На правах рукописи

ЛУКАШЕВИЧ НАТАЛЬЯ ВАЛЕНТИНОВНА

Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа

05.25.05 – Информационные системы и процессы

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора технических наук

Москва 2014

Работа выполнена в лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного Университета им. М.В. Ломоносова

Официальные оппоненты:

заведующий сектором Вычислительного центра РАН им. А.А. Дородницына, доктор технических наук, профессор Хорошевский Владимир Федорович

заведующий кафедрой Высшей школы менеджмента СПбГУ, доктор технических наук, профессор Гаврилова Татьяна Альбертовна

заведующий. лабораторией Института проблем управления РАН, доктор технических наук, профессор Кузнецов Олег Петрович

Ведущая организация: Институт системного анализа РАН

Защита диссертации состоится в ..... часов на заседании диссертационного совета Д 002.026.01 при Всероссийском институте научной и технической информации РАН (ВИНИТИ РАН) по адресу: 125190, Москва, ул. Усиевича, д. 20, комн. 502.
Тел. совета (499) 155-46-21.

С диссертацией можно ознакомиться в библиотеке ВИНИТИ РАН.

Автореферат разослан "____"_____2014 г.

Ученый секретарь диссертационного совета

Общая характеристика работы

Актуальность темы. В настоящее время в связи с огромными объемами электронных документов имеется все возрастающая потребность в обработке неструктурированной текстовой информации, повышению качества и эффективности имеющихся методов обработки текстов. В число активно развивающихся направлений обработки неструктурированной текстовой информации входят такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов и многое другое.

Современные информационно-поисковые и информационно-аналитические системы работают с текстовой информацией в широких или неограниченных предметных областях, поэтому характерной чертой современных методов обработки текстовой информации стало минимальное использование знаний о мире и о языке, опора на статистические методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п.

Недостаточное использование лингвистических и онтологических знаний (знаний о мире), используемых в приложениях информационного поиска и автоматической обработки текстов, приводит к разнообразным проблемам: нерелевантному поиску, некачественно рубрикации и реферированию документов. Эти проблемы усугубляются в специализированных видах информационного поиска такие, как медицинский, патентный, научный поиск.

В то же время внедрение дополнительных объемов знаний о языке и мире в современные методы автоматической обработки текстов является сложной задачей. Это связано с тем, что такие знания необходимо описывать в специально создаваемых компьютерных ресурсах (тезаурусах, онтологиях), которые должны содержать описания десятков тысяч слов и словосочетаний. При применении таких ресурсов обычно необходимо автоматически разрешать многозначность слов, т. е. выбирать правильное значение. Кроме того, поскольку ведение любых ресурсов отстает от развития предметной области, необходимо развитие комбинированных методов, учитывающих как знания, так и лучшие современные статистические методы обработки текстов.

Изначально в качестве ресурсов для информационного поиска получили большое распространение информационно-поисковые тезаурусы. Но они создавались для ручного индексирования документов людьми-индексаторами, и в последние десятилетия их роль резко снизилась. Затем множество экспериментов в области автоматической обработки текстов и информационного поиска проводилось на основе тезауруса WordNet[1]. Однако этот тезаурус создавался в качестве проверки психолингвистической теории, и не учитывает особенностей автоматической обработки текстов, из-за чего имеется много проблем в его использовании в прикладных разработках. Кроме того, многими исследователями была показана недостаточная формализация описаний в вышеуказанных типах тезаурусов, что приводит к серьезным проблемам с автоматическим логическим выводом, необходимым во многих приложениях автоматической обработки текстов и информационного поиска (расширение поискового запроса, вывод рубрики, разрешение многозначности и др.). Проблемы с логическим выводом усиливаются при обработке целых текстов (в противоположность обработке отдельного предложения), которые могут содержать сотни и тысячи слов и имеют сложную внутреннюю структуру.

[1] Miller G. Nouns in WordNet // WordNet – An Electronic Lexical Database / Fellbaum, C (ed). The MIT Press, 1998. P. 23-47.

Одной из современных парадигм компьютерных ресурсов, описывающих знания о мире и предметных областях, являются так называемые формальные онтологии. При этом многие исследователи в этой сфере видят своей целью разработку достаточно сложных формальных подходов в описании, практически аксиоматизированных теорий. Однако автоматическую обработку неструктурированных текстов на естественном языке с их неоднозначностью и неточностью трудно проводить с помощью аксиоматизированных теорий. Кроме того, описания в рамках таких формализмов плохо масштабируются для представления знаний в широких неструктурированных предметных областях.

Вышеуказанные вопросы применения онтологий к автоматической обработке текстов исследовались в трудах многих российских и зарубежных исследователей: П. Воссена, Ю.А. Загорулько, Н.Г. Загоруйко, Д. Г. Лахути, Б. Магнини, А.С. Нариньяни, О.А. Невзоровой, С. Ниренбурга, В. Раскина, В. Ш. Рубашкина, В.Д. Соловьева, С.Ю. Соловьева, М.Г. Мальковского, Х. Феллбаум, Г. Хирста, Э. Хови, В.Ф. Хорошевского и др. В работах Е.М. Бениаминова, Т.А. Гавриловой, Л.А. Калиниченко, А.С. Клещева и др. исследовались вопросы применения онтологий в различных компьютерных приложениях. В работах Р.С. Гиляревского, Г.Г. Белоногова, Д.Г. Лахути, А.И. Черного и многих других обсуждались вопросы улучшения качества информационного поиска на основе дополнительных знаний.

Таким образом, рост потоков неструктурированной информации, необходимость повышения качества ее обработки и представления в информационных системах требует развития моделей представления онтологических и лингвистических знаний в компьютерном ресурсе, предназначенном для эффективного использования в автоматической обработке текстов в широких предметных областях.

Целями исследования, проведенного в диссертации, являются

1) разработка формализованной модели лингвистико-онтологического ресурса (лингвистической онтологии) для описания широких предметных областей, обеспечивающей эффективность широкого круга приложений информационного поиска и автоматической обработки текстов и позволяющей создавать большие ресурсы;
2) разработка алгоритмов для автоматического построения тематического представления содержания текста как иерархической структуры, моделирующей структуру связного текста;
3) разработка методов решения различных задач автоматической обработки текстов в широких предметных областях на основе созданных лингвистических ресурсов и тематического представления текстов;
4) разработка алгоритмов автоматизированного пополнения лингвистической онтологии.

Научная новизна работы. В диссертации разрабатывается система моделей и алгоритмов, направленных на комплексное решение задачи применения знаний о языке и о мире для улучшения качества автоматической обработки текстов в приложениях информационного поиска.

Предложена новая формализованная модель базы знаний онтологического типа – лингвистической онтологии, предназначенной для использования в автоматической обработке текстов в широких предметных областях. Модель основывается на сочетании принципов трех различных методологий разработки компьютерных ресурсов:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet;
- методологии создания формальных онтологий.

Предложенная модель позволяет в короткие сроки создавать онтологические ресурсы в неструктурированных предметных областях. Особенностью предлагаемого подхода к

описанию предметной области является то, что создаваемые предметно-ориентированные базы знаний направлены на эффективное применение в различных задачах информационного поиска, что показано в целом ряде вычислительных экспериментов.

Предложена модель представления тематической структуры текстов на основе согласованного учета свойств лексической и глобальной связности текста. Предложен и реализован алгоритм автоматического построения тематического представления содержания текстов, которое моделирует основное содержание текста посредством выделения тематических узлов – совокупностей близких по смыслу понятий текста.

Предложен метод концептуального индексирования документов для информационно-поисковой системы, базирующийся на знаниях, описанных в предметно-ориентированной базе знаний, и построенном тематическом представлении документов.

Предложен и реализован алгоритм автоматического разрешения лексической многозначности на основе знаний, сочетающий информацию о локальном и глобальном контексте употребления многозначного слова. Метод разрешения многозначности базируется на совокупности различных контекстных признаков и для нахождения их оптимальной комбинации был использован численный метод координатного спуска.

Предложен и реализован алгоритм автоматической рубрикации документов, основанный на использовании тематического представления документов и описании рубрик в виде булевских выражений над понятиями лингвистической онтологии, и способный обрабатывать тексты различных типов (официальные документы, сообщения информационных агентств, газетные статьи). Система рубрикации легко настраивается на новый рубрикатор и новые типы текстов, рубрицирование можно осуществлять сразу по нескольким рубрикаторам. На основе предложенного метода было реализовано более 20 систем автоматической рубрикации текстов с количеством тематических рубрик от 35 до 3000. Возможности быстрой настройки системы рубрикации на новый рубрикатор и достигаемый при этом высокий уровень качества рубрикации был продемонстрирован на Российском семинаре по информационному поиску РОМИП в 2007 и 2010 гг[1].

Предложен и реализован алгоритм автоматического многошагового построения булевского выражения по длинному поисковому запросу на естественном языке, включающий расширение запроса по тезаурусным отношениям, подтвержденным поисковой выдачей. Для обеспечения устойчивости обработки длинного поискового запроса метод построения булевских выражений используется в сочетании с совокупностью различных признаков запроса, документа и коллекции, и для нахождения оптимальной функции соответствия между запросом и документом был использован численный метод координатного спуска.

Предложен и реализован метод автоматического аннотирования отдельного документа, который базируется на тематическом представлении содержания текстов, что позволяет повысить связность создаваемой аннотации. Реализованная система автоматического аннотирования одного документа получила наилучший результат в одной из номинаций на конференции SUMMAC в 1998 г.[2]Предложен и реализован метод автоматического аннотирования новостного кластера на основе тематического представления кластера и моделировании лексической связности. Показано, что предложенная модель позволяет значительно улучшить связность порождаемой аннотации, а

---

[1] Агеев М., Добров Б., Красильников П., Лукашевич Н., Павлов А., Сидоров А., Штернов С. УИС РОССИЯ в РОМИП2007: поиск и классификация // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008.

[2] Mani I., House D., Klein G., Hirshman L., Firmin Th., Sundheim B. SUMMAC: a text summarization evaluation // Natural Language Engineering. 2002. V.8, N 01. P. 43-68.

также снизить повторы информации, ухудшающие восприятие порожденного текста человеком.

Предложена и обоснована многофакторная модель извлечения терминов предметной области из текстов. Реализован новый метод автоматизированного извлечения терминов предметной области для пополнения предметно-ориентированной базы знаний. Метод основывается на вычислении для языковых выражений трех типов статистических характеристик

- характеристик, вычисленных на основе текстовой коллекции предметной области,
- характеристик, вычисленных на основе поисковой выдачи глобальных поисковых систем,
- характеристик, вычисляемых на основе известных терминов предметной области,

что очень важно для пополнения предметно-ориентированной базы знаний, учета появляющихся новых терминов в развивающейся предметной области. Для нахождения оптимальной комбинации статистических характеристик для определения терминологичности выражения применяется метод машинного обучения – логистическая регрессия.

Достоверность результатов обуславливается использованием для их получения фундаментальных принципов представления знаний, теории связного текста, методов формального логики и методов оптимизации, проведением большого числа вычислительных экспериментов по оценке качества работы предложенных методов, в том числе и на общественно доступных коллекциях с тестированием независимыми экспертами.

Практическая значимость. Разработанная модель лингвистической онтологии стала основой для разработки нескольких лингвистических и терминологических ресурсов в ряде предметных областей, в том числе такие ресурсы, как Тезаурус русского языка РуТез и Онтология по естественным наукам и технологиям ОЕНТ.

Для применения разработанных лингвистических ресурсов в автоматической обработке текста был предложен и реализован ряд алгоритмов, которые были объединены в программно-лингвистический комплекс АЛОТ. Созданные лингвистические ресурсы и методы обработки текстов используются для обработки потоков документов в Университетской информационной системе РОССИЯ (uisrussia.msu.ru). Созданные технологии и ресурсы применяются в различных проектах с государственными и коммерческими организациями, включая Государственную Думу Федерального Собрания Российской Федерации, НИИ Восход, ФСБ РФ, Банк России, Счетную палату Российской Федерации, а также в коммерческих организациях: НПО «Гарант-сервис», компания «Рамблер-Медиа».

Апробация. Основные результаты диссертации докладывались на Международной конференции по интеллектуальным технологиям и компьютерной лингвистике Диалог (1996-2012 гг.), Национальной конференции по искусственному интеллекту (1996, 2000, 2002, 2004, 2006, 2010), Российской конференции по электронным библиотекам RCDL (2001-2007, 2010), Всероссийской конференции Знания-Онтологии-Теории (2007, 2009), Международной конференции по Лингвистическим ресурсам и их оценке LREC (2002, 2004, 2006, 2012), симпозиумах Американской ассоциации по искусственному интеллекту (1998, 2002), Международной конференции Текст-Речь-Диалог (TSD, Brno, 1998), Международном семинаре по взаимоотношениям между онтологиями и лексикой OntoLex (2000, 2004), Международной конференции «Знания - Диалог – Решение» (1995, 2001, 2003), семинарах Международной конференции по информационному поиску SIGIR (2002), Международной

конференции по многоязычному информационному поиску CLEF (2003, 2005), Международной конференции по использованию естественного языка в базах данных NLDB (Клагенфурт, Австрия, 1999), Международном конгрессе "Терминология и инженерия знаний" TKE (Инсбрук, Австрия, 1999), Международной конференции по применению статистических методов для обработки текстов JADT (Лозанна, Швейцария, 2000), Международном конгрессе «Русский язык: исторические судьбы и современность» (2001, 2004, 2007), Казанской школе по компьютерной и когнитивной лингвистике TEL (2001-2004, 2006-2009), Международной конференции Всемирной ассоциации ворднетов GWA (Брно, 2004), семинарах по итогам конкурса компании Яндекс «Интернет-математика» (2005, 2007), Симпозиуме «Онтологическое моделирование» (2008, 2010), Международной конференции по концептуальным структурам ICCS (Москва, 2009), Международной конференции по распознаванию образов и машинному обучению PReMI (Москва, 2011), Международной конференции "Новые достижения в автоматической обработке текстов" RANLP (2011, Болгария), Международной конференции по компьютерной лингвистике Coling-2012, семинарах российской секции SIGMOD (2005, 2008), семинаре «Когнитивные аспекты компьютерной лексикографии» (НИВЦ МГУ, 2005, 2008), междисциплинарном семинаре "Лингвистические основы информационных технологий» (ИПИРАН), научном семинаре отдела Интеллектуальных систем ВЦ РАН, научном семинаре Российской ассоциации искусственного интеллекта (2013).

Публикации. Диссертация написана по материалам более ста пятидесяти работ автора; пятьдесят девять основных из них указаны в списке литературы. Опубликована монография, поддержанная грантом РФФИ, пятнадцать из работ опубликованы в журналах из перечня ведущих периодических изданий ВАК, также тринадцать работ указаны в международных системах цитирования из списка ВАК.

Личный вклад автора. Все описанные в диссертации модели и алгоритмы разработаны лично автором; первые версии программ, реализующих предложенные алгоритмы автоматического построения тематического представления текстов, автоматической рубрикации текстов, автоматического разрешения лексической многозначности, автоматического аннотирования, автоматического извлечения терминов из текстов, написаны автором диссертации лично; текущие версии программных модулей, реализующие предложенные в диссертации алгоритмы в рамках различных программно-аппаратных архитектур, написаны под непосредственным контролем автора диссертации.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав и списка литературы. Объем диссертации составляет 312 страниц. Список литературы включает 317 наименований.

Содержание диссертации

Во введении обоснована актуальность темы диссертации и сформулирована цель работы.
В первой главе рассматриваются особенности различных видов онтологических ресурсов и методы их применения в автоматической обработке текстов.

Онтологии представляют собой компьютерные ресурсы, содержащие формализованное описание фрагмента знаний о мире. При имеющихся различиях к определению онтологии большинство авторов соглашаются в наборе основных компонентов онтологии: классы или понятия; атрибуты (свойства); экземпляры (отдельные индивиды),

отношения между классами или экземплярами; аксиомы. Таким образом, формальным определением онтологий может служить следующее:

$$O = <C, E, At, R, A>$$

где C – понятия (классы) онтологии, E – экземпляры онтологии, At – атрибуты понятий и экземпляров онтологии, R – отношения между понятиями (экземплярами), A – аксиомы онтологии.

Термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий разными авторами рассматривается множество различных компьютерных ресурсов, в том числе и известных задолго до начала исследований по онтологиям таких, как рубрикаторы или тезаурусы.

При этом в некоторых типах онтологий некоторые из вышеперечисленных компонентов онтологий могут быть не определены[1]. Так, рубрикаторы обычно не включают экземпляры и атрибуты, т.е. распространенной формальной моделью рубрикаторов является модель вида:

$$O = <C, \varnothing, \varnothing, R, A>=< C, R, A >$$

Наиболее формализованные онтологии представляют собой **логические теории**, построенные на произвольных логических утверждениях о понятиях – аксиомах. Для описания таких формальных онтологий применяются различные логики (дескриптивные логики, модальные логики, логика предикатов первого порядка) и различные языки описания онтологий DAML+OIL, OWL, CycL, Ontolingua. Онтологии, такие, как тезаурусы, рубрикаторы, понятия которых не определяются полностью в терминах формальных свойств и аксиом, иногда называются **легкими онтологиями** (lightweight ontologies).

Разработчики онтологий по-разному трактуют взаимоотношения между онтологией и естественным языком. Некоторые исследователи рассматривают онтологию как структуру, независимую от естественного языка, другие – как структуру, независимую от *конкретного* естественного языка, третьи вводят элементы языкового лексикона в формальное определение онтологии. Известной формальной моделью онтологии с включением единиц естественного языка является следующая модель[2]:

$$O = <L, C, F, G, H, R, A>$$

где $L = L_C \cup L_R$ – словарь онтологии, содержащий набор лексических единиц (знаков) для понятий L_C и набор знаков для отношений L_R ; C – набор понятий онтологии; F и G связывают наборы лексических единиц $\{L_j\} \subset L$ с наборами понятий и отношений данной онтологии; H – фиксирует таксономический характер отношений (связей), при котором понятия онтологии связаны нерефлексивными, ациклическими, транзитивными отношениями $H \subset C$ x C; R – обозначает нетаксономические отношения между понятиями онтологии, A – набор аксиом онтологии.

Известно, что в установлении взаимоотношений между понятиями и словами (выражениями) естественного языка имеется много проблем. Стремление к четкой формализации отношений между понятиями в онтологии чрезвычайно трудно соблюсти в ситуации, когда необходимо создавать сверхбольшие ресурсы. Поэтому значительно большее распространение в приложениях автоматической обработки текстов получили

[1] Гаврилова Т.А., Хорошевский В.Ф., Базы знаний интеллектуальных систем. Санкт-Петербург: Изд-во "Питер", 2000. 382 с.

[2] Maedche A., Staab S. Learning Ontologies for the Semantic Web // In Proceedings of Semantic Web Workshop, Hongkong. 2001.

вышеупомянутые "легкие" онтологии, например, тезаурусы. Тезаурусы представляют собой так называемые **лингвистические онтологии**, т.е. онтологии, опирающиеся в своем построении на значения реально существующих выражений естественного языка. Наиболее известными типами тезаурусов, обсуждаемыми в качестве источников знаний для приложений обработки неструктурированной информации, являются информационно-поисковые тезаурусы и тезаурусы типа WordNet.

Информационно-поисковый тезаурус (в соответствии с определениями стандартов[1]) – это нормативный словарь терминов на естественном языке, явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов. В соответствии с международными и национальными стандартами формальную модель информационно-поискового тезауруса можно представить следующим образом:

$$ИПТ = <D_{th}, T, R_H, R_A, A_T>$$

где D_{th} – набор дескрипторов предметной области, соответствующий понятиям данной предметной области, индекс $_{th}$ означает в данном случае тот факт, что разработчики информационно-поисковых тезаурусов включают в состав дескрипторов термины предметной области, которые необходимы для выражения основных тем документов этой ПО; T – набор терминов предметной области: $D_{th} \subset T$; R_H – иерархические отношения информационно-поискового тезауруса; R_A – ассоциативные отношения информационно-поискового тезауруса; A_T – аксиомы транзитивности иерархических отношений.

Информационно-поисковые тезаурусы, создаваемые в том виде, как это закреплено международными и национальными стандартами, предназначены для использования их в ручном режиме индексирования. По своей сути такой тезаурус является искусственным языком описания, построенным на основе естественного языка; имеется значительная дистанция между лексическим составом документов предметной области и словарным составом информационно-поискового тезауруса в этой предметной области. Поэтому традиционные информационно-поисковые тезаурусы, разработанные для ручного индексирования, сложно использовать при автоматическом индексировании документов, применять в других приложениях информационного поиска, хотя такие тезаурусы содержат в себе много полезной информации о предметной области.

Не случайно большое место в исследованиях по применению тезаурусов в различных приложениях автоматической обработки текстов занимают тезаурусы другого типа – тезаурусы типа WordNet, словарный состав которых является значительно более подробным, значительно более близок лексике документов. Однако описание в рамках WordNet также подвергалось многочисленной критике за раздельное описание частей речи, слишком большой набор не связанных между собой значений, недостаточную формализованность описания единиц (синсетов) и отношений между ними.

Формальную модель ресурса типа WordNet можно представить следующим образом:

$$WN = <LC_{n,adj,v,adv}, R_{n,adj,v,adv,} S, T, M, A_n>$$

где $LC_{n,adj,v,adv} = \{LC_n, LC_{adj}, LC_v, LC_{adv}\}$ – совокупность лексикализованных понятий-синсетов, сгруппированных по разным частям речи (существительные, прилагательные, глаголы и наречия); синсет представляется собой одну лексему (слово в определенном значении) или совокупность синонимичных лексем,

- $R_{n,adj,v,adv} = \{R_n, R_{adj}, R_v, R_{adv}\}$ – наборы отношений между синсетами, различающиеся для разных частей речи,

- T – текстовые выражения (слова и словосочетания), описанные в ресурсе,

[1] Z39.19 – Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO, 2005.

- S – отношения между текстовыми выражениями и синсетами,

- M – совокупность неоднозначных текстовых выражений: $M \subset T$,

- A_n – аксиомы транзитивности и наследования, индекс n отражает тот факт, что аксиомы обсуждаются и используются в подавляющем большинстве случаев только для синсетов существительных.

В данной главе также рассмотрены эксперименты по применению тезауруса WordNet для концептуального индексирования при информационном поиске, а также методы использования тезаурусов в вопросно-ответных системах и системах автоматической рубрикации. В заключении главы делается вывод о сложности согласованного описания онтологических и лингвистических знаний в едином компьютерном ресурсе. Кроме того, подчеркивается, что применение таких ресурсов для автоматической обработки текстов в широких предметных областях требует развития специализированных моделей и методов.

Во второй главе представлена формализованная модель лингвистической онтологии, предназначенной для автоматической обработки текстов в широкой предметной области. В **разделе 2.1** показано, что ресурс для автоматической обработки текстов в информационно-поисковых приложениях в широких предметных областях должен сочетать принципы различных традиций и методологий:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet;
- методологии создания формальных онтологий.

В результате исследований и экспериментов сформулированы следующие принципы создания онтологических ресурсов для автоматической обработки текстов (далее ЛО – лингвистическая онтология для автоматической обработки текстов).

Онтологию ЛО для предметной области D можно формально представить следующим образом:

$$ЛО = < C, Ex, NO, R_{lo}, A_{tr,i}, S, T, M_{m,a}, L, DC >$$

где C – множество понятий онтологии, где понятие обозначает класс сущностей, обладающих одинаковыми свойствами и отношениями к другим классам сущностей;
Ex – множество экземпляров понятий онтологии, задано отображение $E: C \rightarrow 2^{Ex}$;
NO – множество имен понятий и экземпляров в онтологии, имена уникальны;
R_{lo} – набор отношений между понятиями $R \subset C \times C$, специально разработанный для автоматической обработки текстов в широких предметных областях;
$A_{tr,i}$ – аксиомы, основанные на свойствах транзитивности и наследования отношений;
T – множество текстовых входов онтологии – языковых выражений, значения которых представлены в онтологии;
S – множество отношений между языковыми выражениями (T) и понятиями (C): $\{s\ (c_i, t_j)\}$;
$M_{m,a}$ – множество многозначных слов и выражений из T: $M_{m,a} \subset T$; многозначные текстовые входы онтологии делятся на два подвида: M_m – текстовые входы, которые относятся к более, чем одному понятию онтологии, и M_a – текстовые входы, которые многозначны, но в онтологии представлено только одно значение: $M_{m,a} = M_m \cup M_a$;
L – множество лемматических представлений языкового выражения (т.е. представление выражения в виде последовательности слов в словарной форме, например, словосочетания *ценная бумага* представляется в лемматическом виде как *ЦЕННЫЙ БУМАГА*),
DC – это отображение терминологического состава (TD) заданной коллекции предметной области ($Dcoll$) на текстовые входы и понятия онтологии:

$$DC: (Dcoll, TD) \rightarrow (T, C).$$

Отображение *DC* задает критерий минимальной полноты онтологии, которая должна обеспечивать покрытие терминологического состава заданной коллекции предметной области, что собственно и отражает суть лингвистической онтологии.

Таким образом, *лингвистическая онтология предметной области представляет собой базу знаний онтологического типа о понятийной системе и лексико-терминологическом составе предметной области.*

Единицей онтологии ЛО является понятие, как единица в системе понятий, имеющая свои специфические свойства, отличающие данную единицу от других единиц в системе понятий. Такой взгляд соответствует как современной трактовке дескрипторов в информационно-поисковых тезаурусах, так и понятий (классов) в онтологиях. Каждое введенное понятие c_i должно иметь однозначное имя n_i. Именем понятия может являться однозначное слово или словосочетание, значение которого соответствует этому понятию (т.е. один из текстовых входов понятия).

Каждое понятие c_i снабжается набором текстовых входов $\{t_{ij}\}$ – языковых выражений, значения которых соответствуют данному понятию. Такие языковые выражения являются между собой онтологическими синонимами. В текстах может встречаться множество вариантов текстовых входов того или иного понятия, как, например, известно о существовании множественной вариативности терминов предметной области.

В **разделе 2.2** рассматривается предлагаемая модель отношений в лингвистической онтологии. Система отношений ЛО представляет собой небольшой набор отношений, и в этом предлагаемая модель лингвистической онтологии близка к традиционным информационно-поисковым тезаурусам. Однако для установления отношений применяются более строгие онтологические критерии. С каждым отношением связан свой набор аксиом, которые имеют важное значение для различных этапов автоматической обработки текстов и приложений информационного поиска.

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений должны выполнять разнообразные функции: использоваться в классических функциях информационно-поисковых тезаурусов для расширения поискового запроса или вывода рубрики документа; учитываться при разрешения многозначности языковых единиц, включенных в лингвистическую онтологию; применяться для выявления лексической связности в текстах в целях улучшения качества обработки связных текстов.

Для реализации каждой из этих функций необходимо осуществление специализированного логического вывода. В условиях широкой предметной области и, следовательно, необходимости создания лингвистической онтологии большой величины, для обработки текстов, не ограниченных по стилю, жанру, величине, наиболее стабильно можно опираться на те отношения, которые не исчезают, не меняются в течение всего срока существования любого или подавляющего большинства экземпляров понятия, так лес всегда состоит из деревьев. В диссертации показано, что надежные отношения связаны с сосуществованием понятий или их экземпляров.

В качестве аксиом в ЛО используются свойства транзитивности и наследования:

$$r(c_i, c_j) \wedge r(c_j, c_k) \rightarrow r(c_i, c_k) \qquad (A_{tr})$$

$$r(c_i, c_j) \wedge r_1(c_j, c_k) \rightarrow r_1(c_j, c_k) \qquad (A_i)$$

Основными отношениями в предлагаемой модели онтологии ЛО является следующий набор надежных отношений: *выше-ниже*, *часть-целое*, несимметричная ассоциация: *асц*$_1$-*асц*$_2$, симметричная ассоциация – *асц*.

Подраздел 2.2.1. посвящен рассмотрению отношения *выше-ниже,* которое трактуется как отношение *класс-подкласс* в формальных онтологиях, и для этого отношения

традиционно предполагается выполнение свойств транзитивности и наследования. Существенными при построении ЛО являются формальные проверки правильности установления отношения *класс-подкласс,* поскольку при нарушении онтологических принципов не выполняются вышеупомянутые свойства этого типа отношений.

Новым при построении модели отношений в ЛО является существенное использование для установления нетаксономических отношений так называемого *отношения онтологической зависимости,* которое рассматривается в **подразделе 2.2.2.**

Отношение онтологической зависимости возникает тогда, когда существование одной сущности зависит от существования другой сущности[1]:

X онтологически зависит от Y тогда и только тогда, когда X существует только, если Y существует.

$D\ (X,\ Y) = def\ (существует\ (X) \rightarrow существует\ (Y))$

Отношения онтологической зависимости разделяются на несколько подвидов.

При *специфической зависимости* (SD) конкретная сущность e_1 зависит от другой конкретной сущности e_2, если необходимо, чтобы e_2 существовал, если e_1 – существует:

$$SD\ (e_1, e_2) =_{def}\ ((\exists\ t\ \ pre\ (e_1, t)) \wedge \forall t\ (pre\ (e_1, t) \rightarrow\ pre\ (e_2, t)))$$

где $pre\ (e_i, t)$ – предикат существования сущности e_i в заданное время *t.*

Например, существование конкретного человека зависит от существования его мозга, кроме того, мозг не может быть заменен на другой мозг, т.е. это специфическая зависимость.

Отношение специфической зависимости между конкретными сущностями может быть естественно перенесено на специфическую зависимость между понятиями (CSD), т.е. понятие c_1 является специфически зависимым от понятия c_2, если все экземпляры c_1 специфически зависят от экземпляров c_2, т.е.

$$CSD\ (c_1, c_2) =_{def}\ (\forall e_1 \in E(c_1)\ \exists e_2 (e_2 \in E(c_2) \wedge SD\ (e_1, e_2)))$$

При *родовой зависимости* (generic – GD) существование конкретной сущности зависит от существования конкретных сущностей, относящихся к некоторому понятию *c*:

$$GD\ (e, c) =_{def}\ ((\exists t\ \ pre\ (e, t)) \wedge$$
$$(\forall t\ (pre\ (e, t) \rightarrow \exists\ e_c\ (e_c \in E\ (c) \wedge\ pre\ (e_c, t))))$$

Данное отношение также может быть перенесено на отношения между понятиями (CGD).

Кроме того, могут быть выделены *внутренняя онтологическая зависимость*, т.е. зависимость от внутренних свойств или частей сущности, и *внешняя онтологическая зависимость*, т. е. онтологическая зависимость от существования некоторой отдельной сущности. Наконец, может быть выделена онтологическая зависимость по определению: *сказать, что сущность X зависит от Y, это означает сказать, что Y необходимо (eliminably) должно быть использовано в любом определении X*[2].

В диссертации показано, что принципы установления отношений ассоциации, провозглашаемые в различных руководствах и стандартах по разработке информационно-

[1] Guarino N., Welty C. Evaluating ontological decisions with ONTOCLEAN // Communications of the ACM. 2002. V. 45(2). P. 61-65.

[2] Masolo C., Vieu L., Bottazzi E. Catenacci C., Ferrario R., Gangemi A., Guarino N. Social roles and their descriptions // In Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning. AAAI Press. 2004.

поисковых тезаурусов, часто связаны с проверкой условий онтологической зависимости понятий между собой. Кроме того, приводятся результаты экспериментов, показывающие, что результаты использования тезаурусного отношения ассоциации для автоматического расширения запроса при информационном поиске коррелируют с существованием между соответствующими понятиями отношения онтологической зависимости и его конкретным подвидом.

Отношения онтологической зависимости играют существенную роль при описании в ЛО отношений *часть-целое* **(подраздел 2.2.3)** и отношений несимметричной ассоциации **(подраздел 2.2.4).**

Особенностью отношения *часть-целое* является как его широкое распространение в различных предметных областях, так и разнообразие его проявлений. Наиболее известный подтип отношения *часть-целое* относится к взаимоотношениям между физическими объектами, однако это отношение может устанавливаться и между сущностями, длящимися во времени, между группами сущностей, ролями и процессами и др., в результате чего в разных компьютерных реализациях имеются значительные расхождения в принципах установления данного отношения. Важность данного отношения связана предполагаемым свойством транзитивности, при этом имеются многочисленные примеры видимого нарушения этого свойства.

При рассмотрении отношения *часть-целое* важным является выделение его онтологических свойств, т.е. анализ сосуществования части и целого[1].

Так, часть e_1 называется *неотделимой частью* e_2, если e_1 специфически зависит от e_2, и e_1 является частью e_2:

$$IP\ (e_1, e_2) =_{def}\ (\forall t(\ pre\ (e_1, t) \rightarrow\ PP\ (e_1, e_2)))$$

где PP – предикат *быть частью:* e_1 является частью e_2.

Примером неотделимой части является мозг человека, который не может существовать вне своего целого. Сердце человека может быть отделено от конкретного человека и пересажено другому человеку. Но при этом должна существовать сама категория людей. Таким образом, сердце человека зависит от человека родовой зависимостью, и такая зависимость называется *обязательным целым MW*.

$$MW\ \ (e_1, c) =_{def}\ (\forall t(\ pre\ (e_1, t) \rightarrow\ \exists\ e_2(e_2 \in E(c) \wedge PP\ (e_1, e_2))))$$

Эти отношения могут быть перенесены на отношения между понятиями:
Отношение неотделимой части между понятиями (*CIP*):

$$CIP\ \ (c_1, c_2) =_{def}\ (\forall e_1(e_1 \in E(c_1) \rightarrow\ \exists e_2(e_2 \in E(c_2) \wedge IP\ (e_1, e_2))))$$

Отношение обязательного целого между понятиями (*CMW*):

$$CMW\ \ (c_1, c_2) =_{def}\ (\forall e_1(e_1 \in E(c_1) \rightarrow\ \exists e_2(e_2 \in E(c_2) \wedge MW\ (e_1, e_2))))$$

В **подразделе 2.2.3.3.** описываются решения, принятые для представления отношения *часть-целое* в предлагаемой модели ЛО. При описании данного отношения применяются следующие принципы: существование экземпляров понятия-части c_1 зависит от существования экземпляров целого c_2 *специфической или родовой онтологической зависимостью,* т.е. экземпляры понятия-части c_1 представляет собой неотделяемые части для

[1] Simons P. Parts. A study in Ontology. Oxford University Press, 1987.

экземпляров понятия-целого c_2 или экземпляры понятия c_2 являются обязательным целым для экземпляров c_1.

$$целое_{ло}\ (c_1, c_2) =_{def}\ (CIP\ (c_1, c_2) \vee CMW(c_1, c_2))$$

В диссертации показано, что отношение $целое_{ло}$ является транзитивным.

Накладывая вышеперечисленные условия установления отношения *часть-целое*, в предложенной модели ЛО не ограничиваются семантические подвиды частей: как части в ЛО могут рассматриваться физические части (*балкон зала - зрительный зал*), члены множества (*член партии - политическая партия*), характерные свойства *(водоизмещение - судно),* роли в процессах *(инвестор - инвестирование)* и др.

Таким образом, в настоящее время в ЛО используются следующие свойства отношения *часть-целое:*

часть $(c_1, c_2) \leftrightarrow$ *целое* (c_2, c_1)

целое $(c_1, c_2) \wedge$ *целое* $(c_2, c_3) \rightarrow$ *целое* (c_1, c_3) – транзитивность отношения

выше $(c_1, c_2) \wedge$ *целое* $(c_2, c_3) \rightarrow$ *целое* (c_1, c_3) – наследование отношения целое по отношению *выше-ниже.*

Установление третьего типа отношений ЛО – несимметричной ассоциации – также связано с отношением онтологической зависимости **(подраздел 2.2.4).**

В результате анализа результатов экспериментов в работе показано, что в лингвистической онтологии, предназначенной для автоматической обработки текстов, необходимо, прежде всего, отражать так называемую внешнюю родовую зависимость, т.е. зависимость существования понятия от существования другого понятия. Отношение внешней родовой зависимости является несимметричным, и для его обозначения используется отношение несимметричной ассоциации $асц_1 - асц_2$. Отношение $асц_1$ ведет от зависимого понятия к главному понятию отношения родовой зависимости, а отношение $асц_2$ является к нему обратным отношением.

В лингвистической онтологии постулируются следующие свойства отношения несимметричная ассоциация, представляющей в ЛО отношение внешней родовой зависимости:

$асц_1\ (c_1, c_2) \leftrightarrow асц_2 (c_2, c_1)$

Наследование отношения несимметричной ассоциации на виды и части:

выше $(c_1, c_2) \wedge асц_1\ (c_2, c_3) \rightarrow асц_1\ (c_1, c_3)$

целое $(c_1, c_2) \wedge асц_1\ (c_2, c_3) \rightarrow асц_1\ (c_1, c_3)$

В **подразделе 2.2.5** рассматриваются принципы установления отношений симметричной ассоциации, которые в предлагаемой модели ЛО применяются в весьма ограниченном наборе случаев.

В **разделе 2.3** рассматриваются группировки понятий (окрестности) и отношений (пути) в ЛО, которые полезны для различных приложений автоматической обработки текстов, использующие ЛО.

Для каждого понятия $c_i \in C$ может быть определена окрестность понятия $O_i \subset C$, такая, что $c_j \in O_i$, если существует набор понятий $\{c_1, \ldots c_k\}$ такой, что $r_1\ (c_i, c_1), \ldots r_2\ (c_n, c_{n+1}) \ldots r_r\ (c_k, c_j) \in R,$ и на основе аксиом A_{tr}, A_i выводимо отношение $r\ (c_i, c_j)$:

$r_1\ (c_i, c_1), \ldots r_2\ (c_n, c_{n+1}) \ldots r_r\ (c_k, c_j) \mapsto r\ (c_i, c_j)$

На множестве отношений ЛО может быть введено отношение иерархии I по следующим правилам:

выше $(c_1, c_2) \rightarrow I\,(c_1, c_2)$
целое $(c_1, c_2) \rightarrow I\,(c_1, c_2)$
асц1 $(c_1, c_2) \rightarrow I\,(c_1, c_2)$
асц $(c_1, c_2) \rightarrow I\,(c_1, c_2)$

Это отношение означает, что правый элемент отношения считается более высоким по иерархии, чем левый. Для отношения симметричной ассоциации оба члена отношения равноправны.

В окрестности понятия c_i можно определить верхнюю полуокрестность O^+ и нижнюю полуокрестность O^-:

$O^+(c_i) \cup O^-(c_i) = O\,(c_i)$
$c_j \in O^+\,(c_i)$, если $c_j \in O\,(c_i) \wedge I\,(c_i, c_j)$
$c_j \in O^-\,(c_i)$, если $c_j \in O\,(c_i) \wedge I\,(c_j, c_i)$

Пересечение $O^+(c_i)$ и $O^-(c_i)$ может быть непустым из-за существования отношений симметричной ассоциации, входящих в обе полуокрестности. Верхняя полуокрестность понятия c_i также называется *дерево-вверх* понятия c_i, нижняя полуокрестность понятия c_i – *деревом-вниз* понятия c.

Можно определить следующие виды путей между понятиями:

- *Путь по иерархии вверх* $P_{up}\,(c_0, c_{00})$: От понятия c_0 к понятию c_{00} существует путь по иерархии вверх, если $c_{00} \in O^+(c_0)$;

- *Путь по иерархии вниз* $P_{down}(c_0, c_{00})$: От понятия c_0 к понятию c_{00} существует путь по иерархии вниз, если $c_{00} \in O^-(c_0)$;

- *Путь с перегибом вверх* $P_{updown}(c_0, c_{00})$: Между понятиями c_0 и c_{00} такими, что $c_0 \notin O\,(c_{00})$ и $c_{00} \notin O\,(c_0)$, существует путь с перегибом-вверх, если существует точка перегиба – понятие c_i такое, что:

$$\exists\, c_i : c_i \in O^+(c_0) \wedge c_i \in O^+(c_{00})$$

- *Путь с перегибом вниз* $P_{downup}(c_0, c_{00})$: Между понятиями c_0 и c_{00} такими, что $c_0 \notin O\,(c_{00})$ и $c_{00} \notin O\,(c_0)$, существует путь с перегибом-вниз, если существует точка перегиба – понятие c_j такое, что:

$$\exists\, c_j : c_j \in O^-(c_0) \wedge c_j \in O^-(c_{00})$$

Введенные типы путей в ЛО используются в процедурах автоматического разрешения лексической неоднозначности, расширения поискового запроса, вывода рубрик по тексту.

В **разделе 2.4.** указывается, что описанные выше принципы создания лингвистических онтологий в широких предметных областях положены в основу разработки нескольких больших ресурсов для информационного поиска: Общественно-политического тезауруса[1] (1993 – н/в), Тезауруса русского языка РуТез (1997 – н/в), Онтологии по Естественным наукам и технологиям ОЕНТ (2004 – н/в), и ряда других. Вышеперечисленные ресурсы имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними, которые устанавливаются в соответствии

[1]Разработка первой версии Общественно-политического тезауруса была начата автором диссертации в инициативном порядке в 1993 году во время участия в проекте ПОЛИТЕКСТ (1993-1995), которым руководили Юдина Т.Н., Журавлев С.В., Леонтьева Н.Н.

с требованием правомочности расширения запроса по иерархии связей при информационном поиске.

Разнообразие предметных областей, для которых созданы лингвистические онтологии по предложенной модели, доказывает универсальность этой модели, ее способность описывать базовые свойства и отношения понятий, присутствующие в любой предметной области. Объемы созданных ресурсов демонстрируют удобство модели для быстрого наращивания ресурсов.

В третьей главе рассматриваются вопросы моделирования свойств связного текста, автоматического выявления тематической структуры текста на основе знаний, описанных в лингвистической онтологии. Указывается, что распространенной моделью обработки связного текста является модель мешка слов (bag of words), когда предполагается, что все слова в тексте употребляются независимо друг от друга, и, таким образом, значимость слова в тексте определяется как функция от особенностей употребления в тексте именно этого слова, прежде всего, от частоты его употребления в этом тексте.

Такая модель противоречит известной особенности связного текста, заключающейся в том, что если текст посвящен какой-то теме, то в нем употребляется множество слов и выражений, относящихся к этой теме. Наличие большого лингвистико-онтологического ресурса позволяет выявить взаимоотношения между словами. В данной главе описывается модель тематической структуры текста, которая позволяет эффективно преобразовывать информацию о взаимосвязи слов в оценку значимости конкретных слов в тексте и использовать полученные оценки в приложениях автоматической обработки текстов

В разделе 3.1 рассматриваются виды связности текста и существующие методы их автоматического моделирования. Одним из существенных видов связности текста является лексическая связность, которая состоит в повторе одних и тех же и близких по смыслу слов в связном тексте, которые образуют цепочки – так называемые лексические цепочки. Лексическая связность может моделироваться на основе знаний, описанных в тезаурусах и лингвистических онтологиях.

В разделе 3.2 рассматриваются автоматические методы моделирования лексической связности и построения лексических цепочек на основе информации об отношениях между словами, хранимыми в тезаурусах. Подавляющее число экспериментов по построению лексических цепочек было проведено с помощью тезауруса WordNet. В большинстве подходов используется алгоритм построения лексических цепочек, который строит лексические цепочки с начала текста, при этом очередное слово вносится в ту существующую цепочку, с одним из элементов которой имеется наиболее сильная связь по отношениям тезауруса.

Непосредственным применением лексических цепочек для приложений автоматической обработки текстов является автоматическое аннотирование (реферирование) текстов, поскольку автоматически составленная аннотация должна отвечать законам построения связного текста, быть понятной и связной. Методы автоматического аннотирования и применение автоматических лексических цепочек рассматриваются **в разделе 3.3.**

В разделе 3.4 рассматриваются проблемы построения лексических цепочек. Несмотря на то, что изначально представляется, что выявление лексических цепочек в связном тексте является интуитивно понятным, в работе показано, что в этом процессе очень велика субъективность, которая проявляется в том, что люди строят по тексту разные лексические цепочки. Кроме того, если учитывать более обширную систему отношений, позволять элементам входить в разные лексические цепочки, то резко возрастает сложность автоматического моделирования лексических цепочек, их излишнего порождения.

В разделе 3.5 предлагается новый подход к автоматическому выделению лексических цепочек из текстов, объясняющий вышеуказанные проблемы и предлагающий пути их решений. Основным принципом предложенного подхода является взгляд на роль лексических цепочек в тексте с точки зрения другого важного свойства текста – глобальной связности. Глобальная связность текста проявляется в том, что текст имеет единую тему. Тематическая структура текста представляет собой иерархическую структуру тем и подтем. Каждому предложению текста имеется некоторое соответствие в этой тематической структуре. Таким образом, предполагается, что содержание текста выражается в виде совокупности пропозиций: $P^D = \{p_0\ (c_{01}...c_{01}), p_1\ (c_{11}...c_{1n})..\ p_k\ (c_{k1}...c_{kn})\}$. Над этим множеством определено отношение частичного порядка, т.е. выполняются следующие свойства: рефлексивность, транзитивность, - антисимметричность.

У документа имеется основная тема – главная пропозиция $p_0\,(c_{01}...c_{0n})$:

$$\forall p_i\ ((p_i \in P^D) \rightarrow (p_i \preceq p_0))$$

Пропозиции тем (подтем) устанавливают отношения между тематическими элементами $c_1...c_n$. В иерархической тематической структуре главная тема $p_0\ (c_{01}...c_{0n})$ поясняется, характеризуется, дополняется деталями посредством подтем $p_1\ (c_{11},\ ....c_{im})\ \ldots\ p_i\ (c_{i1},\ ...c_{ij}...c_{im})$.

Рассмотрим две пропозиции $p_i\ (c_{i1},\ c_{i2}\ \ldots\ c_{ik})$ и $p_j\,(c_{j1},\ c_{j2}\ \ldots\ c_{jm})$ такие, что $p_i \in P^D$, $p_j \in P^D$, $p_i \preceq p_j$. Такие пропозиции связаны между собой и поэтому должны существовать взаимоотношения между участниками этих пропозиций, т.е.

$$\forall p_i, p_j\,(p_i \preceq p_j \rightarrow \exists\ c_{il}\, c_{jn}\, r_c\,(c_{il},\ c_{jn}))$$

В результате каждый тематический элемент c_{0i} основной пропозиции p_0 имеет представительство в пропозициях нижнего уровня p_j посредством связанных с ним по смыслу элементов пропозиции $l_{0ij.}$ Возникает структура типа узла: основной тематический элемент c_{0i} и связанные с ним элементы $l_{0ij.}$ Назовем такой узел тематическим узлом $tnode_i$: $tnode_i = (c_{0i,}\ \{l_{0i1}...l_{0ij}\})$. Множество всех тематических узлов, выделяемых в тексте, будем обозначать $Tnode=\{tnode_{1...}\ tnode_n\}$.

В диссертации показано, что основная роль лексических цепочек относительно тематической структуры текста состоит в обеспечении представительства тематических элементов более высоких уровней иерархии в подтемах более низкого уровня. По внутренней структуре лексическая цепочка имеет структуру узла с выделенным центральным элементом и некоторой совокупностью лексем, связанных с этим центральным элементом. Среди тематических узлов можно выделить основные тематические узлы и локальные тематические узлы. Основные тематические узлы имеют в качестве центра тематические элементы основной темы документа.

В диссертации также обосновывается, что для качественного построения тематических узлов необходим учет совместной встречаемости элементов тематических узлов в одних и тех же предложениях текста: если c_1 и c_2 часто встречаются в анализируемом тексте в одних и тех же простых предложениях, то это означает, что данный текст посвящен рассмотрению отношений между этими сущностями, т. е. c_1 и c_2 соответствуют разным тематическим элементам основной темы или подтемы текста и должны быть отнесены к разным лексическим цепочкам (тематическим узлам).

Для учета факторов построения тематического представления подходит представление распределения понятий текста в виде мультиграфа, т.е. графа с двумя типами дуг между вершинами. Один тип дуг, R_{sent}, отражает отношения между понятиями в ЛО. Другой тип дуг, R_{text}, отражает совместную встречаемость понятий в предложениях текста. В вершинах мультиграфа указана частотность упоминания соответствующего понятия в тексте. На дугах R_{text} отмечена частота встречаемости данной пары понятий в одних и тех же

предложениях текста. Дуги R_{sent} указывают частотность упоминания данной пары понятия в пределах нескольких предложений, но не в одном предложении текста. Таким образом, мультиграф MG тематического представления может быть определен как $MG = (V, fv, R_{text}, fr_{text}, R_{sent}, fr_{sent})$.

Для определения статуса тематического узла, т. е. определения, относится ли он к главной теме текста, локальной теме или просто упоминался, использованы следующие принципы. Предполагается, что основными тематическими узлами $Tnode^M$ в первую очередь являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями $\forall\, tnode_i,\ tnode_j(((tnode_i \in Tnode^M) \wedge (tnode_j \in Tnode^M)) \rightarrow fr_{text}(tnode_i, tnode_j) > 0)$, т.е. элементы всех пар основных тематических узлов должны обсуждаться в связи с друг другом в некоторых предложениях текста;

- сумма частот текстовых связей между ними максимальна для анализируемого текста:

$\sum_N fr_{text}(\forall\, tnode^M_i, tnode^M_j) \geq \sum_N fr_{text}(\forall\, tnode_i, tnode_j)$, где $N = \binom{m}{2}$ - *число сочетаний из величины, равной количеству основных текстовых узлов* $m = |Tnode^M|$.

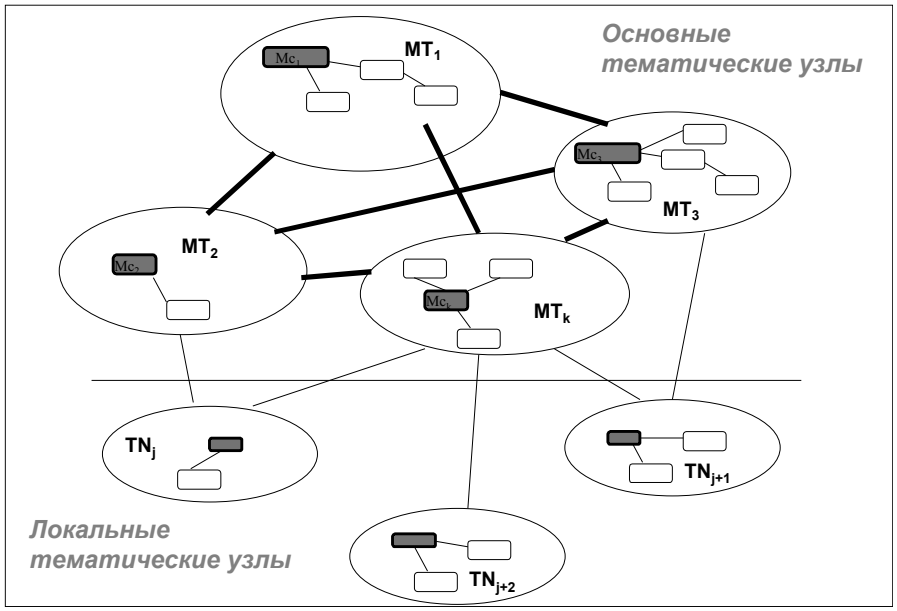

Рис. 1. Структура тематического представления. Линии между тематическими узлами представляют собой текстовые связи R_{text} в предложениях текста. Между основными тематическими узлами существуют наиболее частотные текстовые связи.

Таким разбиением тематических узлов *Tnode* на основные и локальные задается разбиение понятий, упомянутых в тексте, на следующие пять классов по их важности для анализируемого текста:

- главные понятия основных тематических узлов C^M (основные темы);
- другие понятия основных тематических узлов EC^M;
- главные понятия локальных тематических узлов C^L (локальные темы);
- другие понятия локальных тематических узлов EC^L;

- упоминавшиеся понятия C^O.

Таким образом, построено тематическое представление *TR* текста, в котором понятия лингвистической онтологии, упоминавшиеся в тексте, разбиты на тематические узлы *Tnode*. Между тематическими узлами фиксируются текстовые связи R_{text}. (рис. 1). В зависимости от принадлежности понятия к тому или иному классу понятий вычисляется значимость (вес) понятия.

В четвертой главе рассматриваются методы и алгоритмы применения изложенной модели лингвистической онтологии для автоматической обработки текстов и автоматически создаваемого тематического представления текста в приложениях информационного поиска.

В разделе 4.2 рассматривается предложенные методы автоматического разрешения многозначности на основе знаний, описанных в лингвистической онтологии. Основой методов является оценка семантической близости между возможными значениями, с одной стороны, и окружающим текстовым контекстом, с другой стороны. Лучшие результаты автоматического разрешения многозначности показал метод LocGlob, учитывающий как локальный контекст употребления многозначного слова, так и глобальный контекст (целый текст).

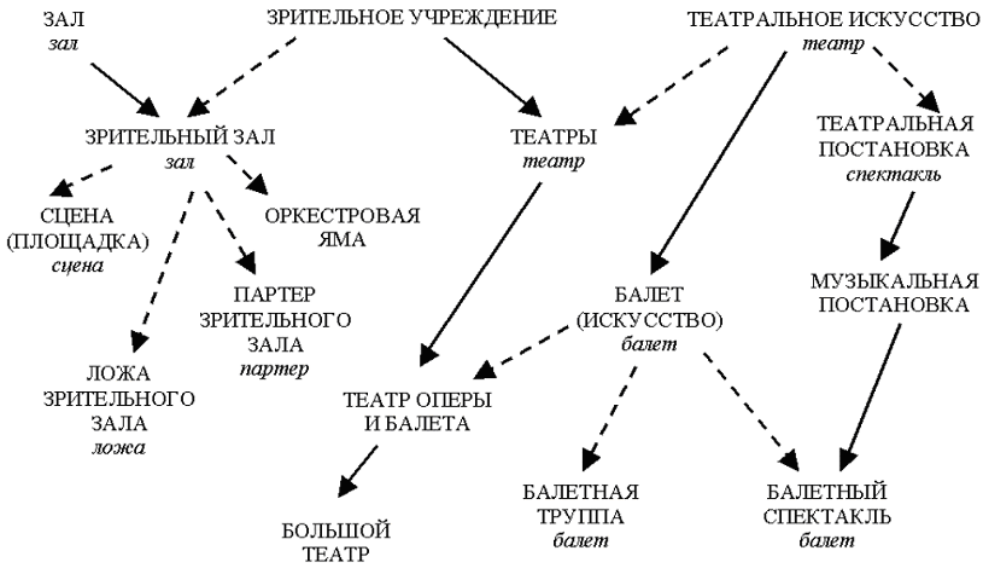

Рис. 2. Фрагмент тезауруса РуТез, демонстрирующий описание многозначных слов *балет, зал, сцена, театр* и концептуальные пути между соответствующими понятиями

Семантическая близость F_{sim} между двумя понятиями c_1 и c_2 оценивается на основе рассмотрения пути отношений, который существует между этими единицами ЛО. При рассмотрении путей вводятся ограничения на конфигурации путей между понятиями c_1 и c_2, которые рассматриваются при оценке семантической близости понятий, а именно, либо путь должен состоять из совокупности иерархических отношений, направленных в одну сторону, например, последовательность отношений от вида к роду (иерархический путь) (P_{up} или P_{down}), либо такой путь должен включать ровно один перегиб, т.е. изменение направления движения (путь с перегибом P_{updown} или P_{downup}) (рис. 2).

Выбор значения многозначного слова производится на основе вычисления весов значений в данном контексте, который представляет собой линейную функцию, зависящую от 47 параметров, включая различные конфигурации путей между понятиями в онтологии. Параметры алгоритма подбирались на основе размеченной лексическими значениями коллекции текстов *методом координатного спуска.*

Таблица 1. Точность разрешения лексической многозначности по источникам публикаций

Источник	N_{doc}	N_{amb}	$P_{locglob}$, %	P_{glob}, %
Известия	44	2525	**75.23**	72.00
Ведомости	62	2697	**77.89**	73.41
Независимая газета	42	2776	**68.14**	66.50
Комсомольская правда	49	2240	**66.74**	63.04
Яндекс-Новости	30	450	**75.05**	68.00
Всего	227	10688	**73.37**	68.77

Результаты работы алгоритмов разрешения многозначности по каждому из источников показаны в Табл. 1, где N_{doc} – число документов, N_{amb} – число вхождений неоднозначных терминов, $P_{locglob}$ – точность по алгоритму LocGlob, P_{glob}– точность по алгоритму Glob с более простой, булевской функцией определения F_{sim}.

При тестировании алгоритмов разрешения лексической многозначности на базе тезауруса РуТез для алгоритма LocGlob была получена точность разрешения многозначности – 57.14%, с учетом разрешения многозначности за счет попадания в словосочетания, описанные в тезаурусе – 63.4%. Результаты в сопоставимой задаче автоматического разрешения многозначности для английского языка, которая тестировалась в рамках конференции SENSEVAL-3 имеет точность – 50.89%.

В разделе 4.3 рассматривается метод создания концептуального индекса документов в информационно-поисковой системе, т.е. индекса понятий ЛО, в котором (в отличие от пословного индекса) выражения-синонимы сведены к одному элементу индекса, а разные значения неоднозначных выражений сопоставлены разным элементам индекса. Вес понятия c в концептуальном индексе документа D вычисляется на основе статуса понятия в тематическом представлении документа и частоты упоминания данного понятия в документе:

$$\mu(c, D) = \lambda \cdot \nu^*(c, D) + (1-\lambda) \cdot freq(c, D) \cdot [freq^*(D)]^{-1}$$

где $\nu^*(c, D) = max_{Th(c, D)} \nu(c, D)$ – максимум из весов понятия c в тематических узлах Tnode;

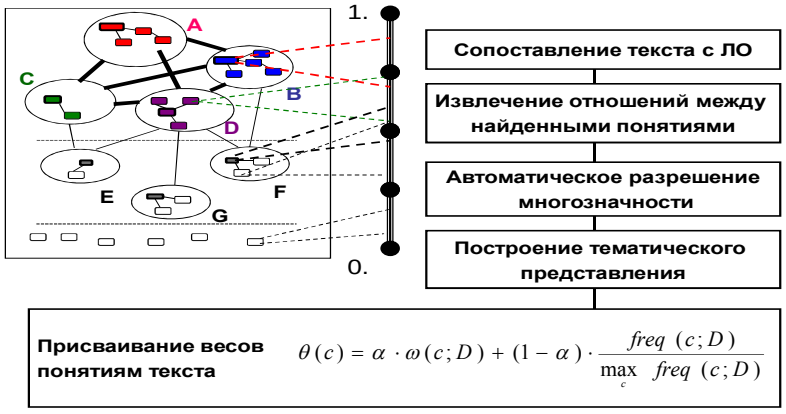

Рис. 3. Схема построения тематического представления и концептуального индекса документа

freq (*c*, *D)* – частота понятия *c* в документе D, *freq*(D) = max* $_{d \in D}$ *freq (d, D)* – максимальная частота среди понятий документа D; λ – подбираемый параметр формулы (рис. 3).

Представлена также модель определения близости запроса и документа с учетом концептуальных окрестностей понятий, упомянутых в запросе.

Модель поиска по концептуальному индексу была протестирована на коллекции нормативно-правовых документов УИС РОССИЯ (uisrussia.msu.ru). Концептуальный индекс строился на основе Общественно-политического тезауруса. В качестве запросов использовались случайно выбранные рубрики из Классификатора правовых актов[1]. Тестирование показало, что средняя точность поиска по таким запросам на основе концептуального индекса (с учетом синонимов и многословных выражений) возрастает более, чем в полтора раза по сравнению с качеством поиска по отдельным словам. Данный результат подтверждает также и качество предложенной модели описания отношений в ЛО.

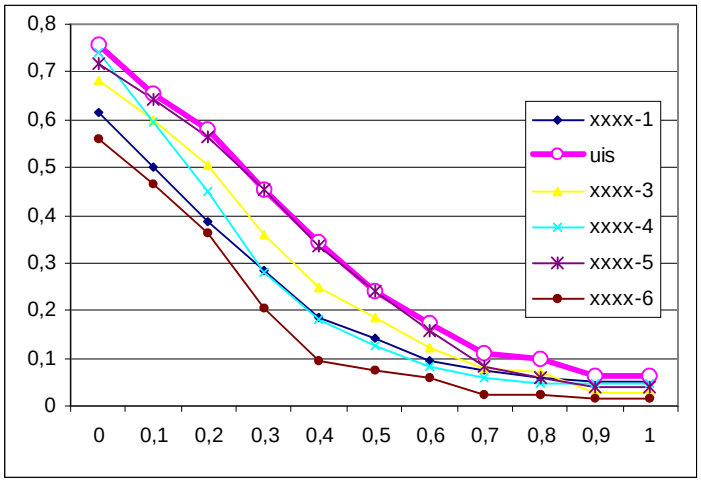

Рис. 4. Результаты тестирования поиска по длинным запросам на семинаре РОМИП-2008

В рамках семинара РОМИП-2008 тестировалась комплексная модель информационного поиска по правовым документам, важнейшим составным компонентом которой является использование концептуального индекса, построенного по лингвистической онтологии и тематическому представлению. Представленная модель показала лучший результат из представленных алгоритмов, получив на первых 35 документах, которые были полностью оценены людьми-оценщиками, показатель средней точности MAP, который превышает показатель следующего участника (27.6%) на 7% (рис. 4).

В **разделе 4.3.4.** описывается специализированный алгоритм обработки сверхдлинных вопросов, описывающих сложную ситуацию, требующую консультативной помощи, так называемая феноменологическая модель.

Феноменологическая модель преобразует заданный вопрос в булевский поисковый запрос типа конъюнкция дизъюнкций над понятиями лингвистической онтологии:

$$\bigcap_i \bigcup_j c_{i,j},$$

где $c_{i,j}$ — понятия ЛО.

[1] УКАЗ Президента РФ от 15.03.2000 N 511. О Классификаторе правовых актов.

Элементами дизъюнкции могут быть понятия ЛО, которые рассматриваются как близкие по смыслу – они связаны между собой путями отношений определенного вида. Формула описания запроса наращивается по шагам. Если в выдаче слишком много документов, то наращиваются конъюнкции булевского выражения, если слишком мало – дизъюнкции. В качестве понятий, которыми могут быть дополнены дизъюнкты, могут использоваться: понятия ЛО, имеющие разрешенные пути отношений к начальным понятиям дизъюнкций D_i^0, при условии выполнения дополнительных условий (наличия в формулировке запроса, наличие в информере выдачи).

Феноменологическая модель работает в составе многофакторной модели. Значимость факторов модели определялась на основе *метода координатного спуска*. Качество комбинированной модели, включая феноменологическую модель, тестировалось на 165 запросах типа «формулировка проблемы» в юридической области экспертами-юристами на коллекции документов, отвечающих на такие вопросы (40 тысяч документов). Оценка производилась по показателю точности по первым документам - Precision(n). В табл. 2 приводится точность для разных методов по 5, 10, 15, 20 документам.

Метод	5	10	15	20
Поиск по комбинированной векторной модели	55.88	49.45	44.28	40.55
Максимальный результат, полученный по леммам (векторная модель + упорядочение по предложениям + замешивание полученных весов) – то есть без всякого участия знаний ЛО	68.00	57.70	52.89	47.70
Комбинированная модель, включая феноменологическую модель	**76.48**	**65.21**	**57.54**	**51.82**
Яндекс-сервер (лучший результат по настройкам)	50.84	43.07	38.85	35.37

Табл. 2. Точность по 5 первым документам по различным применяемым моделям

В разделе 4.4 рассматривается предложенный метод автоматической рубрикации на основе тезаурусных знаний и автоматически порождаемого тематического представления текстов. При создании лингвистического профиля рубрикатора каждая рубрика R описывается дизъюнкцией альтернатив, каждый элемент дизъюнкции представляет собой конъюнкцию:

$$R = \bigcup_i D_i \; ; \qquad D_i = \bigcap_j K_{ij} \; ,$$

Элементы конъюнкции в свою очередь описываются экспертами с помощью так называемых «опорных» понятий ЛО. Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее, каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия. Опорное понятие может быть как «положительным», т. е. добавлять нижерасположенные понятия в описание элемента конъюнкции, так и «отрицательным», т. е. вырезать из описания рубрики свои подчиненные понятия. Результатом применения расширения опорных понятий является совокупность понятий ЛО, полностью описывающая элемент конъюнкции:

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn}) = \bigcup_k d_{ijk} \; .$$

Вес элемента конъюнкции рассчитывается по формуле:

$$\theta(K_{ij}) = \max_k \left(\theta(d_{ijk})\right),$$

где d_{ijk} – понятия ЛО, полученные из опорных понятий, приписанных элементу конъюнкции экспертом, посредством применения функции расширения; $\theta(d_{ijk})$ – вес понятия, полученный на основе построенного тематического представления. Вес конъюнкции в целом предназначен учитывать не только сумму весов составляющих его понятий, но и оценку близости упоминания понятий в тексте.

Система автоматической рубрикации тестировалась в задаче классификации Web-страниц в рамках семинара РОМИП 2007. Работа по описанию 247 рубрик задания была выполнена за 8 часов рабочего времени. В опорных булевских выражениях было использовано около 900 понятий тезауруса, в расширенных булевских выражениях содержится около 40 тысяч понятий (с повторениями).

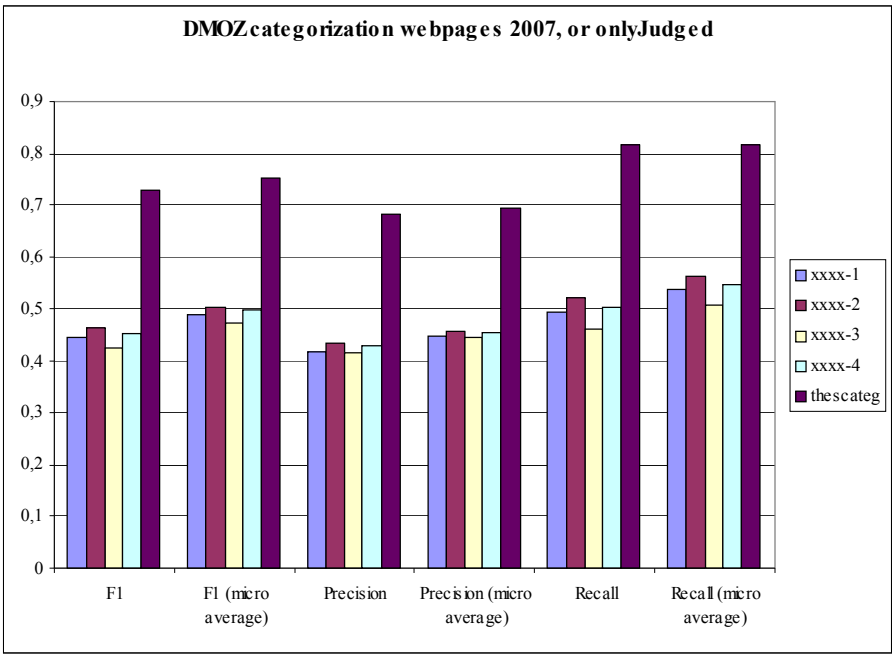

Рис. 5. Результаты рубрикации веб-страниц на основе предложенной технологии в экспериментах РОМИП-2007 (система thescateg)

На выбранных организаторами для тестирования 19 рубриках созданная система рубрикации показала наивысшие показатели классификации по F1-мере. По метрике OR (документ считается относящимся к рубрике, если хотя бы один из оценщиков отнес его к данной рубрике) для оцененных документов величина F1 составила 72%, что более чем на 56% превысило показатели следующей по качеству результатов системы 46% (рис. 5). На основе созданной технологии рубрикации создано более 20 разных систем автоматической рубрикации для разных организаций и проектов.

В разделе 4.5 рассматриваются предложенные методы автоматического аннотирования текстов, т.е. составления краткого реферата отдельного текста или совокупности текстов, на основе тезаурусных знаний и тематического представления текста.

Основным принципом предложенного метода автоматического аннотирования является следующий: то новое и важное, что несет в себе текст и что должна отразить в себе аннотация, это именно то, каким образом взаимодействуют между собой эти главные участники. Отсюда следует основной принцип составления аннотаций: важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те

предложения текста, которые содержат, по крайней мере, два понятия, входящих в состав разных основных тематических узлов текста.

Для составления **аннотации отдельного документа** существенным условием сбора связной аннотации является то, для каждой пары выявленных основных тематических элементов текста (основных тематических узлов) в аннотацию выбираются предложения, содержащие первое вхождение этой пары, следуя по порядку текста.

Качество технологии автоматического аннотирования отдельного документа тестировалось на конференции SUMMAC по оценке методов автоматического аннотирования. Реализованная система автоматического аннотирования имела лучший показатель F-меры для аннотаций наилучшей длины[1].

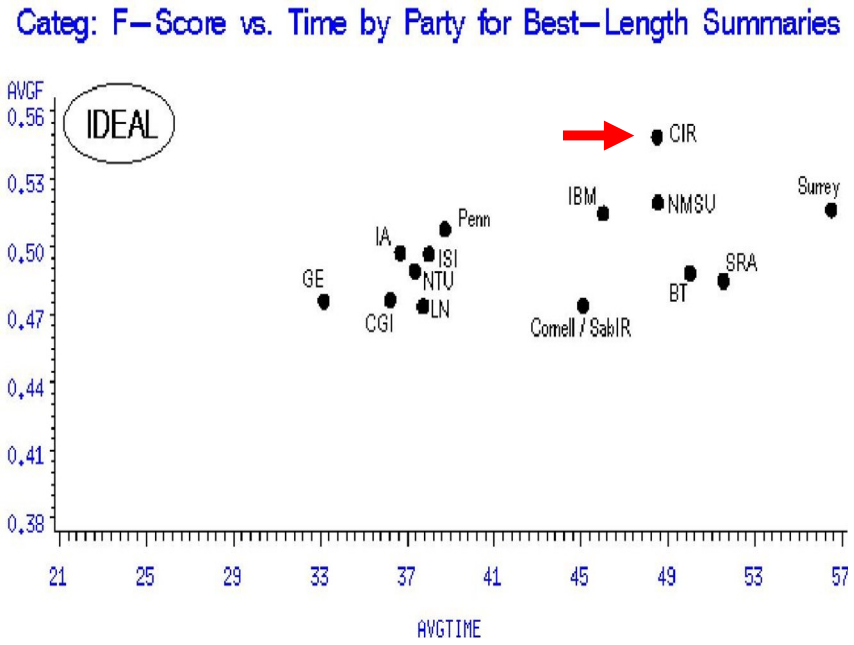

Рис. 6. Результаты работы предложенного метода аннотирования, продемонстрированные на международном тестировании SUMMAC.

График (рис. 6) отражает соотношение качества аннотации (F-мера по оси Y) и времени на принятие решения. Стрелкой показан результат предложенного алгоритма.

При обработке современных новостных потоков большое значение имеет **автоматическое аннотирование новостных кластеров**, совокупностей сообщений на одну и ту же тему. Новостной кластер представляет собой совокупность тематически близких документов. Поэтому тематическая структура новостного кластера так же, как и отдельного элемента выявляется за счет построения тематического представления этого кластера, и это представление используется для управления набором предложений в аннотацию кластера, а именно для решения таких задач как обеспечение полноты, снижения повторов, а также обеспечения связности аннотации кластера.

Для проверки предложенной модели аннотирования новостного кластера был проведен следующий эксперимент. Аннотация представляла собой заголовок и четыре предложения. Для новостных кластеров были получены их тематические представления. Далее ручные аннотации были размечены на предмет наличия основных тематических узлов

[1] Mani I., House D., Klein G., Hirshman L., Firmin Th., Sundheim B. SUMMAC: a text summarization evaluation // Natural Language Engineering. 2002. V.8, N 01. P. 43-68.

для данного кластера и именованных сущностей. Результатом проведённого анализа явился тот факт, что 83% предложений реальных ручных аннотаций (от общего числа предложений), сделанных экспертами-лингвистами, удовлетворяют сделанным предположениям. Особенность оставшихся 17% предложений состоит в том, что все они являлись последними предложениями ручной аннотации. Проведенный эксперимент доказывает, что сделанные предположения в методе автоматического аннотирования новостных кластеров имеют высокую корреляцию со структурой человеческих аннотаций.

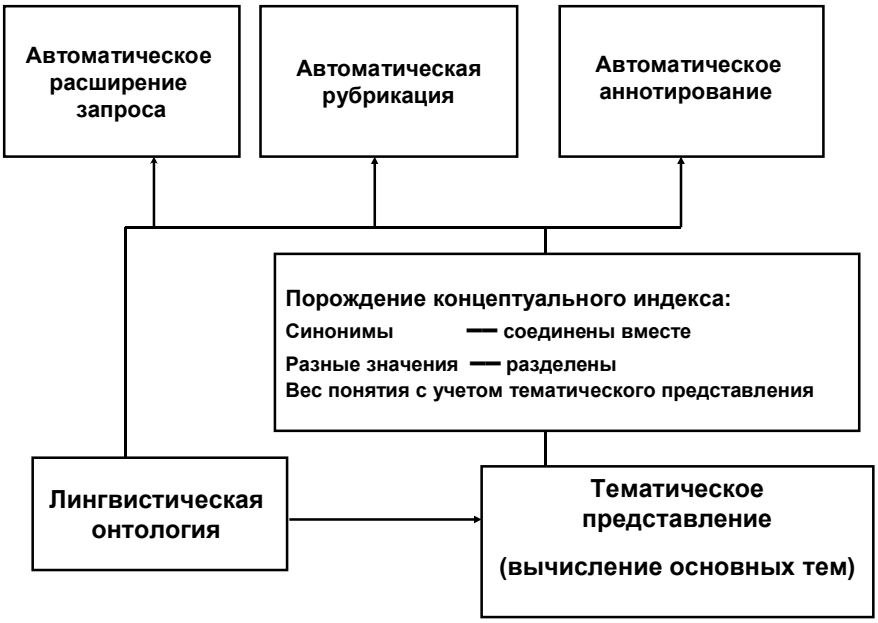

Рис. 7. Схема приложений автоматической обработки текстов на основе ЛО

В разделе 4.6 описывается применение описанных технологий в программном комплексе АЛОТ[1]. Программно-лингвистический комплекс АЛОТ производит автоматическую обработку поступающих на вход информационной системы потоков документов. Получая на входе файлы в формате HTML, АЛОТ на выходе выдает текстовые файлы в специальном формате, содержащие морфологический (нормализованные слова документа) и тематический индексы (термины и рубрики), предназначенные для дальнейшей загрузки в базу данных. На основе исследований, описанных в данной работе в программном комплексе АЛОТ, выполняются следующие этапы автоматического анализа текстов:

- сопоставление единиц ЛО с морфологическим представлением текста;
- автоматическое разрешение лексической неоднозначности,
- построение тематического представления текста,
- формирование концептуального индекса понятий ЛО с весами, полученными на основе построенного тематического представления
- автоматическое рубрицирование,
- автоматическое аннотирование (рис. 7).

[1] Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н., Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе РОССИЯ. // Электронные библиотеки, 2002 - Том.5 - Выпуск 2

Глава 5 описывает многофакторную модель для автоматического извлечения терминов из текстов для автоматизированного наращивания состава лингвистической онтологии.

В работе показано, что автоматический отбор терминов должен базироваться на совокупности различных признаков слов и словосочетаний, которые должны быть объединены в многофакторную модель. Вместе с тем, такие многофакторные модели должны быть переносимы с одной предметной области на другую. В данном исследовании предлагается подход по выявлению большого количества признаков для автоматического извлечения терминов из текстов и комбинирования этих признаков методами машинного обучения (рис. 8).

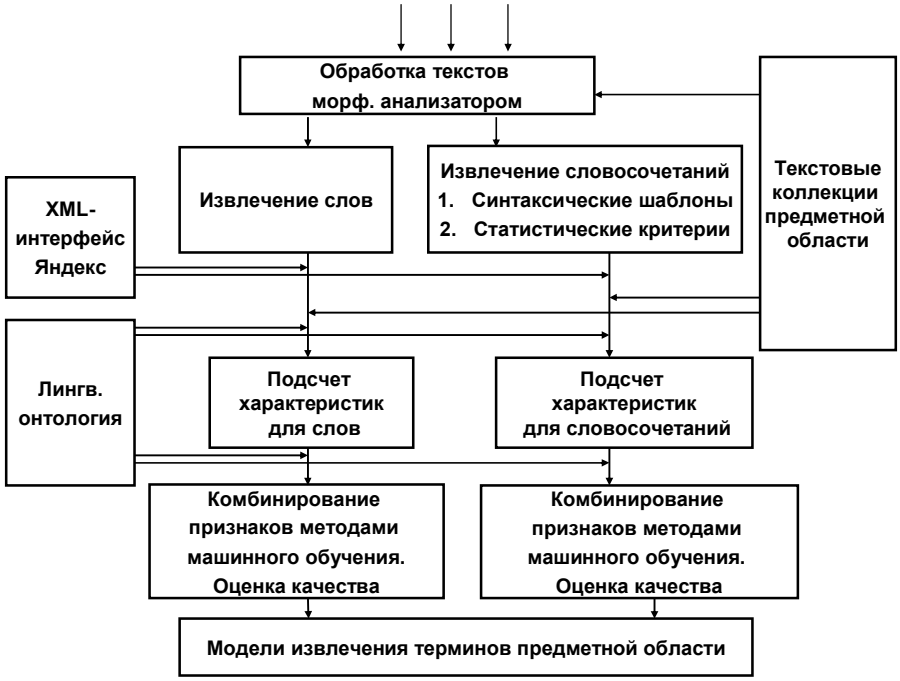

Рис. 8. Схема предложенного метода автоматического извлечения терминов. Лингвистические онтологии используются на двух этапах: как один из источников признаков и как средство для оценки качества извлечения

В предлагаемой модели используется три типа признаков для извлечения терминов:

- признаки, построенные на основе текстовой коллекции предметной области;
- признаки, полученные на основе информации глобальной поисковой машины,
- признаки, полученные на основе заданного тезауруса предметной области.

Так моделируется ситуация развития существующего тезауруса, при которой знания, описанные в текущей версии тезауруса, должны использоваться для автоматического извлечения новых терминов. Наборы признаков отличаются для отдельных слов, словосочетаний из двух слов и словосочетаний с большим количеством слов.

Для комбинирования выделенных признаков для наилучшего извлечения терминов предметной области применяются методы машинного обучения. Задачей применения методов является переупорядочение исходного списка слов (первоначально упорядоченного по мере снижения частотности) так, чтобы в начало списка попало как можно больше терминов. Таким образом, наилучшее переупорядочение списка снизит трудозатраты эксперта по вводу терминов в терминологические ресурсы – эксперт будет меньше просматривать слова, не являющиеся терминами. Для оценки качества такого упорядочения

используется мера, заимствованная из информационного поиска – так называемая средняя точность – AvP.

Для нахождения комбинации признаков, наилучшим образом отделяющих термины от нетерминов, используется метод машинного обучения – логистическая регрессия. При применении метода для извлечения терминов предполагается, что языковые объекты (слова и выражения) описываются n числовыми признаками f_j: $X \to R$, $j=1,.., n$.. Для построения классификатора решается задача минимизации эмпирического риска с функцией потерь вида:

$$Q(w) = \sum_{1}^{m} \ln(1+\exp(-y_i < x_i, w >)) \to \min_w$$

Кроме того, можно оценивать апостериорные вероятности принадлежности объекта x классам следующим образом:

$P(y|x) = \sigma(-y < x, w >), y \in Y$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ – сигмоидная функция.

Поскольку на практике данная апостериорная вероятность часто трактуется как оценка удаленности объектов от границ классов, то такая оценка позволяет, в нашем случае, упорядочивать языковые выражения в порядке снижения вероятности отнесения этого выражения к классу терминов и применять меру средней точности AvP для оценки результирующего качества упорядочения списка выражений по доле, содержащихся в нем терминов.

Для оценки качества предложенного комбинированного метода извлечения терминов проводились эксперименты в двух предметных областях. Одной из областей является широкая область по естественным наукам и технологиям, второй областью – банковская область. Для каждой предметной области имеются соответствующие текстовые коллекции, из которых извлечены слова и словосочетания – кандидаты в термины. В результате экспериментов было показано, что предложенные признаки терминов и их комбинация позволяют значительно улучшить качество выделения терминов в двух различных предметных областях.

В заключении диссертационной работы перечисляются ее основные результаты.

Основные оригинальные результаты, полученные в диссертации:

1. Предложена новая формализованная модель базы знаний онтологического типа – лингвистической онтологии, предназначенной для использования в автоматической обработке текстов в широких предметных областях. Модель неоднократно использовалась для создания сверхбольших лингвистико-онтологических ресурсов в разных предметных областях.

2. Предложена новая модель представления тематической структуры текстов на основе согласованного учета свойств лексической и глобальной связности текста. Предложен и реализован алгоритм автоматического построения тематического представления содержания текстов, который моделирует основное содержание текста посредством выделения тематических узлов – совокупностей близких по смыслу понятий текста.

3. Предложен и реализован метод концептуального индексирования документов для информационно-поисковой системы, базирующийся на понятиях лингвистической онтологии и тематическом представлении текста. В состав метода входит процедура автоматического разрешения лексической многозначности, основанная на информации о локальном и глобальном контексте употребления многозначного слова.

4. Предложен и реализован метод автоматического многошагового построения булевского выражения для длинного поискового запроса на естественном языке, включающий итерационное расширение запроса по отношениям лингвистической онтологии, подтвержденным поисковой выдачей.

5. Предложены и реализованы методы автоматической обработки текстов на основе концептуального индекса, включая:

- метод автоматической рубрикации документов, основанный на использовании тематического представления документов и описании рубрик в виде булевских выражений над понятиями лингвистической онтологии. На основе метода реализовано более 20 систем автоматической рубрикации;

- метод автоматического аннотирования отдельного документа и совокупности тематически близких документов на базе выделения из текстов наиболее содержательных предложений. В экспериментах показана высокая связность создаваемых аннотаций в сочетании с не уступающей другим методам полнотой представления информации.

Качество данных алгоритмов было экспериментально проверено в процессе независимой экспертизы в сравнении с другими методами на общественно доступных данных. Программы построения тематического представления текстов, порождения концептуального индекса, автоматической рубрикации и аннотирования объединены в единый программный комплекс тематического анализа текста.

6. Предложена многофакторная модель извлечения терминов предметной области. Реализованный в соответствии с предложенной моделью метод извлечения терминов основывается на вычислении для языковых выражений трех типов статистических характеристик и комбинировании их методами машинного обучения.

Основные результаты диссертации опубликованы в следующих изданиях:
Монография, поддержанная грантом РФФИ:

1. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского университета, 2011.

Публикации в изданиях из перечня ВАК:

1. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. 1995. N 3. С.21-24.
2. Лукашевич Н.В., Салий А.Д. Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер.2. 1996. N 1. С. 1-6.
3. Лукашевич Н.В. Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. 1996. N 10. С. 22-30.
4. Лукашевич Н.В. Салий А.Д., Представление знаний в системе автоматической обработки текстов // НТИ. Сер.2. 1997. N3. С. 1-6.
5. Лукашевич Н.В., Добров Б.В. Модификаторы концептуальных отношений в тезаурусе для автоматического индексирования // НТИ, Сер.2. 2001. N 4. С. 21-28.
6. Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования практической онтологии // Известия РАН. Теория и системы управления. 2004. N 2. С. 58-68.
7. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Ученые записки Казанского Государственного Университета. Серия Физико-математические науки. 2007. т.149. книга 2. С.49-72.

8. Лукашевич Н.В. Моделирование отношения ЧАСТЬ-ЦЕЛОЕ в лингвистическом ресурсе для информационно-поисковых приложений // Информационные технологии. 2007. N12. С. 28-34.
9. Агеев М.С., Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2008. Том 150. книга 4. С. 25-40.
10. Лукашевич Н.В., Логачев Ю.М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. разд. 2. 2010. С. 108-116.
11. Лукашевич Н.В. Понятия в формальных и лингвистических онтологиях // Научно-техническая информация, сер.2. 2011. N 7. С. 1-8.
12. Лукашевич Н.В., Четверкин И.И. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. разд. 2. 2011. С. 73-81.
13. Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. 2011. N 4. С. 95-103.
14. Алексеев А.А., Лукашевич Н.В. Комбинирование признаков для извлечения тематических цепочек в новостном кластере // Труды Института системного программирования РАН. 2012, Т. 23. С. 257-276**.**
15. Лукашевич Н.В. Отношения часть-целое: теория и практика // Нейрокомпьютеры: разработка, применение. 2013. N1. С. 7-12.

Основные публикации, указанные в международных системах цитирования из списка ВАК:

1. Dobrov B., Loukachevitch N., Nevzorova O., Fedunov B. Methods of automated design of application ontology // Journal of Computer and Systems sciences international. 2004. V. 43. I. 2. P. 213-222. (Web of Science)
2. Loukachevitch N., Dobrov B. Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains // Proceedings of Second International WordNet Conference GWC-2004. 2004. P.163-168. (Web of Science)
3. Loukachevitch N., Dobrov B. Large-Scale Linguistic Ontology as a Basis for Text Categorization of Legislative Documents //Legal Knowledge And Information Systems: Jurix 2005, the Eighteenth Annual Conference. IOS Press, 2005. V. 134. P. 109-110. (Web of Science)
4. Ageev M., Dobrov B., Loukachevitch N. Sociopolitical Thesaurus in Concept-based Information Retrieval: Ad-hoc and Domain Specific Tasks // Cross-Language Evaluation Forum. Results of the CLEF 2005 Cross-Language System Evaluation Campaign / Eds.: C.Peters, V.Quochi. Springer Verlag, 2006. LNCS-4022. P. 141-150. (Scopus, Web of Science)
5. Loukachevitch N. Concept Formation in Linguistic Ontologies. Conceptual Structures: Leveraging Semantic Technologies // In Proceedings of ICCS-2009 / Eds Sebasian Rudolph, Frithjof Dau, Sergei O. Kuznetsov. Springer Verlag, 2009. LNAI-5662. P. 2-22. (Scopus)
6. Loukachevitch Natalia. Multigraph representation for lexical chaining // Proceedings of SENSE workshop, 2009. P. 67-76. (Scopus)
7. Loukachevitch N., Dobrov B. Combining Evidence for Automatic Extraction of Terms // In Proc. of 4th International conference on Pattern Recognition and Machine Intelligence, Springer Verlag, 2011. V. 6744. P. 234-240. (Web of Science)

8. Loukachevitch N. Establishment of taxonomic relationships in linguistic ontologies // Knowledge processing and data analysis. Springer Verlag, 2011. LNCS-6581. P.232-242. (Scopus).
9. Dobrov B., Loukachevitch N. Multiple evidence for term extraction in broad domains // International Conference Recent Advances in Natural Language Processing, RANLP-2011, pp. 710-715. (Scopus)
10. Alekseev A.A., Loukachevitch N.V. The automatic retrieval of news entities based on the structure of a news cluster // Scientific and Technical Information Processing, 2012. V 39. N 6. P. 303-309. (Scopus)
11. Alekseev A.,, Loukachevitch N. Use of multiple features for extracting topics from news clusters // Proceedings of SYRCODIS-2012, 2012. P. 3-11. (Scopus)
12. Chetviorkin I. I., Loukachevitch N. V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // Proceedings of COLING-2012, 2012, P. 593–610. (Scopus)
13. Bolshakova E., Loukachevitch N., Nokel M. Topic Models Can Improve Domain Term Extraction // International conference on Information Retrieval ECIR-2013, Springer Verlag, 2013. LNCS-7814, P.684-687. (Scopus)

Основные публикации в других научных изданиях

1. Лукашевич Н.В., Добров Б.В. Построение и использование тематического представления содержания документов // Труды 5ой Национальной конференции КИИ-96. Казань, 1996. С. 130-134.
2. Лукашевич Н.В. Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. Москва, 1997. С. 188-191.
3. Лукашевич Н.В. От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Труды международного семинара Диалог-99, Том 2. 1999. С. 184 -190.
4. Loukachevitch N., Dobrov B. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems // Machine Translation Review. 2000. N 11. p. 10-20.
5. Добров Б.В., Лукашевич Н.В. Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск, 2001. С.78-82.
6. Лукашевич Н.В., Добров Б.В. Автоматическое выявление лексической связности текста // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2001. Вып. 6. Казань: Отечество, 2001. С. 19-38.
7. Loukachevitch N., Dobrov B. Development and Use of Thesaurus of Russian Language RuThes // In Proc. of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation. (LREC2002) / Dimitris N. Christodoulakis. 2002. pp. 65-70.
8. Лукашевич Н.В., Добров Б.В. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2002 / Под ред. А.С.Нариньяни. М.: Наука, 2002. Т.2. С.338-346.
9. Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту КИИ-2002. М.: Физматлит, 2002. Т.1. С.178-186.

10. Лукашевич Н.В., Добров Б.В. Организация тезаурусного поиска в Университетской информационной системе РОССИЯ // Русский язык в Интернете / Под ред. В.Д.Соловьева. Казань: Отечество, 2003. С.84-96.
11. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. 2003. С. 201-210.
12. Loukachevitch N., Dobrov B. Development of Ontologies with Minimal Set of Conceptual Relations // In Proc. of Fourth International Conference on Language Resources and Evaluation / Eds: M.T.Lino et al. 2004. vol. VI. P. 1889-1892.
13. Loukachevitch N., Dobrov B. Development of Bilingual Domain-Specific Ontology for Automatic Conceptual Indexing // In Proc. of Fourth International Conference on Language Resources and Evaluation / Eds: M.T. Lino et al. 2004. vol. VI. P. 1993-1996.
14. Loukachevitch N., Dobrov B. Ontological Types of Association Relations in Information Retrieval Thesauri and Automatic Query Expansion // Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments / Eds: A.Oltramari et al. 2004. P. 24-29.
15. Лукашевич Н.В., Добров Б.В. Взаимодействие лексики и терминологии в общезначимой сфере языка // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог’2004 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. М.: Наука, 2004. С. 172-178.
16. Агеев М.С., Добров Б.В., Лукашевич Н.В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды шестой Всероссийской научной конференции. Пущино, 2004. С. 216-225.
17. Лукашевич Н.В, Добров Б.В. Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. Т2. М.: Физматлит. 2004. С. 544-551.Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: описания понятий и лексических значений. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2006 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. М.: Наука. 2005. С. 138-142.
19. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2010. М.: Физматлит. 2010. С. 489-497.
20. Лукашевич Н.В., Добров Б.В. Разрешение лексической многозначности на основе тезауруса предметной области. Компьютерная лингвистика и интеллектуальные технологии. // Труды международной конференции «Диалог 2007». М.: Наука. 2007. С. 400-406.
21. Лукашевич Н.В. Проблемы установления родовидовых отношений в лингвистических онтологиях // Материалы Всероссийской конференции «Знания-Онтологии-решения» (ЗОНТ-07). 2007. С. 211-220.
22. Лукашевич Н.В. Типы и роли в лингвистических онтологиях // Труды Казанской школы по компьютерной лингвистике TEL-2006. Казань: Отечество, 2007. С. 49-64.
23. Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний // Интернет-математика-2007: Сборник работ участников конкурса. Екатеринбург: Изд-во Урал. ун-та, 2007. С.108-117.

24. Лукашевич Н.В. Описание понятий-ролей в лингвистических и онтологических ресурсах // Материалы Всероссийской конференции RCDL-2007. 2007.
25. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. М.: Изд-во ИНТУИТ, 2008. 176 с.
26. Добров Б.В., Лукашевич Н.В. Транзитивные нетаксономические отношения в онтологическом моделировании // Труды симпозиума Онтологическое моделирование. Институт проблем информатики РАН, 2008. С.229-259.
27. Агеев М.С., Добров Б.В., Лукашевич Н.В., Штернов С.В. УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008.
28. Агеев М.С., Добров Б.В., Красильников П., Лукашевич Н.В, Павлов А., Сидоров А., Штернов С.В. УИС РОССИЯ в РОМИП2007: поиск и классификация // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008.
29. Лукашевич Н. В., Добров Б. В. Автоматическое аннотирование новостного кластера на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии по материалам ежегодной Международной конференции «Диалог 2009». 2009. Вып. 8 (15). С. 299-305.
30. Loukachevitch N. Multigraph representation for lexical chaining // In Proc. of SENSE workshop. 2009. P. 67-76.
31. Лукашевич Н.В., Логачев Ю.М. Использование методов машинного обучения для извлечения слов-терминов // Труды Конференции по искусственному интеллекту, КИИ-2010. 2010.