**Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Санкт-Петербургский государственный университет»**

На правах рукописи

Добров Алексей Владимирович

АВТОМАТИЧЕСКАЯ РУБРИКАЦИЯ НОВОСТНЫХ СООБЩЕНИЙ СРЕДСТВАМИ СИНТАКСИЧЕСКОЙ СЕМАНТИКИ

Специальность: 10.02.21 — прикладная и математическая лингвистика

**Автореферат на соискание ученой степени
кандидата филологических наук**

**Санкт-Петербург
2014**

Работа выполнена на кафедре математической лингвистики федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет»

Научный руководитель: доктор филологических наук, профессор кафедры математической лингвистики Санкт-Петербургского государственного университета
Григорий Яковлевич Мартыненко

Официальные оппоненты: доктор технических наук, профессор, ведущий научный сотрудник Лаборатории информационных систем математических наук (ЛИСМН) Научно-исследовательского вычислительного центра, Московский государственный университет (НИВЦ МГУ)
Нина Николаевна Леонтьева (г. Москва)

кандидат филологических наук, член Гильдии лингвистов-экспертов по документационным и информационным спорам (ГЛЭДИС)
Игорь Вениаминович Жарков
(г. Санкт-Петербург)

Ведущая организация: ФГБОУ ВПО «Сыктывкарский государственный университет».

Защита состоится “ ____” _____ 2014 г. в _____ часов на заседании Совета Д 212.232.23 по защите диссертаций на соискание ученой степени доктора наук, на соискание ученой степени кандидата наук при ФГБОУ ВПО «Санкт-Петербургский государственный университет» по адресу: 199034, г. Санкт-Петербург, Университетская наб., д.11, ауд._____.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета (Санкт-Петербург, Университетская наб., 7/9).

Автореферат разослан «____» _____ 2014 года.

Ученый секретарь диссертационного совета: К.В. Манерова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В диссертационном исследовании рассматривается возможность использования средств автоматического синтаксического и семантического анализа текстов новостных сообщений при решении задачи повышения эффективности их автоматической рубрикации.

В последние десятилетия возрос интерес к созданию эффективных инструментов работы с текстовой информацией, основанных на автоматической обработке текстов — систем информационного поиска, машинного перевода, автоматической рубрикации и классификации текстов, автоматического реферирования, систем фактографического анализа. Этот интерес в значительной степени обусловлен резким увеличением объема текстовой информации в электронной форме, приводящим к необходимости автоматизации различных видов деятельности, относящейся к поиску и структурированию информации, до сих пор выполнявшейся вручную.

Актуальность темы исследования определяется тем, что погрешность статистических моделей, стоящих за применявшимися до сих пор алгоритмами рубрикации, несмотря на постоянные усовершенствования этих моделей, становится все более существенной проблемой. Благодаря развитию сети Интернет, объемы текстовой информации резко возросли, и возникла необходимость в принципиально новых технологиях, обеспечивающих более качественный и точный анализ современного медиатекста, чем существующие средства автоматической рубрикации. Недостаточная разработанность инструментов работы с информационными потоками, прежде всего — с «новостными лентами», приводит к появлению новых научных исследований, направленных на поиск наиболее эффективных и точных методов автоматической обработки медиатекста и на развитие структурно-лингвистических моделей, необходимых для корректной работы этих методов. Востребованность таких методов приводит к росту интереса исследователей к развитию научных лингвистических подходов, основанных на выявлении принципов и объяснении особенностей функционирования языковой системы и позволяющих повысить эффективность автоматической обработки текстов и, в частности, их автоматической рубрикации. Тем не менее, вопрос о возможностях использования таких дополняющих друг друга методов лингвистической обработки текста, как синтаксический и семантический анализ, с целью повышения качества автоматической рубрикации до сих пор не подвергался детальному изучению.

Степень разработанности исследуемой проблемы. Комплексный формально-лингвистический подход активно применялся и применяется в областях машинного перевода (Л.Н. Беляева, М.И. Откупщикова), информационного поиска (И.П. Панков, В.П. Захаров), стилеметрии и атрибуции текстов (О.Н. Гринбаум, Г.Я. Мартыненко, М.А. Марусенко), автоматического реферирования текстов (В.В. Богданов, И.П. Севбо), но до сих пор не было попыток применить этот подход к задаче автоматической рубрикации текстов. В ряде работ группы «УИС Россия» (М.С. Агеев, Б.В. Добров, Н.В. Лукашевич и др.) и некоторых иных исследовательских групп рассматривался вопрос о применении морфологического анализа и лингвистического тезауруса к автоматической рубрикации текстов и было показано значимое повышение ее эффективности при использовании этих инструментов, однако возможности дальнейшего повышения качества автоматической рубрикации путем одновременного использования синтаксического и семантического анализа до сих пор не изучались.

Если в 50-е и 60-е годы XX века интерес к созданию комплексных моделей языка (Ю.Д. Апресян, А.К. Жолковский, И.А. Мельчук, Н. Хомский и др.) был крайне высоким, то позднее, в силу недостаточной эффективности создававшихся

автоматизированных систем и высокой трудоемкости их разработки, наступил период смещения интереса в область более простых статистических (в меньшей степени лингвистических) подходов к решению прикладных задач, связанных с обработкой текстов. Эти подходы не всегда предполагали необходимость даже морфологического анализа текстов, однако именно они позволили создать действовавшие системы автоматической обработки текстов в относительно короткие сроки (например, модель Р. Шенка, системы SHRDLU, LUNAR, LIFER/LADDER и др.). Качество результатов работы этих систем было ограничено возможностями моделей, лежащих в основе их реализаций. Далеко не всегда учитывалась такая важная особенность языковых единиц, как морфологическая, синтаксическая и лексико-семантическая неоднозначность, что часто усугублялось осознанным отказом от учета высокочастотной служебной лексики, крайне важной для грамматики, но несущественной для статистических эвристик. Тем не менее, к 90-м годам XX века подходы к автоматической обработке текста, основанные на полностью статистических моделях, стали господствующими (методы, основанные на N-граммах, методы кластерного анализа, нейронные сети, методы машинного перевода, основанные на механизме «памяти переводов» и др.). Лишь к концу 90-х годов XX века, в связи с широким распространением новых поколений вычислительной техники, появилась возможность создания высокопроизводительных систем автоматической обработки текстов, основанных на структурно-лингвистических подходах: формальных грамматиках, формально-семантических моделях и собственно лингвистических алгоритмах анализа и синтеза текстов на естественных языках. В эти годы активно развиваются компьютерные реализации моделей, созданных еще в 60-е годы XX века — система «ЭТАП», реализовавшая модель «Смысл ↔ Текст» (Ю.Д. Апресян, Л.Л. Иомдин, В.З. Санников, Л.Л. Цинман и др.), различные варианты синтаксических парсеров, основанных на порождающих грамматиках непосредственных составляющих и грамматиках зависимостей (системы ФРАП, ПОЛИТЕКСТ, ДИАЛИНГ, Link Grammar Parser, пакет NLTK и др.). Модель «Смысл ↔ Текст» дала толчок развитию самостоятельных моделей естественного языка (Н.Н. Леонтьева, В.А. Тузов, З.М. Шаляпина и др.) и компьютерных реализаций этих моделей. Кроме того, в конце 90-х — начале 2000-х годов, в условиях высокой популярности идеи «семантической паутины» (Т.Б. Ли), возникло множество новых подходов к моделированию лексической семантики, основанных на компьютерных онтологиях.

Непрерывный рост количества разработанных моделей в различных областях компьютерной лингвистики привел к о собой актуальности изучения методов объективной оценки эффективности создаваемых технологий и сопоставления различных статистических, структурно-лингвистических и комбинированных подходов путем сравнения показателей эффективности результатов работы их компьютерных реализаций. Широкое распространение статистических методов оценки этих показателей позволило выявить ряд объективных проблем в некоторых областях компьютерной лингвистики, в частности, в области автоматической рубрикации (классификации) документов (М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, И.С. Некрестьянов, А.В. Антонов, С.Г. Баглей и др.).

Экспериментальные исследования, направленные на сравнение эффективности возможных способов решения этих проблем, показали наибольшую перспективность с точки зрения эффективности автоматической рубрикации текстов комплексных инженерных (в терминах инженерии знаний) и лингвистических подходов к организации систем автоматической обработки текстов и, следовательно, к описанию и математическому моделированию единиц различных уровней языковой системы. Подходы, основанные только на статистических методах машинного обучения, а также комбинированные подходы, предполагающие приоритет отдельных языковых уровней или отдельных аспектов языковых единиц, на сегодняшний день

характеризуются более низкими показателями качества, чем альтернативные им полностью инженерные подходы.

Одним из таких комплексных подходов, которые предполагают одновременный учет наибольшего количества аспектов языковых явлений, является подход, характерный для Петербургской лингвистической школы и ее последователей в области математической, структурной и прикладной лингвистики (В.Г. Адмони, Л.Н. Беляева, В.В. Богданов, В.Д. Буторов, А.С. Герд, Л.Р. Зиндер, Г.Я. Мартыненко, М.И. Откупщикова, И.П. Панков, Р.Г. Пиотровский, В.Ш. Рубашкин, С.Я. Фитиалов, Г.С. Цейтин и др.). Этот подход предполагает всестороннее рассмотрение языковых единиц при их моделировании, сочетающееся с максимальной детализацией как семантических (когнитивных, концептуальных, онтологических и др.), так и формальных (синтаксических, морфологических, фонологических и др.) аспектов создаваемых лингвистических моделей. Безусловно, исследователи, придерживающиеся данного подхода, часто отмечают центральную роль семантики в языковой системе, поскольку семантика «пронизывает» практически все уровни языка. Вместе с тем, именно поэтому семантике не отводится роль отдельного языкового уровня, а при моделировании языковых единиц детализируются как их семантические свойства, так и формальные, с максимально возможной степенью достоверности. В русле такого подхода выполнено настоящее диссертационное исследование.

Объектом исследования являются методы автоматического синтаксического и семантического анализа текстов новостных сообщений, позволяющие повысить точность и полноту их автоматической рубрикации.

Предмет исследования — способы использования средств автоматического синтаксического и семантического анализа текстов новостных сообщений при повышении эффективности их автоматической рубрикации.

Гипотеза исследования состоит в том, что эффективность автоматической рубрикации текстов, основанной на анализе синтаксической семантики[1], выше, чем эффективность автоматической рубрикации текстов, не учитывающей синтаксическую семантику.

Цель диссертационного исследования состоит в установлении принципов и разработке алгоритмов автоматического синтаксического и семантического анализа и рубрикации текстов новостных сообщений и в определении степени эффективности их автоматической рубрикации, основанной на комплексном лингвистическом анализе. Цель исследования предусматривает решение следующих **задач**:

1. Определить основные принципы математического моделирования языковых единиц в области автоматической обработки текстов, характеризующиеся высокими показателями эффективности применения создаваемых моделей к анализу текстов новостных сообщений.

2. Выявить основные положения структурно-лингвистических подходов к описанию и моделированию синтаксической семантики, применимые к автоматической обработке текстов новостных сообщений и обеспечивающие наиболее высокие показатели эффективности такой обработки.

3. Проанализировать существующие методы автоматической обработки текстов, их автоматической рубрикации, способы оценки их эффективности и основные проблемы, выявляемые при такой оценке.

4. Создать компьютерную модель синтаксиса русского языка, включающую в себя информацию о синтаксических единицах, о структурных отношениях между

[1] Прим. автора: под термином «синтаксическая семантика», согласно определению В.Г. Адмони, в диссертационном исследовании понимается семантика синтаксических структур.

этими единицами, о допустимых способах их линейного развертывания и об их семантических свойствах, достаточную для повышения эффективности автоматической рубрикации новостных сообщений.

5. Разработать алгоритмы морфологического, синтаксического и семантического анализа текстов, необходимые для компьютерного анализа текстов новостных сообщений на уровне синтаксической семантики; реализовать указанные алгоритмы в виде единой компьютерной системы.

6. Разработать систему автоматической рубрикации новостных сообщений, основанную на комплексном лингвистическом анализе текстов на уровне синтаксической семантики, и, в частности, систему образов рубрик (правил отнесения текстов к рубрикам), соответствующих набору рубрик, используемых информационным агентством, опубликовавшим анализирующиеся в исследовании новостные сообщения.

7. Экспериментально апробировать и произвести оценку эффективности автоматической рубрикации текстов, обеспечиваемой разработанной в результате исследования системой.

Научная новизна результатов исследования заключается в следующем:

1. Впервые исследованы возможности одновременного применения компьютерного синтаксического и семантического анализа текстов новостных сообщений к их автоматической рубрикации.

2. Создана инновационная компьютерная модель иерархии синтаксических составляющих русского языка, одновременно отражающая информацию о синтаксических зависимостях, об отношениях между единицами этой иерархии и о системе значений синтаксических составляющих, обеспечивающая возможность производить эффективный автоматический синтаксический анализ текстов новостных сообщений на русском языке.

3. Разработан новый, универсальный алгоритм лингвистического анализа, применимый к языкам различных типов, предполагающий строгое разделение алгоритмического ядра, независимого от языка, и подключаемых языковых модулей.

4. Дано научное обоснование архитектуры системы автоматической обработки текстов, обеспечивающей высокопроизводительный комплексный лингвистический анализ текстов новостных сообщений.

5. Впервые создана система автоматической рубрикации новостных сообщений, основанная на комплексном анализе текстов на уровне синтаксической семантики.

6. Установлена эффективность использования компьютерного синтаксического и семантического анализа текстов новостных сообщений при их автоматической рубрикации.

Теоретическая значимость результатов исследования определяется тем, что в нем:

1. Обоснована целесообразность моделирования синтаксических структур при помощи нестрого бинарных комбинированных структур составляющих и зависимостей с точки зрения соответствия модели языковому материалу; создано математическое исчисление контактных и разрывных составляющих, отражающее возможные степени нарушения проективности и альтернативное модели трансформационной грамматики; выявлены основные принципы моделирования лексической семантики при помощи компьютерных онтологий, необходимые для корректной ав-

томатической обработки текстов.

2. Расширены существующие представления о возможностях моделирования синтаксической семантики в части обоснования необходимости уточнения пропозициональных моделей семантики предложения путем сведения их к сетевому представлению и в части возможностей отражения семантической перспективы при сетевом представлении семантики предложения и высказывания.

3. Дано научное обоснование универсальности метода межуровневого взаимодействия при решении проблемы неоднозначности на различных языковых уровнях при анализе текста.

4. Обоснована целесообразность использования компьютерного синтаксического и семантического анализа текстов для их автоматической рубрикации; конкретизированы научные представления о рубриках, соответствующих медиатопикам, как о предметных областях и концептах компьютерной онтологии, и образах рубрик, используемых при автоматической рубрикации текстов.

Практическая значимость результатов исследования заключается в том, что созданная система автоматической рубрикации новостных сообщений может широко применяться в практической деятельности информационных агентств и новостных порталов и служить средством для существенного ускорения и упрощения работы экспертов, осуществляющих структурирование новостных потоков. Применение созданной системы автоматической рубрикации новостных сообщений при поиске новостей может ускорить и упростить работу пользователя, состоящую в фильтрации результатов поиска информации по конкретной тематике. Созданная система автоматической обработки текстов может применяться в системах машинного перевода, информационного поиска и автоматического реферирования текстов. Кроме того, полученные результаты могут быть использованы в курсах по синтаксической семантике, генеративной грамматике, уровням лингвистического анализа, математическим моделям языка, информационным технологиям, информационному поиску, а также при разработке спецкурсов, посвященных автоматической обработке текстов и автоматической рубрикации текстов.

Методология исследования. Теоретико-методологической основой исследования являются, прежде всего, труды отечественных и зарубежных исследователей в области синтаксической семантики, а также работы, посвященные проблемам автоматической обработки текстов и, в частности, их автоматической рубрикации. Для диссертационного исследования наиболее значимыми являются следующие положения.

1. Наиболее перспективен комплексный подход к изучению синтаксической семантики, основывающийся на приоритете принципов концептоцентрического анализа падежно-ролевого типа при учете онтологического фактора (Н.Д. Арутюнова, В.В. Богданов, Дж. Катц, Ч. Филлмор и др.).

2. Семантика предложения складывается из значений его частей и способа их соединения в соответствии с принципом композициональности (Г. Фреге, Р. Монтегю, Б. Парти и др.).

3. Синтаксическая структура предложения может моделироваться при помощи древовидных структур непосредственных составляющих, содержащих в себе информацию о зависимостях между отдельными частями предложения (Л. Блумфилд, З. Харрис, Н. Хомский и др.); древовидные структуры непосредственных составляющих бинарны (Дж.Б. Йоханессен, Р.С. Кейн, Н. Хомский и др.).

4. Семантика представляет собой сложный целостный объект (А.В. Бондар-

ко), пронизывает практически все уровни языка и тем самым не представляет собой отдельного уровня (А.С. Герд); автоматическая обработка текста должна быть функциональной моделью человеческого понимания этого текста и включать в себя анализ текста на всех уровнях языка, а не только один из видов анализа (А.В. Гладкий, А.К. Жолковский, Н.Н. Леонтьева, И.А. Мельчук и др.)

5. Существующие инженерные подходы к автоматической рубрикации текстов, а также подходы, основанные на машинном обучении, приводят к неразрешимым проблемам, связанным с невозможностью учета в рамках этих подходов полисемии и омонимии лексических единиц и с возникновением так называемых ложных корреляций, обусловленных игнорированием синтаксических связей между лексическими единицами в тексте, что требует разработки и апробации нового подхода к автоматической рубрикации текстов (М.С. Агеев, Б.В. Добров и Н.В. Лукашевич).

Методы исследования выбраны с учетом специфики объекта, языкового материала, целей и задач работы. В работе применяются методы лингвистического анализа языкового материала (метод анализа структур непосредственных составляющих и зависимостей, методы комплексного анализа синтаксической семантики), методы и приемы компьютерной лингвистики и статистические методы оценки и сравнения эффективности автоматической рубрикации текстов (в том числе — разработанный автором настоящего диссертационного исследования экспериментальный метод, основанный на сравнении машинной рубрикации с несколькими эталонами).

Основным **материалом** исследования являются данные эксперимента, позволяющего установить значения показателей эффективности разработанной системы: коллекция новостных сообщений агентства «РИА Новости» объемом 24327 документов, 16450 отнесений 165 испытуемыми 200 случайных текстов из указанной коллекции новостных сообщений к 10 рубрикам, 2000 отнесений текстов к рубрикам, выполненных системой автоматической рубрикации.

На защиту выносятся следующие положения.

1. Анализ синтаксической семантики в существенной мере решает проблемы морфологической, синтаксической и лексической неоднозначности и ложной корреляции, возникающие при использовании методов автоматической рубрикации новостных сообщений, основанных на ключевых словах и сочетаниях слов, при этом эффективность автоматической рубрикации новостных сообщений повышается при использовании средств синтаксического и семантического анализа.

2. Алгоритмы синтаксического и семантического анализа и модели синтаксических структур и их семантики, используемые при автоматической рубрикации текстов, могут быть универсальными и совпадать с аналогичными алгоритмами, используемыми в системах машинного перевода и информационного поиска;

3. Те рубрики новостных сообщений, которые соответствуют конкретным предметным областям, в наибольшей мере характеризуются единообразием оценок их соответствия текстам; семантические отношения между такими рубриками и относящимися к ним значениями языковых единиц имеют детерминированную логическую интерпретацию и потому могут моделироваться при помощи компьютерных онтологий наравне с иными семантическими отношениями; при этом образом рубрики является совокупность всех концептов онтологии, привязанных к предметной области, соответствующей этой рубрике, или ее подобластям.

Достоверность и научная обоснованность теоретических и практических результатов исследования обеспечивается:

1. Использованием материалов как традиционных, так и новейших отечественных и зарубежных фундаментальных исследований в области компьютерной, структурной и прикладной лингвистики.

2. Выбором методов анализа языкового материала, адекватных цели и задачам исследования.

3. Применением математических методов оценки эффективности работы систем автоматической рубрикации; методов математической статистики, в частности — критерия согласия Пирсона для проверки гипотезы о нормальности выборочного распределения и t-критерия Стьюдента для сравнения выборочного среднего с заданным значением для нормально распределенных выборок.

Апробация результатов исследования

Основные положения диссертации были представлены на международных конференциях «Востоковедение и африканистика в диалоге цивилизаций» (Санкт-Петербург, СПбГУ, апрель 2009 г.), «Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы» (Йошкар-Ола, Республика Марий-Эл, апрель 2011 г.), «VI Международная научно-практическая конференция «психолингвистика в современном мире» (Переяслав-Хмельницкий, Украина, октябрь 2011 г.), собраниях научного коллектива лаборатории информационных лингвистических технологий ИЛИ РАН (2004-2008 г.), собраниях научного коллектива лаборатории интеллектуальных систем отдела свободного программного обеспечения инновационного центра Санкт-Петербургского Государственного Университета Телекоммуникаций (2008-2011 г.). По теме диссертации опубликовано 7 работ общим объемом 4,3 п.л., в том числе 3 статьи в научных журналах и изданиях, включенных в перечень рецензируемых научных журналов и изданий для опубликования основных научных результатов диссертаций, рекомендованный ВАК РФ, и одна статья в зарубежном издании. 3 работы опубликованы в материалах международных конференций.

Объем и структура диссертации. Работа состоит из введения, четырех глав, заключения, списка сокращений и условных обозначений, словаря терминов, списка затекстовых ссылок, списка использованной литературы, включающего 210 наименований, в том числе 58 на иностранных языках, и списка иллюстративного материала. К диссертации прилагаются исходные коды программ системы автоматической рубрикации новостных сообщений, основанной на синтаксическом и семантическом анализе текстов (Приложение А), материалы эксперимента (Приложение Б) и расчеты оценки эффективности разработанной системы автоматической рубрикации (Приложение В). Общий объем работы составляет 417 машинописных страниц печатного текста: основное содержание изложено на 250 страницах, 167 страниц занимают Приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **Введении** обосновывается актуальность, научная новизна, теоретическая и практическая значимость работы, предлагаются рекомендации по использованию результатов исследования, определяется его методологичесвкая основа, объект и предмет, цель и задачи, характеризуется материал и методы исследования, формулируются гипотеза и положения, выносимые на защиту.

В **Первой главе** «Проблемы математического моделирования языковых единиц в области автоматической обработки текстов» определяются основные принципы математического моделирования языковых единиц в области автоматической обработки текстов (далее — АОТ), характеризующиеся высокими показателями эффективности применения создаваемых моделей.

Рассматриваются различные понимания задачи АОТ, анализируется предложенное Н.Н. Леонтьевой противопоставление АОТ и автоматического понимания текстов (АПТ) и возникающие в связи с этим противопоставлением проблемы моделирования понимания текстов (И.А. Мельчук, М.И. Откупщикова) и недостатки игнорирования при таком моделировании квантитативных аспектов текста, важных для его понимания (О.Н. Гринбаум,Г.Я. Мартыненко и М.А. Марусенко). Изучаются области применения и различные определения понятий морфологического, синтаксического и семантического анализа, представленные в работах Л.Н. Беляевой, В.П. Гергеля, И.П. Кузнецова, Н.Н. Леонтьевой, И.А. Мельчука, В.В. Окатьева, М.И. Откупщиковой, А.В. Сокирко, Н.В. Сомина и др. Рассматриваются различные подходы исследователей к вопросу о том, где проходит граница между синтаксическим и семантическим анализом текста.

Подчеркивается взаимосвязь областей АОТ и искусственного интеллекта (далее — ИИ) в связи с задачей моделирования знаний. Отмечаются важнейшие идеи, сформулированные в областях машинного перевода (далее — МП),информационного поиска (ИП) и автоматического реферирования (далее — АР). В области МП была обнаружена необходимость создания семантической модели, отражающей значения лексических единиц и их связь с объектами реального мира. Было произведено разграничение морфологического, синтаксического, семантического и прагматического уровней анализа. В области АР была установлена недостаточная эффективность использования статистических моделей для выявления основного содержания текста и необходимость для решения этой задачи учета синтаксиса связного текста. В области ИП была предложена идея использования дескрипторов при индексировании и поиске, в дальнейшем повлиявшая на возникновение информационно-поисковых тезаурусов, а впоследствии — семантических сетей и компьютерных онтологий.

Способы представления знаний в системах АОТ и ИИ зависят от используемых алгоритмов решения задач. В рамках ИИ были предприняты попытки создания универсальных моделей таких алгоритмов — лабиринтная (А. Ньюэлл, Г. Саймон) и реляционная (Д.А. Поспелов, В.Н. Пушкин) модели, однако эти модели не являются общепринятыми и характеризуются недостаточной самообучаемостью. Проблема обучения решается по-разному в зависимости от подхода: при «инженерном» подходе осуществляются попытки решить ее при помощи той или иной системы правил, при «нейробионическом» подходе утверждается, что система ИИ должна моделировать взаимодействие нейронов и синапсов в высших отделах коры головного мозга человека.

Помимо того, что не существует общепринятого способа представления знаний в системах ИИ и АОТ, не имеет однозначного решения и проблема разграничения знаний о мире и о языке. Как отмечает И.П. Панков, для полноценного понимания текста часто требуется множество знаний о мире, внешних по отношению к языку. Формулируется вывод об оптимальности использования компьютерных онтологий для моделирования знаний в системах АОТ.

В данном диссертационном исследовании в качестве основного критерия выбора той или иной математической модели рассматривается ее способность объяснить семантические явления, свойственные моделируемым единицам. С этой точки зрения рассматриваются три основные разновидности математических моделей синтаксических структур — структуры непосредственных составляющих, структуры зависимостей и комбинированные структуры составляющих и зависимостей.

Изучение различных разновидностей структур непосредственных составляющих показывает преимущество бинарных структур по сравнению со структурами, допускающими узлы с локальными степенями, большими двух (небинарными), со-

стоящее в том, что бинарные структуры обеспечивают возможность более точного моделирования иерархических отношений (J.B. Johannessen, R.S. Kayne, R. Larson и др.) и, как следствие, семантических связей. Отмечается неизбежность семантически не нагруженных узлов в нестрого бинарных (допускающих отсутствие второго дочернего узла) структурах, вместе с тем, обнаруживается довольно низкая эффективность практических воплощений (G.M. Kobele) идей минимализма (Н. Хомский), предполагающего строго бинарное ветвление.

В связи с особенностями структур непосредственных составляющих в диссертационном исследовании рассматривается трансформационная порождающая грамматика Н. Хомского. Рассматривается модель деривации синтаксических структур и разделение грамматик непосредственных составляющих на контекстно-свободные и контекстно-зависимые. В ходе анализа предлагаемых Н. Хомским трансформационных правил обнаруживается, что эти правила не всегда соответствуют особенностям языка и часто приводят к порождению некорректных предложений, что соотносится с наблюдающейся в более поздних модификациях теории Н. Хомского тенденцией к устранению трансформаций (см. Я.Г. Тестелец, J, Emonds, H, van Riemsdijk, E. Williams), а в связи с обнаруженными в ходе развития генеративизма языковыми фактами, доказывающими недостаточность трансформаций для объяснения всех возможных порядков слов, — к вытеснению трансформационного компонента так называемым стилистическим, что, однако, не привело к отказу генеративистов от первоначальной аксиоматизации контактности дочерних составляющих при их линейном развертывании. Вместе с тем, многократные попытки построения универсальной модели правил переупорядочивания, способной объяснить особенности порядков слов в различных языках, до сих пор не увенчались успехом, поэтому в реферируемом исследовании принимается предположение, что аксиома о контактности дочерних составляющих и, как следствие, сама идея трансформаций не вполне корректны и, в соответствии с концепциями некоторых исследователей (А.В. Гладкого, И.Б. Долининой и др.), допускается возможность разрывности составляющих. Данное предположение снимает необходимость различения глубинных и поверхностных синтаксических структур, однако требует формализации ограничений, накладываемых на характер разрывности составляющих.

Сточкизренияпервоначальносформулированногосемантическогокритериявыбора математических моделей языковых единиц, среди структур непосредственных составляющихоптимальнымидлязадачреферируемогоисследованияпризнаютсянестрогого бинарные структуры непосредственных составляющих, допускающие разрывность некоторыхединиц.

Структуры зависимостей (Л. Теньер, И.А. Мельчук и др.) содержат информацию о направлении семантических связей, отсутствующую в структурах составляющих. Вместе с тем, в структурах зависимостей теряется информация об иерархии синтаксических и, как следствие, семантических связей. Данная проблема решается в реферируемой работе путем комбинации двух моделей: в структуру составляющих добавляется разграничение ядерных и неядерных дочерних составляющих и, тем самым, задается отношение зависимости. Среди комбинированных структур составляющих и зависимостей рассматривается модель А.В. Гладкого.

Таким образом, в реферируемом исследовании отдается предпочтение комбинированным структурам, являющимся разновидностью нестрого бинарных структур составляющих, допускающих разрывность и содержащих метки дочерних составляющих, позволяющиеразграничиватьглавныеизависимыесоставляющие.

Помимо вопроса о выборе моделей синтаксических структур, в работе формулируется проблема моделирования грамматической семантики. Отмечается, что мно-

гие аспекты грамматического значения, детально описанные А.В. Бондарко и его последователями в теории функциональной грамматики, редко находят отражение в математических моделях. К числу таких аспектов относится неоднозначность категориальных значений (ядерные и периферийные значения, грамматическая полисемия), например, неоднозначность категориального значения родительного падежа. В математических моделях редко производится разграничение грамматических значений по признакам формальности и интенциональности; в частности, редко моделируется механизм выражения иллокутивных функций высказывания грамматическими значениями отдельных словоформ. За редким исключением (Дж. Лакофф), в математических моделях не учитываются несоответствия грамматических и семантических категорий, вызванные межкатегориальным взаимодействием. Формулируются способы представления отмечаемых в теории функциональной грамматики фактов в данной работе, основанные на разграничении классов синтаксических составляющих не только по формальным, но и по функционально-семантическим критериям.

В связи с вопросом о принципах моделирования лексической семантики и полученным ранее выводом об оптимальности использования компьютерных онтологий для представления знаний в системах АОТ и ИИ изучается история появления термина «онтология» в области АОТ, разграничиваются понятия «онтология» и «компьютерный тезаурус», приводится 12 определений понятия «онтология», соответстующих 12 различным ответам на вопрос о том, какие единицы и в каком виде должны храниться в онтологиях. Сравниваются понятия «концепта онтологии» и концепта в общелингвистическом понимании. Рассматриваются различные подходы к моделированию структуры концепта (Е.С. Кубрякова, J.A. Fodor, R. Jackendoff). На основе анализа различных подходов к организации онтологий и моделированию концептов разрабатывается комплексный подход, оптимальный, как представляется, с точки зрения задач АОТ.

Проведенный в реферируемом исследовании анализ различных подходов к построению математических моделей языковых единиц, необходимых для АОТ, позволил выработать комплексный подход, дающий возможность выявить оптимальные способы построения комбинированных моделей как синтаксических структур, так и грамматических и лексических значений.

Во **Второй главе** «Теории синтаксической семантики и их значение для автоматической обработки текста» рассматриваются различные теории, моделирующие синтаксическую семантику, и степень их применимости к задачам АОТ.

Вслед за В.В. Богдановым выделяется 5 групп подходов к моделированию семантики предложения: онтологоцентрический, концептоцентрический, синтактикоцентрический, антропоцентрический и комплексный подход.

Онтологоцентрический подход (П. Адамец, Е.В. Падучева и др.) предполагает моделирование семантики предложения через экстралингвистическую ситуацию или референцию к такой ситуации. При онтологоцентрическом подходе представляется затруднительным математическое моделирование механизма порождения высказывания, поскольку этот механизм должен включать в себя осмысление говорящим выражаемой экстралингвистической ситуации, иными словами — механизм понимания.

Концептоцентрический подход предполагает моделирование семантики предложения не через экстралингвистичесакую ситуацию, а через концептуальную структуру, обусловленную языковыми факторами. В рамках концептоцентрического подхода выделяется пропозициональное направление, в котором под концептуальной структурой понимается пропозиция, т.е. некая формула предикатно-аргументной структуры. В теориях генеративной семантики, основанных на декомпозиции предикатов и нумерации аргументов (Дж. Лакофф, Дж. Макколи и др.), пред-

полагается, что более сложные предикаты могут выражаться через более простые и что семантические отношения между аргументом и предикатом могут быть определены по номеру аргумента. Общие правила интерпретации номеров, однако, не формулируются и вряд могут быть сформулированы.

В теориях синтаксической семантики (Ю.Д. Апресян, У. Чейф, Ч. Филлмор, Р. Шенк и др.), основанных на теории падежных ролей (Г. Суит, А. де Гроот, Э. Бенвенист, Л. Ельмслев, Р. Якобсон, Е. Курилович и др.), аргументы не нумеруются, а помечаются падежными ролями. Падежные роли не нуждаются в интерпретации, но их количество, инвентарь и иерархия не являются общепринятыми. В диссертации подробно анализируются критерии, в соответствии с которыми различные исследователи выделяют падежные роли, особенности перераспределения падежных ролей при деривации предикатов, кумуляции аргументов, изменении перспективы.

В рамках концептоцентрического подхода также выделяется непропозициональное направление, в рамках которого семантика предложения рассматривается через призму отношений между субъектом и предикатом в терминах Аристотелевой логики (Н.Д. Арутюнова) или синтаксического инварианта (базового С-показателя, Дж. Катц).

Синтактикоцентрический подход (Н.Ю. Шведова) рассматривает семантику предложения через призму структурных схем. Механизм построения семантической структуры предложения из структурной схемы и лексических значений заполнивших ее слов не формализован и потому представляется не вполне ясным. Основным недостатком выбора структурных схем как отправной точки при моделирования семантики предложения видится неоднозачность структурных схем: при различном лексическом наполнении одна и та же структурная схема может иметь различные значения.

Основная идея антропоцентрического подхода (З. Вендлер, П. Грайс, Л. Линский, Дж. Остин, Дж. Р. Сёрл, П.Ф. Стросон и др.) состоит в том, что семантика предложения не может быть удовлетворительно описана без учета особенностей конкретного речевого акта, в котором оно было употреблено в виде высказывания. Наиболее значимым для моделирования семантики предложения является фактор говорящего: в любом предложении проявляется коммуникативная интенция говорящего, выражающаяся в иллокутивной функции. Одним из наиболее существенных наблюдений в рамках антропоцентрического подхода для моделирования семантики предложения является принцип кооперации и следующие из него коммуникативные импликатуры (Г.П. Грайс). Импликатуры включаются в семантику предложения и влияют на его интерпретацию, в некоторых случаях меняя ее на противоположную. Иллокутивные функции и импликатуры возникают при помещении предложения в прагматический контекст, т.е. при превращении предложения в высказывание. Таким образом, антропоцентрический подход позволяет объяснить отличия семантики высказывания от семантики предложения.

Комплексные подходы к моделированию семантики предложения основаны на комбинировании приемов остальных подходов. Так, подход В.В. Богданова основан на совмещении концептоцентрического подхода падежно-ролевого типа с антропоцентрическим (в части прагматической рамки), онтологоцентрическим (в части референциальных статусов пропозиций) и синтактикоцентрическим (в части модификаторов и коннекторов, связывающих пропозиции между собой). В реферируемом исследовании анализируется модель В.В. Богданова и предлагается альтернативная версия комплексного подхода, специально разработанная для задач АОТ. В рамках предлагаемого подхода концептуальная структура имеет не пропозициональное, а сетевое представление в виде концептуального графа (J. Sowa). Сетевое представление

семантики предложения использовалось И.А. Мельчуком, Ю.Д Апресяном и др. в модели «Смысл ↔ Текст», в диссертации анализируются особенности этого семантического представления и его связь с синтаксическим представлением. Формулируется вывод о том, что данная модель не может быть напрямую использована в разрабатываемой системе АОТ.

При моделировании семантики высказывания неизбежно возникает проблема учета его контекста.В диссертации рассматриваются различные виды анафорических отношений; вслед за Е.В. Падучевой выделяются кореферентные типы (субстанциальное и концептуальное тождество, уподобление, различение и распределение), анализируются некоторые не учтенные Е.В. Падучевой типы анафорических отношений, семантически сходные с отношениями, выражаемыми конструкциями с родительным падежом, обнаруженные в ходе исследования.

Учет анафорических отношений позволяет отличать «новое» от «старого», тем самым в некоторой степени производя актуальное членение высказывания. Рассматриваются различные подходы к выполнению данной процедуры; вслед за В.Б. Касевичем делается вывод о том, что одно и то же предложение может получить различные интерпретации с точки зрения актуального членения, но только в различных контекстах или в отрыве от контекста. С точки зрения АОТ, данное свойство актуального членения представляет собой проблему: чтобы произвести актуальное членение, необходима информация о контексте, но для корректности информации о контексте необходимо его актуальное членение. Проблема усугубляется также тем, что коммуникативное структурирование текста в существенной мере зависит от его коммуникативного стиля (Е.В. Ягунова), выявление которого также требует в некоторых случаях актуального членения.

Рассматривается модель тематических прогрессий (Ф. Данеш), основанная на различных комбинациях анафорических отношений и вариантов актуального членения. Отмечается, что 5 видов тематических прогрессий, выделенных Ф. Данешем, можно свести к двум основным типам — цепному и кустовому. Различные исследования (К.А. Филиппов, А.С. Штерн) показывают, что прогрессии кустового типа гораздо легче для восприятия, чем прогрессии цепного типа. Данный факт согласуется с идеей тематической доминанты, рассматриваемой в следующей главе.

Структуры, сходные с тематическими прогрессиями, использовались в области АР (И.П. Севбо). Цепочки анафорических отношений (в терминах И.П. Севбо — нанизывания) позволяют вычленить в тексте наиболее значимые компоненты. С точки зрения И.П. Севбо, нанизывания являются синтаксическими отношениями, а тот факт, что они связывают между собой элементы разных предложений, говорит о том, что необходимо расширить понятие «синтаксис» и включить в него раздел «синтаксис связного текста». Такой подход позволяет рассматривать анафорические отношения как разновидность синтаксических и интерпретировать их методами, сходными с методами, применяемыми при моделировании семантики синтаксических связей. В процессе построения математической модели системы АР И.П. Севбо предполагает, что конструкции с однородностью семантически эквивалентны цепочкам предложений, и предлагает конкретные алгоритмы разбиения однородных рядов. При этом обнаруживаются некоторые нюансы, противоречащие идее правомерности такого преобразования. И.П. Севбо разрабатывает алгоритмы, компенсирующие возникающие при таком разбиении побочные эффекты, однако в реферируемом исследовании показано, что эти алгоримы предотвращают далеко не все возможные последствия.

Таким образом, анализ различных теорий, моделирующих синтаксическую семантику, показал, что

а) с точки зрения точности и полноты моделирования, оптимальным представлением семантики высказывания является концептоцентрическое сетевое представление, включающее в себя информацию об относительном референциальном статусе входящих в него пропозиций и прагматическую рамку;

б) семантические представления разных высказываний одного и того же текста могут быть взаимосвязаны благодаря анафорическим отношениям, содержательно близким к синтаксическим;

в) значимость компонента текста зависит от количества кореферентных ему иных компонентов этого текста.

В **Третьей главе** «Методы автоматической обработки текста и способы оценки их эффективности» рассматриваются различные методы и алгоритмы АОТ, в том числе — автоматической рубрикации текстов, и способы оценки эффективности работы этих методов и алгоритмов.

При рассмотрении методов морфологического анализа (далее — МА) анализируются проблемы морфологической неоднозначности, многообразия возможных форм и отсутствия общепринятого инвентаря грамматических категорий (С.А. Коваль). Рассматриваются конкретные системы МА (Диалинг, mystem и Starling).Декларативные методы МА предполагают наличие специальной базы данных, содержащей в себе заранее указанные результаты разбора всех возможных словоформ; системы, основанные на декларативных методах МА, ограничены объемами этой базы. Процедурные методы МА, напротив, не предполагают хранения словоформ в базе данных и основаны на разбиении словоформ на известные системе основы или корни и аффиксы. Процедурные методы могут характеризоваться меньшей производительностью, чем декларативные, однако, в отличие от них, теоретически способны к распознаванию словоформ неизвестных системе лексических единиц. Построение компьютерных моделей морфологии и морфонологии, необходимых для процедурного МА, может быть весьма трудоемким, поскольку число нерегулярностей в морфемике и словообразовании весьма велико и существенно превышает число нерегулярностей в формообразовании и словоизменении.

В реферируемой работе рассматриваются некоторые методы снятия морфологической неоднозначности — методы, основанные на межуровневом взаимодействии, методы, основанные на правилах, статистические методы, основанные на марковских цепях, и комбинированные методы, основанные на правилах и на марковских цепях. В системах, выполняющих не только МА, но и все остальные виды анализа на более высоких уровнях, оптимальными являются методы межуровневого взаимодействия, поэтому в данной работе используются именно эти методы.

В работе приводятся способы оценки эффективности морфологического анализа, основанные на стандартных мерах точности, полноты и F-меры.

При рассмотрении методов синтаксического анализа (далее — СА) анализируются проблемы синтаксической неоднозначности, проблемы восстановления эллипсиса и проблемы разрешения анафорических связей. Синтаксическая неоднозначность может быть обусловлена как морфологической неоднозначностью, так и регулярными свойствами грамматики, допускающими несколько интерпретаций одной и той же цепочки грамматических форм. В последнем случае чаще всего причина неоднозначности состоит в возможности восстановления эллипсиса. Проблемы восстановления эллипсиса рассматриваются отдельно, поскольку эллиптированные элементы представляют собой либо анафоры, либо отсылки к общему фону знаний, причем характер этих отсылок зависит от лексики. Сложности, связанные с восстановлением анафорических отношений, также рассматриваются наряду с прочими

проблемами синтаксического анализа, хотя в работе демонстрируется, что при выбранной в данной диссертации модели синтаксической структуры более целесообразно относить восстановление анафор к семантическому анализу.

В зависимости от подхода к решению проблемы синтаксической неоднозначности, методы СА можно разделить на одноцелевые, многоцелевые и комбинированные. Одноцелевые методы СА предполагают, что на выходе СА всегда должна быть одна гипотеза разбора, т.е. снятие неоднозначности осуществляется теми или иными способами в процессе самого СА. Многоцелевые методы предполагают, что на выходе СА должны быть все возможные гипотезы разбора, а снятие неоднозначности осуществляется за рамками СА. К комбинированным методам СА относятся различные варианты метода фильтров (И. Лесерф, Л.Н. Иорданская, Ю.Д. Апресян, В.В. Окатьев, В.П. Гергель, В.А. Галактионов, А.М. Мусатов). Метод фильтров состоит в том, что в процессе СА строятся все гипотезы синтаксического разбора, однако в процессе разбора применяются те или иные фильтры, позволяющие блокировать некорректные гипотезы. В реферируемой работе для снятия неоднозначности используются методы межуровневого взаимодействия, т.е. в качестве фильтра в процессе СА выступает семантический анализ.

Методы СА могут быть направлены на построение структуры зависимостей, структуры составляющих или комбинированной структуры. В реферируемой работе анализируются особенности конкретных систем, реализующих методы каждого типа. Методы ранжирования гипотез в процессе СА могут быть формально-грамматическими и вероятностно-статистическими. Поскольку методы обоих типов редко дают положительный результат, в данной работе не используется ранжирование версий СА, все гипотезы СА считаются равновероятными.

В диссертации анализируются различные методы снятия синтаксической неоднозначности. О.В. Митренина выделяет 8 типов методов, которые в данной работе рассматриваются как подтипы трех основных видов методов снятия синтаксической неоднозначности: 1) методы, основанные на семантических ограничениях, 2) методы, основанные на грамматических ограничениях, 3) статистико-вероятностные методы. В реферируемой работе используются методы первых двух видов.

В области семантического анализа в данной работе анализируются проблемы лексической неоднозначности и проблемы, связанные с отсутствием необходимой информации в компьютерных онтологиях. Производится анализ четырех опубликованных ресурсов, содержащих различную семантическую информацию, используемых в АОТ — онтологий OpenCyc и SUMO, тезауруса WordNet и семантической сети «FrameNet».

В рамках анализа существующих методов автоматической рубрикации (классификации) текстов производится разграничение методов автоматической рубрикации, предполагающих наличие заданного множества рубрик, и методов кластерного анализа, направленного на автоматическое выделение кластеров; реферируемая работа посвящена проблеме автоматической рубрикации. Методы автоматической рубрикации текстов подразделяются на «инженерные» методы и методы, основанные на машинном обучении. Инженерные методы предполагают, что образы рубрик создаются вручную разработчиками системы автоматической рубрикации; методы, основанные на машинном обучении, предполагают, что рубрицирование производится путем предварительного «обучения» статистического алгоритма на отрубрицированной коллекции текстов. Демонстрируется фактическое преимущество «инженерных» методов.

Качество автоматической рубрикации текстов определяется тем, насколько успешно решаются проблемы морфологической и лексической неоднозначности и

проблемы ложной корреляции (например, отнесение к рубрике «Экономическая реформа» текста, содержащего в себе слова «экономический» и «реформа», между которыми в тексте отсутствует семантическая связь). Задача автоматической рубрикации рассматривается как частный случай задачи информационного поиска: образ рубрики представляет собой аналог сложного поискового запроса. Обосновывается применение средств АОТ, используемых в информационном поиске, к задаче автоматической рубрикации. Приводятся общепринятые методы оценки эффективности информационного поиска и автоматической рубрикации по мерам точности и полноты. В случае автоматической рубрикации оценка этих мер осуществляется путем сравнения выдачи системы автоматической рубрикации с эталонной рубрикацией.С точки зрения этих показателей анализируются различные существующие системы автоматической рубрикации. Формулируется вывод о том, что наилучшие результаты демонстрируют системы, основанные на инженерных подходах, но эти результаты, как правило, достигают не более 75% F-меры. Кроме того, анализируются проблемы, связанные с самой методикой расчета F-меры: эталонная рубрикация — это результат деятельности одного человека, однако отнесение текста к рубрике человеком может быть субъективным. Предлагается более сложный математический аппарат для оценки эффективности автоматической рубрикации, основанной на сравнении результатов работы машины с ответами нескольких человек.

Рассматриваются особенности новостных сообщений как материала для оценки эффективности автоматической рубрикации средствами синтаксической семантики. Анализируются общие свойства медиатекстов, изучаемые медиалингвистикой (Т.Г. Добросклонская, В.Г. Костомаров, Н.Н. Кохтев, Ю.В. Рождественский, Д.Э. Розенталь, Г.Я. Солганик, Д.Н. Шмелёв, Т. Ван Дейк, М. Монтгомери, А. Белл, Н. Фейерклаф, Р.-Фаулер), анализируется классификация медиатекстов. Новостные тексты рассматриваются как разновидность медиатекстов, характеризующаяся официально-деловым или публицистическим функциональным стилем и доминированием информационной функции сообщения. Новостные сообщения рассматриваются как самостоятельные новостные тексты, в наибольшей степени относящиеся к обиходно-деловому функциональному стилю и реализующие, в общем случае, только информационную функцию. В соответствии с концепцией Т. Ван Дейка, анализируется соотношение тематической структуры дискурса и структуры релевантности новостных текстов. Демонстрируется меньшая степень вероятности расхождений между этими двумя структурами в случае новостных сообщений по сравнению с другими разновидностями новостных текстов. Производится сравнение новостных сообщений и юридических документов с точки зрения возможности их использования в качестве материала для оценки эффективности автоматической рубрикации средствами синтаксической семантики.

Таким образом, анализ различных методов и алгоритмов АОТ и способов оценки эффективности работы этих методов и алгоритмов показал, что:

1. С учетом специфики поставленных задач, оптимальным для данной работы методом снятия морфологической неоднозначности является метод межуровневого взаимодействия, а оптимальным методом СА — метод фильтров (в роли фильтра выступает семантический анализ), направленный на построение комбинированных структур составляющих и зависимостей, осуществляющий снятие синтаксической неоднозначности при помощи семантических и грамматических ограничений.

2. Для решения поставленной в данной работе задачи семантического анализа опубликованных онтологий недостаточно, поскольку существующие онтологии не содержат в себе необходимой информации о семантических ограничениях.

3. Методика оценки эффективности системы автоматической рубрикации должна быть основана на сравнении результатов автоматической рубрикации од-

новременно с несколькими «эталонными» рубрикациями одной и той же коллекции документов.

4. Для задачи оценки эффективности автоматической рубрикации средствами синтаксической семантики оптимальным материалом являются новостные сообщения.

В **Четвертой главе** «Система автоматической рубрикации новостных сообщений средствами синтаксического и семантического анализа текстов» описывается разработанные в рамках данного исследования компьютерная модель синтаксиса русского языка, алгоритмы морфологического, синтаксического и семантического анализа текстов и созданная на базе этих алгоритмов система АОТ; система, выполняющая автоматическую рубрикацию новостных сообщений путем комплексного лингвистического анализа этих документов на уровне синтаксической семантики, а также результаты эксперимента, направленного на оценку эффективности созданной системы.

Анализируются возможности применения комплексного лингвистического анализа текста при его автоматической рубрикации. Предлагаются способы решения проблем морфологической и лексической неоднозначности и ложной корреляции при помощи комплексного лингвистического анализа текстов и указания в образах рубрик понятий, а не ключевых слов.

Объясняется универсальный принцип работы созданного в рамках данного исследования лингвистического процессора. В связи с этим анализируются проблемы морфологического анализа узуальных и окказиональных словоформ текста, описываются алгоритмы анализа деривационных отношений в морфологии, словообразовании и синтаксисе, формулируется универсальный алгоритм распознавания комбинированных структур составляющих и зависимостей. Описывается алгоритм семантического анализа.

Рассматриваются основные модули лингвистического процессора: центральный модуль (tproc), модуль кеша атомарных единиц (cache), модуль обработки цифровых последовательностей (digits), модуль обработки последовательностей пунктуационных символов (punctuation), модуль грамматики (grammar), модуль онтологии (ontology), модуль привязки кеша атомарных единиц к онтологии (onto_cache), модуль планирования (agenda), модуль языковых единиц (signal), модуль классов языковых единиц (sigclass), модуль концептов и концептуальных графов (concgraph), модуль онтологии (ontology), модуль концептуального связывания (concrouting).

Приводится разработанная система грамматических категорий, включающая в себя как общепринятые морфологические категории, так и служебные категории, необходимые для анализа пунктуации и чисел. В созданной системе грамматических категорий некоторые компоненты имеют нестандартный вид, в частности, в связи с особенностями соотношений категории вида и лица русского глагола из системы его категорий исключается категория времени; исторически перфектные формы прошедшего времени трактуются как причастные, что в значительной мере упрощает моделирование взаимодействия грамматических категорий как на уровне морфологии, так и на уровне синтаксиса.

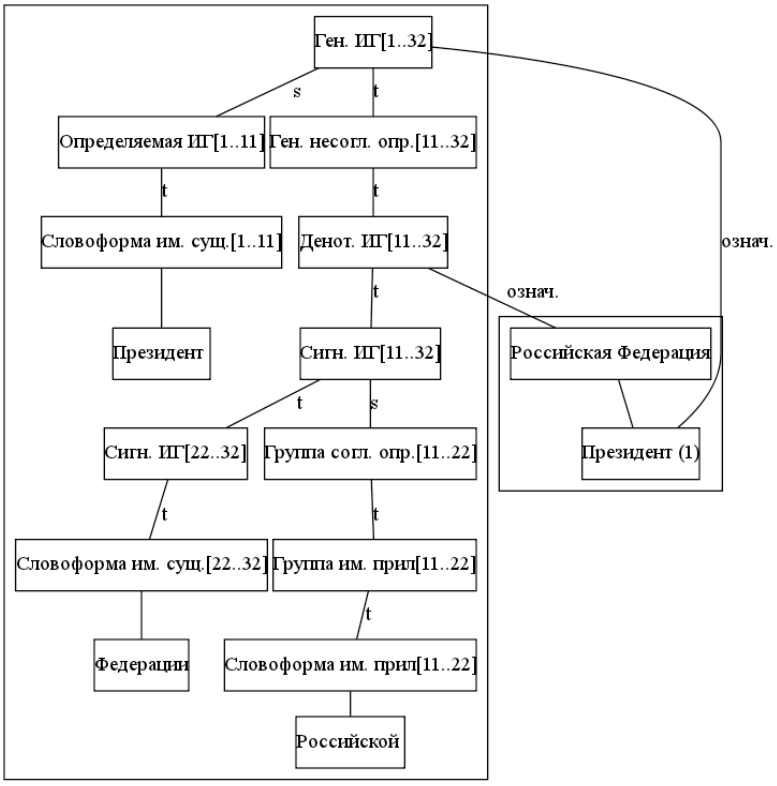
Описывается созданный морфологический словарь, содержащий 121636 лемм. Приводятся использованные источники, указываются конкретные компоненты словаря, созданные в процессе данного исследования без обращения к внешним источникам.

Рассматривается разработанная в рамках данного исследования комбиниро-

ванная грамматика непосредственных составляющих и зависимостей (321 класс непосредственных составляющих). Формулируются принципы моделирования классов непосредственных составляющих (далее — КНС) и их семантических, грамматических, структурных и линейных свойств. Приводится разработанная в рамках данного исследования математическая модель линеаризации разрывных составляющих, обосновывается возможность однозначного числового отражения степени нарушения проективности; указывается, что в рамках данного исследования не было обнаружено случаев нарушения проективности, имеющих степень, превышающую 2, что дает возможность свести все многообразие линейных порядков к шести типам. При описании способов отражения семантических свойств КНС производится разграничение модификационных и аргументных семантических отношений. Модификационные отношения далее подразделяются на функциональные и атрибутивные. Среди функциональных отношений отдельно выделяются отношения, предполагающие операцию семантического запроса. Указанные разграничения позволяют выделить шесть базовых классов КНС. Обнаруживается устойчивая взаимосвязь между семантическим типом КНС и его формально-грамматическими свойствами. Подробно анализируются основные КНС, созданные в рамках данного исследования. Пример соотношения между синтаксической структурой и ее значением приводится на рисунке 1[2].

§

Рисунок 1: Структура составляющих и соответствующие ей концепты



Описывается используемая для целей данного исследования онтология (19287 концептов). Эта онтология разрабатывается коллективом разработчиков (лингви-

[2] Прим. автора: В левой части рисунка 1 изображена нестрого бинарная структура составляющих, отражающая информацию о зависимостях в метках дуг (метка «t» соответствует главной составляющей, метка «s» – зависимой). В метках узлов в сокращенной форме указаны наименования КНС. В квадратных скобках указаны линейные границы составляющих. В правой части изображено семантическое представление, получаемое в результате обработки синтаксической структуры. Концепт компьютерной онтологии 'Российская Федерация' получен напрямую из онтологии, в то время как связь между этим концептом и концептом 'Президент (1)' установлена благодаря синтаксической структуре.

стов и программистов) под руководством автора реферируемой диссертации в соответствии в принципами, сформулированными в первой главе работы. Поскольку автор не является единственным разработчиком онтологии, а цель исследования не предполагает ее создание, в рамках реферируемой работы дается лишь функциональное описание созданной онтологии.

Рассматриваетсясозданнаяутилитаавтоматическойрубрикации(aire_classifier)и алгоритм ее работы, основанный на принципах, сформулированных в предыдущих разделахдиссертации.

Описывается эксперимент, проведенный с целью оценки эффективности разработанной утилиты автоматической рубрикации. Эксперимент состоял в сравнении результатов работы утилиты «aire_classifier» с несколькими эталонами на материале репрезентативной выборки (200 новостных сообщений), взятой из коллекции объемом 24327 текстов, загруженной с веб-ресурса агентства «РИА Новости» с сохранением той рубрикации, которая была указана на этом веб-ресурсе. Изначально тексты были привязаны к 10 рубрикам: «В мире», «Наука», «Спорт», «Главное», «Культура», «Общество», «Политика», «Экономика», «Безопасность», «Происшествия». В коллекцию вошло 20 текстов из каждой рубрики. Перед проведением эксперимента в онтологии лингвопроцессора для указанных рубрик были составлены образы, схема построения которых подробно изложена в диссертационном исследовании. В эксперименте приняли участие 165 экспертов, специализирующихся в различных предметных областях: 13 — в области математической лингвистики, 20 — в области радиофизики, 16 — в области социологии, 20 — в области педагогики, 27 — в области логопедии, 11 — в области романской филологии, 18 — в области финно-угроведения и 40 — в области онтолингвистики. Каждому испытуемому предлагалось выполнить рубрицирование 10 случайно выбранных текстов из коллекции, таким образом, чтобы каждый из 200 текстов был обработан не менее, чем пятью людьми. В среднем каждый текст обрабатывался 8 различными испытуемыми, максимальное количество испытуемых на текст составило 13 человек. Каждый испытуемый заполнял анкету, в которой требовалось указать пол и возраст испытуемого, после чего оценить по шкале от нуля до ста степень соответствия каждого из 10 текстов каждой из 10 рубрик. В случае отсутствия соответствия требовалось в явном виде указывать нулевую степень соответствия. К сожалению, незначительная часть испытуемых все же выполнила рубрицирование не до конца, поэтому 5 текстов получили меньшее количество привязок, чем предполагалось изначально. Всего в ходе эксперимента было выполнено 16450 привязок документов к рубрикам с указанием степени соответствия документа рубрике.

После обработки анкет в роли еще одного «испытуемого» выступила разработанная система автоматической рубрикации. Расчеты F-меры производились в соответствии с предложенной методикой оценки F-меры, основанной на t-критерии Стьюдента для сравнения выборочного среднего с заданным значением: сравнивается выборка оценок испытуемыми степени соответствия одного текста одной рубрике с аналогичной оценкой созданной системы. Поскольку данный статистический критерий основан на предположении о нормальности выборочного распределения, выборки подвергались предварительной фильтрации по критерию согласия Пирсона с параметризованным нормальным распределением. В соответствии с этим критерием 95 выборок были исключены из дальнейшего рассмотрения. Кроме того, 445 выборок имели нулевую дисперсию, поэтому их средние сравнивались с реакциями созданной системы по строгому арифметическому критерию. Таким образом, из 2000 выборок по t-критерию Стьюдента с реакциями системы сравнивались только 1460. Эти же выборки, в соответствии с требованиями предлагаемой методики, сравнивались с нулем. В зависимости от величин, выдаваемых созданной системой и экспертами, и результатов проверки гипотез по t-критерию Стьюдента, выявлялись случаи истинных положи-

тельных, ложных положительных и ложных отрицательных срабатываний, на основании которых и производился расчет точности, полноты и F-меры. Средняя величина F-меры составила 87%, что существенно превышает опубликованные результаты для аналогичных систем (не более 75%).

Таким образом, данные экспериментального исследования показывают, что предлагаемый комплексный лингвистический подход к автоматической рубрикации не только эффективен,, но и более эффективен, чем подходы, не предполагающие синтаксического и семантического анализа.

ЗАКЛЮЧЕНИЕ

Исследование возможностей использования средств автоматического синтаксического и семантического анализа текстов новостных сообщений при решении задачи повышения эффективности их автоматической рубрикации позволило сделать следующие выводы:

1. Выбор комбинированных моделей синтаксических структур, включающих в себя сведения не только об иерархии непосредственных составляющих, но и о синтаксических зависимостях между ними, обеспечивает наиболее высокую эффективность семантической интерпретации результатов синтаксического анализа. Иерархия непосредственных составляющих определяет порядок выполнения операций семантического связывания, в то время как информация о зависимостях используется при определении направлений семантических отношений.

2. С точки зрения корректности автоматического разрешения неоднозначности и соответствия выявляемых семантических структур особенностям контекста и существующим языковым правилам, одним из наиболее эффективных способов моделирования лексической семантики является ее представление в виде единиц компьютерной онтологии, отражающей не только часто регистрируемые в тезаурусах семантические отношения между значениями лексических единиц, но и отношения между соответствующими этим значениям понятиями, необходимые для моделирования семантических валентностей.

3. Особенности линейного развертывания синтаксических структур в виде предложений, характерные как для новостных сообщений, так и для иных видов текстов, могут однозначно моделироваться путем разделения линейных порядков дочерних составляющих на контактные и разрывные с указанием степени нарушения проективности и не требуют использования механизма синтаксических трансформаций, часто существенно усложняющего процесс синтаксического анализа и структуру грамматики, а в некоторых случаях понижающего степень ее соответствия языковому материалу.

4. Оптимальным с точки зрения соответствия языковому материалу представлением семантики предложения является комплексное сетевое представление, отражающее взаимосвязи между концептами, соответствующие пропозициям, а также отношения между этими взаимосвязями (коннекторы, относительные референциальные статусы). Данное представление может отражать особенности перспективы (по Ч. Филлмору), если оно сопровождается указанием центрального концепта (точки отсчета), характеризующегося достижимостью всех остальных концептов, входящих в семантическое представление. Информация, необходимая для преобразования семантики предложения в семантику высказывания (прагматическая рамка, анафорические отношения, импликатуры), также может моделироваться при помощи вершин и ребер направленного ациклического графа, соответствующего указанному семантическому представлению. Такие модели семантической структуры предложения могут использоваться не только при совершенствовании алгоритмов автоматической рубрикации новостных

сообщений, но и в более общих задачах информационного поиска, машинного перевода, автоматического реферирования и т. д.

5. Для эффективного автоматического семантического анализа текста необходимо учитывать все возможные варианты синтаксического анализа каждого предложения, поэтому при применении комплексного подхода к автоматической обработке текстов необходимо использовать многоцелевые методы синтаксического анализа. С точки зрения производительности, наибольшей эффективностью характеризуется метод межуровневого взаимодействия, предполагающий одновременное выполнение анализа на различных языковых уровнях и использование данных вышестоящих уровней на более низких уровнях для фильтрации некорректных интерпретаций. Эффективность семантического анализа текста зависит от качества морфологического и синтаксического анализа и полноты используемой компьютерной онтологии. Система автоматической рубрикации новостных сообщений, основанная на таком семантическом анализе текстов, может быть встроена в структуру информационно-поисковых систем и систем машинного перевода.

6. Методы оценки точности и полноты автоматической рубрикации текстов, основанные на строгом разграничении релевантных и нерелевантных тексту рубрик, не позволяют учитывать вариативность выделения рубрик различными экспертами и многочисленные случаи неполного соответствия рубрики тексту. Поэтому формулы для расчета точности и полноты автоматической рубрикации целесообразно уточнить за счет введения шкалируемого показателя степени соответствия рубрики тексту (веса рубрики) и сопоставления этого показателя с автоматически вычисленным весом по Т-критерию Стьюдента для сравнения выборочного среднего с заданным значением.

7. Эффективность автоматической рубрикации новостных сообщений, выполняемой существующими системами, оценивается в пределах 75 процентов F-меры. Существенная часть ошибок рубрикации обусловлена игнорированием реализующихся в тексте синтаксических и семантических связей и ограничений, требующих компьютерного моделирования.

8. Разработанная система автоматической рубрикации новостных сообщений средствами синтаксической семантики по данным проведенного эксперимента и уточненным формулам оценки F-меры характеризуется эффективностью 87%[3]. Обнаруженные различия между оценками испытуемых и автоматической рубрикацией обусловлены как неполнотой используемого морфологического словаря и компьютерной онтологии, так и ошибками испытуемых. Случаев расхождений между оценками испытуемых и результатами автоматической рубрикации новостных сообщений, обусловленных неоднозначностью языковых единиц или ложной корреляцией, выявлено не было.

Таким образом, была достигнута цель исследования и решены его основные задачи. Гипотеза о том, что эффективность автоматической рубрикации текстов, основанной на анализе синтаксической семантики, выше, чем эффективность автоматической рубрикации текстов, не учитывающей синтаксическую семантику, была подтверждена.

По теме диссертации опубликованы следующие работы:

1.*Добров, А.В.* Опыт создания арабско-русского электронного словаря и системы поддержки перевода [текст] / Т.А. Рыженкова, А.В. Добров // "Востоковедение и африканистика в диалоге цивилизаций": XXV международная конференция по источниковедению и историографии стран Азии и Африки, 22-24 апреля 2009 г.:

[3] Прим. автора: Расчет меры эффективности (F-меры) приведен в Приложении 3 диссертации

Тезисыдокладов — СПб.,2009 — С381—382. — 0.1 п. л. — (авторствонеразделено).

2. ***Добров, А.В.*** **Технологии интеллектуального поиска и способы оценки их эффективности [текст] / А.В. Добров // Структурная и прикладная лингвистика. — СПб.: Издательство СПбГУ, 2010. — Вып. 8. — С219—232.** — 1 п. л.

3. *Добров, А.В.* К вопросу о создании компьютерного представления абхазского языка [текст] / А.В. Добров // Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы. Материалы международной конференции. — Йошкар-Ола, 2011 — С.60—63. — 0.2 п. л.

4. ***Добров, А.В.*** **Комплексный лингвистический подход к автоматической рубрикации новостных сообщений [текст] / А.В. Добров // Политическая лингвистика — Екатеринбург, изд-во УРГПУ: 2011 — Вып. 3(37) — с.202-209. —** 1.3 п. л.

5. ***Добров, А.В.*** **Автоматическая рубрикация текстов средствами комплексного лингвистического анализа [текст] / А.В. Добров // Структурная и прикладная лингвистика. — СПб.: Издательство СПбГУ, 2012. — Вып. 9. —** 0.9 п. л.

6. *Добров, А.В.* К вопросу о методике оценки эффективности автоматической рубрикации текстов: психолингвистический аспект [текст] / А.В. Добров // Психолінгвістика: [зб. наук. праць ДВНЗ "Переяслав-Хмельницький державний педагогічний університет імені Григорія Сковороди"] = Психолингвистика [сб. науч. трудов ГВУЗ «Переяслав-Хмельницкий государственный педагогический университет имени Григория Сковороды»]. — Переяслав-Хмельницкий: ПП "СКД", 2012. — Вып. 9. — С.173—178. — 0.3 п. л.

7. *Добров, А.В.* К вопросу об универсальном представлении концептуальных структур в системах индексирования и автоматической рубрикации текстов [текст] / А.В. Добров // Материалы XLI международной филологической конференции — секция прикладной и математической лингвистики 26—31 марта 2012 г. — СПб.: Филологический факультет, 2012. — 0.5 п. л.