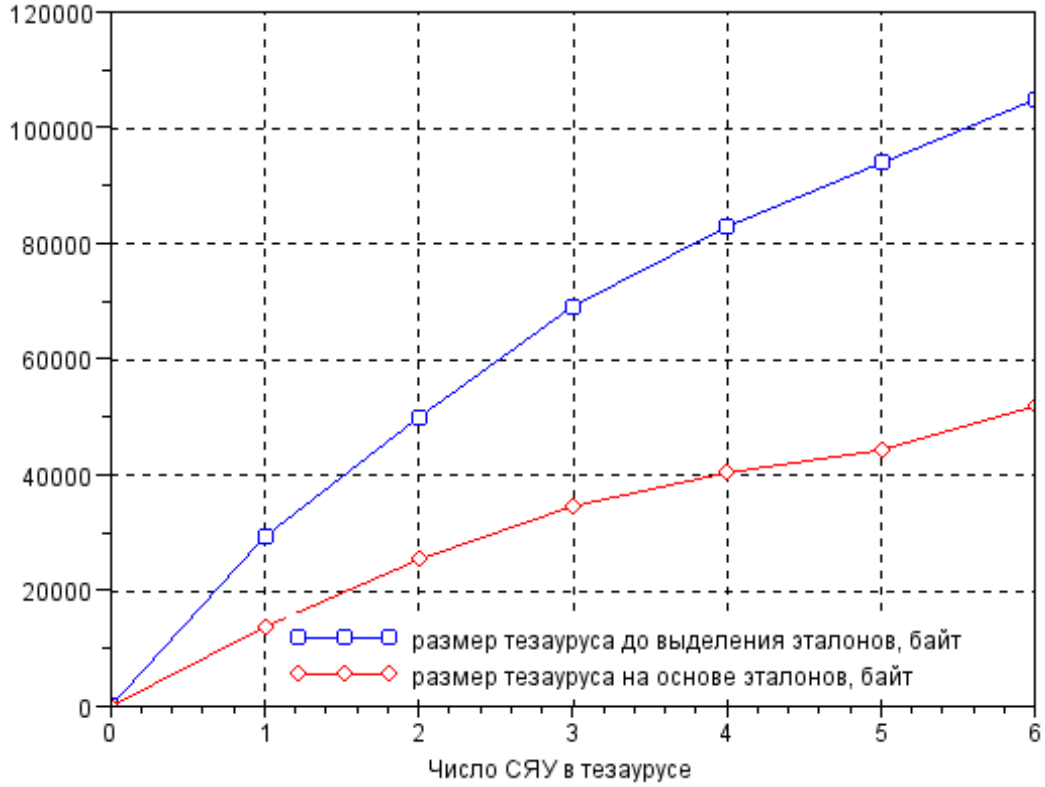

Рис. 1. Размер тезауруса для разного числа СЯУ

Для сравнения в табл. 5 представлены значения числа СЭ-фраз, задающих ситуацию языкового употребления (N_1), фраз, представляющих эталон ситуации языкового употребления (N_2), исходного числа объектов (N_3) и признаков ситуации языкового употребления (N_4), числа объектов (N_5) и признаков эталона (N_6).

Таблица 5

Смысловые эталоны

i	1	2	3	4	5	6
N_1	56	28	29	30	6	10
N_2	8	9	7	9	1	2
N_3	18	17	15	13	12	14
N_4	177	186	173	162	94	81
N_5	9	12	12	11	8	12
N_6	82	90	80	69	35	53

Точность формирования эталона повышается введением согласования знаний относительно разных СЯУ, которое определятся следующим образом. Пусть b_j – основа слова w, f_j – его флексия, выделенные относительно СЯУ S_j. Предположим, что $w = b_1 \bullet f_1$ для СЯУ S_1, $w = b_2 \bullet f_2$ для СЯУ S_2, причём $b_1 = b_2 \bullet suf$, где suf содержит минимум один символ. Тогда относительно S_1 основа b_1 будет заменена на b_2, флексия f_1 – на $f_3 = suf \bullet f_2$, но только в том случае, если частоты встречаемости флексий f_3 и f_2 в отношениях, представляемых формальным контекстом (5.2) тезауруса заданной ПО, не уменьшаются при выполнении указанных замен.

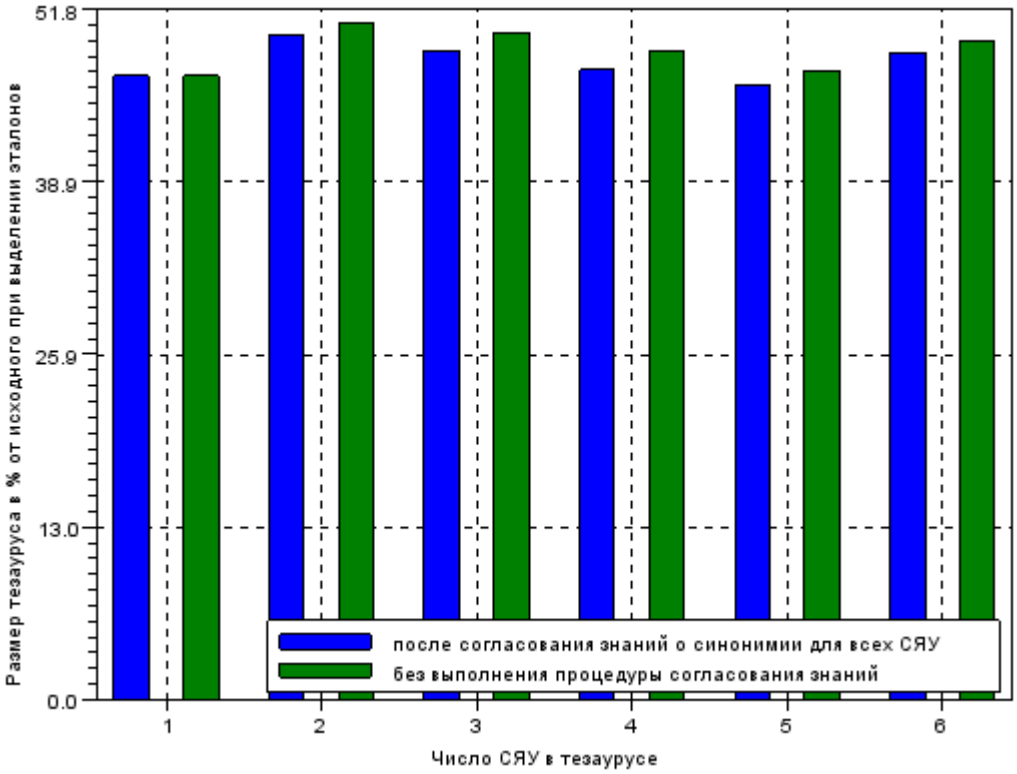

Рис. 2. Сокращение размеров тезауруса согласованием знаний по разным СЯУ

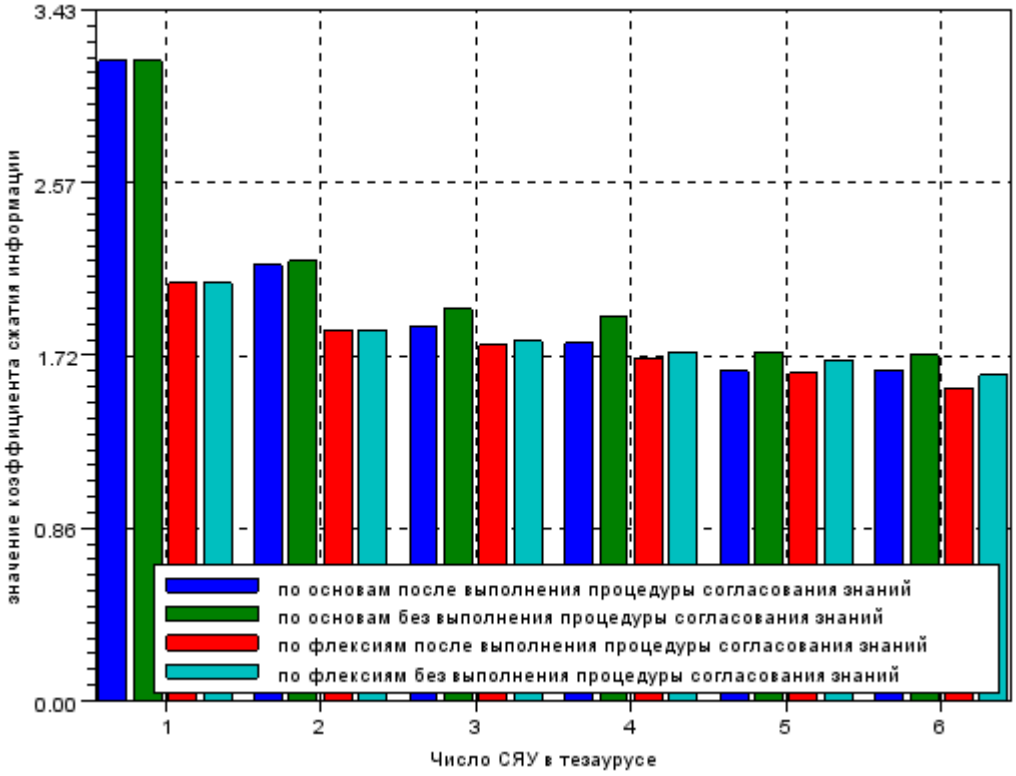

Рис. 3. Сжатие информации тезауруса (эталоны выделены)

Диаграмма на рис. 2 иллюстрирует дополнительное сокращение размеров тезауруса в среднем на 1,5% при выполнении указанной процедуры для ситуаций языкового употребления из табл. 4. Рост специфичности формальных понятий в решётке тезауруса иллюстрируется постепенным уменьшением коэффициентов сжатия информации (рис. 3), аналогичных коэффициентам (6.4) и (6.5) для ФК вида (5.1).

Использование СЯУ в качестве единицы предварительного сжатия информации позволяет сократить резервируемый объём памяти ЭВМ для хранения текстов с учётом возможных видов синонимии. На сегодняшний день за такую оценку для отдельной фразы из n слов берётся значение $vol(n)=n!$. Метод и алгоритмы выделения эталона СЯУ, представленные в диссертации, позволяют оценивать данный объём сверху как $vol_1(n)=l_1\cdot n$ и снизу как $vol_2(n)=l_2\cdot n$, где l_1 – число СЭ-фраз из задающих СЯУ, из которых l_2 определяют эталон. Соотношение указанных оценок для СЯУ из табл. 4 представлено в табл. 6.

Таблица 6

Оценка объёма памяти для хранения ЕЯ-фразы

i	1	2	3	4	5	6
n	12	15	16	17	10	14
$vol(n)$	$4.790\cdot10^8$	$1.308\cdot10^{12}$	$2.092\cdot10^{13}$	$3.557\cdot10^{14}$	$3.629\cdot10^6$	$8.718\cdot10^{10}$
$vol_1(n)$	648	795	416	442	20	42
$vol_2(n)$	168	225	80	187	20	42

В **разделе 6.4** приводится описание архитектуры системы контроля знаний, реализующей предложенные в работе принципы, методы и алгоритмы. На рис. 4 представлен её интерфейс, а также интерпретация ответа на вопрос о влиянии переподгонки на частоту ошибок дерева принятия решений. Демо-версия системы представлена вместе с полным текстом работы в подразделе “Участник:Dmitry.Mikhaylov” раздела “Страницы участников” профессионального информационно-аналитического ресурса www.machinelearning.ru, акты о результатах опытной эксплуатации приводятся в **приложении 2**. Были реализованы следующие компоненты: формирование эталонов и базы лексико-синтаксических связей на основе формальных контекстов (5.1) и (5.2), тезаурус, подготовка и выполнение теста. В целях более гибкой интерпретации ответа испытуемого оценки вида (5.5) вычисляются для случаев неполного ответа, орфографических ошибок, лишних слов, которые не фигурируют в лексико-синтаксических связях, представленных в базе знаний системы.

Рассмотрим более подробно каждый из трёх указанных случаев.

Случай 1. Неполный ответ – для всех слов и словосочетаний из ответа испытуемого нашлись прообразы в наиболее близком варианте правильного ответа, но для части слов правильного ответа не нашлось прообразов в ответе испытуемого.

Ненулевое значение оценки (5.6) будет только для тех из упущенных слов, которые в варианте правильного ответа являются синтаксически зависимыми по отношению к некоторым другим словам, присутствующим в анализируемом ответе. Здесь мы имеем обобщение оценки (5.6) на случай, когда для одного из сравниваемых объектов (основы упущенного слова) не определены признаки из множеств Mx_5 (указание на флексию зависимого слова), Mx_4 (сочетание флексий зависимого и главного слова), M_6 (указание на основу зависимого слова), M_7 (сочетание основы и флексии зависимого слова), M_8 (сочетание основ зависимого и главного слова).

Случай 2. Орфографические ошибки (из допустимых) – слово из ответа испытуемого и слово из варианта правильного ответа являются формами одного и того же

слова, допустимыми в рамках одной лексико-синтаксической связи из известных системе. В этом случае оценка (5.6) для рассматриваемой пары слов вычисляется аналогично общему случаю, описанному в **разделе 5.5**.

Случай 3. “Лишние” слова. Здесь имеется в виду ситуация, когда все слова из варианта правильного ответа нашли свой прообраз в ответе испытуемого, но в анализируемом ответе имеются слова, которые не нашли себе прообразов в правильном “варианте” (в том числе и на уровне словосочетаний). В этом случае ответ не будет засчитан как неверный только тогда, когда “лишние” слова не фигурируют ни в одной лексико-синтаксической связи из представленных в базе знаний системы. При этом значение оценки (5.6) для каждого “лишнего” слова принимается равным нулю.

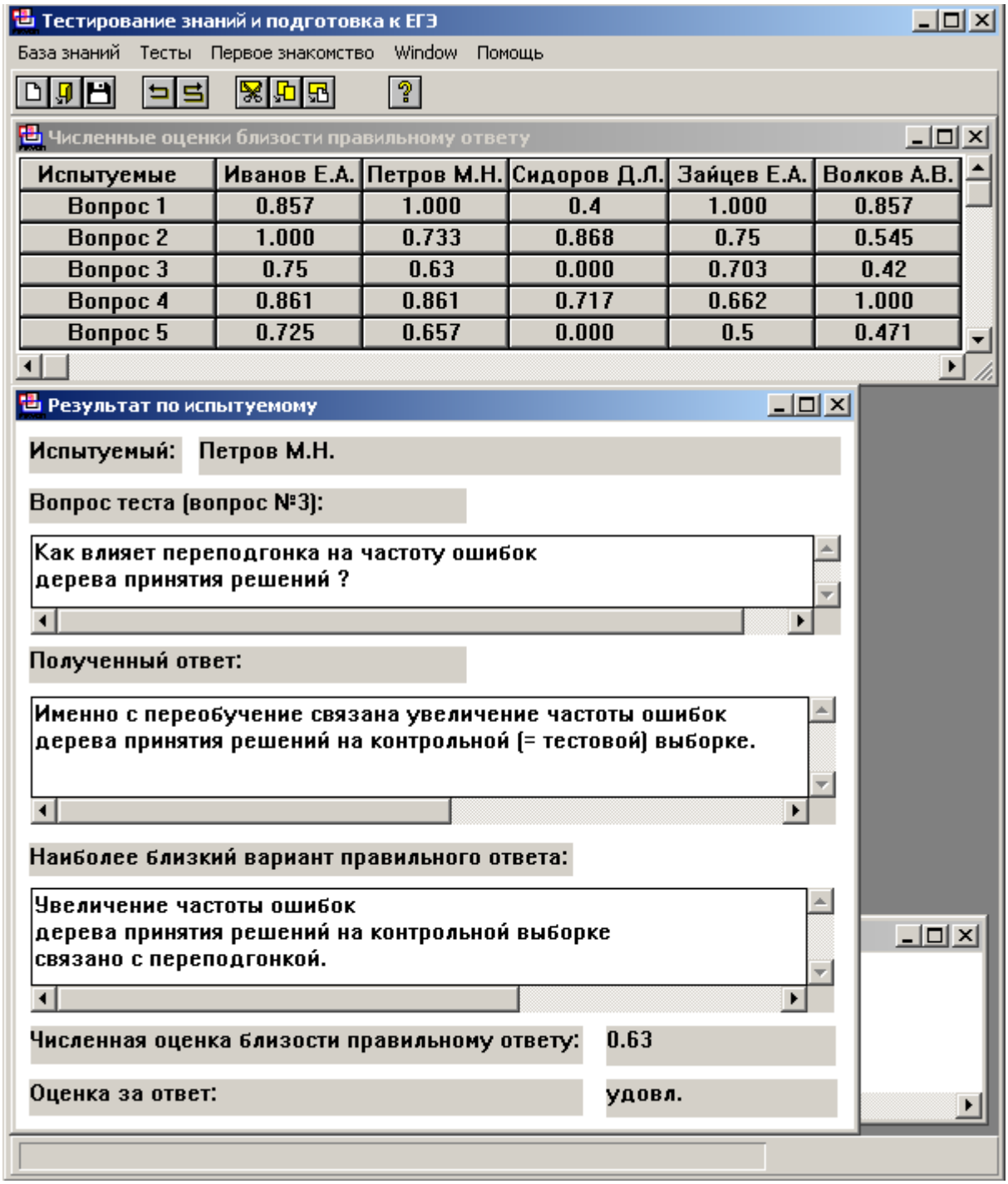

Рис. 4. Пример интерпретации ответа на ТЗОФ

Таким образом, в шестой главе предложен метод компрессии текстовой базы знаний на основе смысловых эталонов и последующего разделения предметных и языковых знаний. При этом наибольший интерес для задач тестирования знаний представляет выделение смыслового эталона на множестве СЭ-фраз на основе принципа формирования и кластеризации семантических отношений, разработанного автором и описанного в **разделах 3.5** и **5.1**.

Заключение

Основные научные результаты работы в области *разработки принципов и методов извлечения данных из текстов на естественном языке* состоят в следующем.

1. На основе теории *анализа формальных понятий* предложена *методика* автоматизированного формирования и экспериментальной оценки знаний, фиксируемых совокупностями классов семантической эквивалентности текстов в рамках ситуаций употребления естественного языка.

Новизной решения является *теоретико-решеточное представление* СЯУ в качестве информационной единицы тезауруса предметной области. За счёт использования формального понятия в качестве базового элемента информационного ресурса предложенное представление тезауруса решеткой формальных понятий позволяет оперировать данными на семантическом уровне без потери или недопустимого упрощения объектов и их признаков.

2. Сформулирован и теоретически обоснован *принцип* формирования и кластеризации семантических отношений на основе описаний ситуаций действительности множествами эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка.

Новизна решения заключается в сравнении символьных последовательностей, составляющих *эквивалентные по смыслу* описания одного и того же *объекта* (ситуации) на заданном языке, с выделением изменяемых и неизменяемых частей для последующего анализа взаимного расположения фрагментов последовательностей в языковых конструкциях с разными логическими акцентами относительно одной и той же ситуации. Предложенная методика выявления закономерностей сосуществования словоформ в линейном ряду позволяет выделять для заданного естественного языка лучший способ выражения нужной мысли, который составляет основу смыслового эталона. Сказанное актуально как для разработки стратегий и правил синтаксического анализа, так и для ролевой идентификации сущностей при формировании признаков сравниваемых текстов. Предложенный принцип формирования и кластеризации семантических отношений реализован в рамках демонстрационного варианта системы контроля знаний.

3. Разработаны *метод и алгоритмы* автоматизированного формирования *смыслового эталона* в виде *решётки формальных понятий*, а также *метод компрессии текстовой базы знаний* на основе выделенных эталонов.

Вне зависимости от пути формирования эталона его выделение сокращает размер базы знаний для оценки семантической схожести текстов предметно-ограниченного естественного языка текстов не менее чем на 40–50%.

В области *разработки и исследования методов и алгоритмов анализа текста* основной научный результат работы есть *метод численной оценки* семантической схожести *текстов* предметно-ограниченного естественного языка относительно ситуаций его употребления.

При этом *семантическая схожесть* текстов *оценивается* по *числу* признаков, которые характеризуют сочетаемость слов и разделяются *объектами* сравниваемых СЯУ относительно тезауруса, что немаловажно, в частности, при интерпретации результатов теста открытой формы в системах контроля знаний.

В области *разработки основ математической теории языков и грамматик* основной научный результат – это решение задачи построения системы целевых выводов в грамматике деревьев (Δ-грамматике).

В отличие от традиционных подходов к формализации преобразований помеченных деревьев, с целью нахождения последовательности преобразований с заданными свойствами автором *исследуется динамика функционирования совокупности правил Δ-грамматики в рамках её динамической информационной модели на основе аппарата ограниченных сетей Петри*. Такое решение учитывает недетерминированный характер порождения множества помеченных деревьев, а построение целевого вывода сводится к классическим задачам теории сетей Петри.

Таким образом, основные научные результаты диссертации можно квалифицировать как решение научной проблемы автоматизации накопления информации о языке как средстве передачи знаний от человека к человеку, имеющей важное значение для обработки данных на ЭВМ в социально-экономических, научных и культурных задачах.

Список основных публикаций автора по теме диссертации

Монография

1. Михайлов Д.В. Теоретические основы построения открытых вопросно-ответных систем. Семантическая эквивалентность текстов и модели их распознавания: монография / Д.В. Михайлов, Г.М. Емельянов; НовГУ им. Ярослава Мудрого. Великий Новгород, 2010. 286 с.

Статьи в рецензируемых научных журналах, включенных в реестр ВАК МОиН РФ

2. Михайлов Д.В. Распознавание сверхфразовых единств при установлении эквивалентности смысловых образов высказываний в общей задаче моделирования языковой деятельности / Г.М. Емельянов, Д.В. Михайлов // Известия СПбГЭТУ “ЛЭТИ”, сер. “Информатика, управление и компьютерные технологии”. СПб., 2003. Вып. 1. С. 65–73.
3. Михайлов Д.В. Информационно-логическая модель системы правил Δ-грамматики / Д.В. Михайлов, Г.М. Емельянов // Известия СПбГЭТУ “ЛЭТИ”, сер. “Информатика, управление и компьютерные технологии”. СПб., 2003. Вып. 3. С. 96–102.
4. Михайлов Д.В. Построение модели объекта информационного пространства применительно к исследованию динамики функционирования Δ-грамматик / Д.В. Михайлов, Г.М. Емельянов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. “Технические науки”. 2004. № 26. С. 131–136.
5. Михайлов Д.В. Представление смысла в задаче установления семантической эквивалентности высказываний / Д.В. Михайлов, Г.М. Емельянов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. “Технические науки”. 2004. № 28. С. 106–110.

6. Михайлов Д.В. Семантическая кластеризация текстов предметных языков (морфология и синтаксис) / Д.В. Михайлов, Г.М. Емельянов // Компьютерная оптика. 2009. Т. 33, № 4. С. 473–480.
7. Михайлов Д.В. Формирование смысловых эталонов и интерпретация результатов открытых тестов в системах контроля знаний / Д.В. Михайлов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. “Технические науки”. 2011. № 65. С. 83–87.
8. Михайлов Д.В. Смысловые эталоны в моделях распознавания и компрессии текстов / Д.В. Михайлов // Вестник Новгородского государственного университета имени Ярослава Мудрого. 2012. № 68 (в печати).
9. Mikhailov D. V. Synonymic Transformations in Analysis of Semantic Pattern Equivalence at the Superphrase Unity Level / G. M. Emelyanov, D. V. Mikhailov, E. I. Zaitseva // Pattern Recognition and Image Analysis. 2003. Vol. 13, N 1. P. 21–23.
10. Mikhailov D. V. Recognition of Superphrase Unities in Texts while Establishing Their Semantic Equivalence / G. M. Emelyanov, D. V. Mikhailov, E. I. Zaitseva // Pattern Recognition and Image Analysis. 2003. Vol. 13, N 3. P. 447–451.
11. Mikhailov D. V. Updating the Language Knowledge Base in the Problem of Equivalence Analysis of Semantic Images of Statements / G. M. Emelyanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2005. Vol. 15, N 2. P. 384–386.
12. Mikhailov D. V. Filling in the Government-Pattern Dictionary in the Analysis of Equivalence for Sense Images of Statements / G. M. Emel’yanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2007. Vol. 17, N 2. P. 268–273.
13. Mikhailov D. V. Clusterization of Semantic Meanings in the Problem of Sense Equivalence Situation Recognition / G. M. Emel’yanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2009. Vol. 19, N 1. P. 92–102.
14. Mikhailov D. V. Formation and clustering of noun contexts within the framework of Splintered Values / D. V. Mikhailov, G. M. Emelyanov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2009. Vol. 19, N 4. P. 664–672.
15. Mikhailov D. V. Sense’s Standards and Machine Understanding of Texts in the System for Computer-Aided Testing of Knowledge / G. M. Emelyanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2011. Vol. 21, N 4. P. 705–719.
16. Mikhailov D. V. Semantic Clustering and Affinity Measure of Subject-Oriented Language Texts / D.V. Mikhailov, G.M. Emel’yanov // Pattern Recognition and Image Analysis. 2010. Vol. 20, N 3. P. 376–385.
17. Корнышов А.Н. Концептуально-ситуационное моделирование высказываний естественного языка в задаче анализа их смысловой эквивалентности / А. Н. Корнышов, Д.В. Михайлов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. “Технические науки”. 2005. № 34. С. 76–80.
18. Emelyanov G.M. Development of Recognition System of Analysis of Semantic Images of Natural Language Statements / G.M. Emelyanov, E.I. Zaitseva, D.V. Mikhailov, E.P. Kurashova // Pattern Recognition and Image Analysis. 2003. Vol. 13, N 2. P. 251–253.
19. Emelyanov G. M. Semantic Relation Analysis for Classification of the Meaning Patterns of Utterances / G. M. Emelyanov, D. V. Mikhailov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2005. Vol. 15, N 2. P. 382–383.
20. Emel’yanov G. M. Analysis of Semantic Relations in Classification of Sense Images of Statements / G. M. Emel’yanov, D. V. Mikhailov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2007. Vol. 17, N 2. P. 274–278.

Доклады на международных конференциях

21. Михайлов Д. В. Применение аппарата ограниченных сетей Петри для построения динамической модели естественного языка / Г. М. Емельянов, Е. И. Зайцева, Д. В. Михайлов // Интеллектуализация обработки информации: тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины, Таврический национальный университет. Симферополь, 2002. С. 121–122.
22. Михайлов Д. В. Установление смысловой эквивалентности высказываний: на пути к решению проблемы / Г. М. Емельянов, Д. В. Михайлов // Интеллектуализация обработки информации: тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины. Симферополь, 2004. С. 70.
23. Михайлов Д. В. Модель сортовой системы языка в задаче построения семантического образа высказывания на уровне глубинного синтаксиса / Д. В. Михайлов, Г. М. Емельянов // Интеллектуализация обработки информации: тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины. Симферополь, 2006. С. 148–150.
24. Михайлов Д. В. Формирование и кластеризация понятий на основе множества ситуационных контекстов / Д. В. Михайлов, Г. М. Емельянов, Н. А. Степанова // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. / Крымский научный центр НАН Украины. Симферополь, 2008. С. 168–170.
25. Михайлов Д. В. Семантическая схожесть текстов в задаче автоматизированного контроля знаний / Д. В. Михайлов, Г. М. Емельянов // 8-я Международная конференция “Интеллектуализация обработки информации” (ИОИ-2010): Сборник докладов. М., 2010. С. 516–519.
26. Mikhailov D. V. Updating of the language knowledge base in the problem of statement’s semantic images’s equivalence’s analysis / G. M. Emelyanov, D. V. Mikhailov // 7th Int. Conf. on Pattern Recognition and Image Analysis: new Information Technologies (PRIA-7-2004). Conf. Proc. / SPbETU. St. Petersburg, 2004. Vol. II. P. 462–465.
27. Mikhailov D. V. Formalization of the word’s lexical meaning in a problem of recognition of natural language’s statements’s synonymy’s situations / G. M. Emelyanov, D. V. Mikhailov // 8th Int. Conf. “Pattern Recognition and Image Analysis: new Information Technologies” (PRIA-8-2007). Conf. Proc. / Mari State Technical University. Yoshkar-Ola, 2007. Vol. 2. P. 253–257.
28. Mikhailov D. V. Formation and clustering of Russian’s nouns’s contexts within the frameworks of splintered values / D. V. Mikhailov, G. M. Emelyanov // 9th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-9-2008). Conf. Proc. / N.I. Lobachevsky State University of Nizhni Novgorod. Nizhni Novgorod, 2008. Vol. 2. P. 39–42.
29. Mikhailov D. V. Semantic clustering in a problem of text information’s compression / D. V. Mikhailov, G. M. Emelyanov // 10th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-10-2010). Conf. Proc. St. Petersburg, 2010. Vol. 2. P. 193–196.
30. Емельянов Г. М. Синонимические преобразования в задаче анализа эквивалентности смысловых образов высказываний на уровне сверхфразовых единств / Г. М. Емельянов, Д. В. Михайлов, Е. И. Зайцева // Распознавание образов и анализ изображений: новые информационные технологии (РОАИ-6-2002): труды 6-й Междунар. конф. / НовГУ им. Ярослава Мудрого. Великий Новгород, 2002. Т. 1. С. 215–219.
31. Емельянов Г. М. Концептуально-ситуационное моделирование процесса перифразирования высказываний Естественного Языка как обучение на основе прецедентов /

Г. М. Емельянов, А. Н. Корнышов, Д. В. Михайлов // Интеллектуализация обработки информации: тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины. Симферополь, 2006. С. 78–79.
32. Корнышов А. Н. Иерархизация системы предикатов семантических отношений / А. Н. Корнышов, Д. В. Михайлов // Интеллектуализация обработки информации: тезисы докладов Международной научной конференции / Крымский научный центр НАН Украины. Симферополь, 2008. С. 130–131.
33. Emelyanov G.M. Semantic relation analysis for classification of meaning pattern of utterances / G.M. Emelyanov, D.V. Mikhailov, N.A. Stepanova // 7th Int. Conf. on Pattern Recognition and Image Analysis: new Information Technologies (PRIA-7-2004). Conf. Proc. / SPbETU. St. Petersburg, 2004. Vol. II. P. 460–461.

Доклады на всероссийских конференциях

34. Михайлов Д. В. Вопросы моделирования семантической связанности для систем автоматизированного тестирования знаний / Г. М. Емельянов, Д. В. Михайлов // Доклады X Всероссийской конференции “Математические методы распознавания образов” (ММРО-10). М., 2001. С. 53–56.
35. Михайлов Д. В. Применение семантических полей словаря РОСС в задаче построения модели управления предикатного слова / Д. В. Михайлов, Г. М. Емельянов // 12-я Всероссийская конференция “Математические методы распознавания образов” (ММРО-12): сборник докладов. М., 2005. С. 382–385.
36. Михайлов Д. В. Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности / Д. В. Михайлов, Г. М. Емельянов // 13-я Всероссийская конференция “Математические методы распознавания образов” (ММРО-13). М., 2007. С. 500–503.
37. Михайлов Д. В. Морфология и синтаксис в задаче семантической кластеризации / Д. В. Михайлов, Г. М. Емельянов // 14-я Всероссийская конференция "Математические методы распознавания образов" (ММРО-14): сборник докладов. М., 2009. С. 563–566.
38. Михайлов Д. В. Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний / Г. М. Емельянов, Д. В. Михайлов // 15-я Всерос. конф. "Математические методы распознавания образов" (ММРО-15): сб. докл. М., 2011. С. 581–584.

Свидетельство об официальной регистрации программы для ЭВМ

39. Свидетельство об официальной регистрации программы для ЭВМ № 2010617263. Программа формирования синтаксических отношений на множестве семантически эквивалентных фраз / Залешин М. В., Михайлов Д. В., Емельянов Г. М.; заявитель и правообладатель “Новгородский государственный университет имени Ярослава Мудрого”. Заявка № 2010615398; заявл. 02.09.10.; зарег. 29.10.10.

Наиболее значимые публикации в других изданиях

40. Михайлов Д.В. Построение динамической модели естественного языка применительно к разработке языковой базы знаний / Г.М. Емельянов, Е.И. Зайцева, Д.В. Михайлов // Искусственный интеллект. 2002. № 2. С. 443–446.

41. Михайлов Д. В. Установление смысловой эквивалентности высказываний: на пути к решению проблемы / Г. М. Емельянов, Д. В. Михайлов // Искусственный интеллект. 2004. № 2. С. 86–90.
42. Михайлов Д. В. Построение модели управления предикатного слова на основе его лексикографического толкования / Г. М. Емельянов, Д. В. Михайлов // Таврический вестник информатики и математики. 2005. № 1. С. 35–48.
43. Михайлов Д. В. Модель сортовой системы языка в задаче построения семантического образа высказывания на уровне глубинного синтаксиса / Д. В. Михайлов, Г. М. Емельянов // Таврический вестник информатики и математики. 2006. № 1. С. 79–90.
44. Михайлов Д. В. Формирование и кластеризация понятий на основе множества ситуационных контекстов / Д. В. Михайлов, Г. М. Емельянов, Н. А. Степанова // Таврический вестник информатики и математики. 2008. № 2. С. 79–88.
45. Михайлов Д. В. Формирование и кластеризация контекстов для существительных русского языка в рамках конверсивных замен / Д. В. Михайлов, Н. А. Степанова, И. И. Юрченко // Физика и механика материалов: приложение к научно-теоретическому и прикладному журналу “Вестник Новгородского государственного университета имени Ярослава Мудрого”. 2009. № 50. С. 31–34.
46. Емельянов Г. М. Концептуально-ситуационное моделирование процесса перефразирования высказываний естественного языка как обучение на основе прецедентов / Г. М. Емельянов, А. Н. Корнышов, Д. В. Михайлов // Искусственный интеллект. 2006. № 2. С. 72–75.