На правах рукописи

Шелманов Артем Олегович

ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ И РАЗРАБОТКА ИНТЕГРИРОВАННОЙ СИСТЕМЫ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗА

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Москва, 2015

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного анализа Российской академии наук.

Научный руководитель: **Осипов Геннадий Семенович,** доктор физико-математических наук, профессор

Официальные оппоненты: **Лахути Делир Гасемович**, доктор технических наук, старший научный сотрудник, руководитель учебно-научного центра программного и лингвистического обеспечения интеллектуальных систем Российского государственного гуманитарного университета (РГГУ)

Лукашевич Наталья Валентиновна, кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского вычислительного центра Московского государственного университета имени М.В. Ломоносова (НИВЦ МГУ)

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук (ИПУ РАН)

Защита состоится «7» октября 2015 г. в 15 часов 00 минут на заседании диссертационного совета Д 002.073.01 при Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д.44, корп.2 (конференц-зал).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН и на официальном сайте ФИЦ ИУ РАН: http://www.ipiran.ru/announce/ .

Автореферат разослан «___» августа 2015 г.

Ученый секретарь
диссертационного совета Д 002.073.01,
доктор технических наук, профессор С.Н. Гринченко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Компьютерный лингвистический анализ текстов на естественном языке – перспективная быстроразвивающаяся область искусственного интеллекта. Одна из его главных целей заключается в построении структурированного представления текста, на основе которого можно решать прикладные задачи. Для многих методов решения задач вопросно-ответного поиска, извлечения информации и знаний из текстов, автоматического реферирования необходимо такое структурированное представление, которое строится в результате глубокого лингвистического анализа, включающего синтаксический и семантический анализ.

Существует значительное число разновидностей методов как синтаксического, так и семантического анализа, которые основаны на разных моделях синтаксической структуры предложения и различном понимании семантики. В настоящей диссертации исследуются методы построения синтаксических деревьев зависимостей и методы определения ролевых структур высказываний (semantic role labeling).

Деревья зависимостей моделируют синтаксическую структуру предложений в виде иерархии слов, связанных дугами, обозначающими синтаксическое подчинение между главным и зависимым словами. Подчинение обуславливается набором общих принципов, которые в целом сводятся к тому, что зависимое слово в предложении является уточняющим, необязательным, менее важным для передачи смысла высказывания, чем главное.

Модель семантики, основанная на ролевой структуре предложения, позволяет абстрагироваться от синтаксических деревьев предложений и сопоставлять разным грамматическим конструкциям одинаковые смысловые структуры. Задача определения ролевых структур высказываний включает в себя поиск предикатных слов, которые описывают в предложении ситуации (это, например, глаголы, причастия, отглагольные существительные), поиск их семантических аргументов – синтаксических конструкций, которые выражают в предложении участников ситуации, а также определение значений аргументов, т.е. назначение им семантических ролей, которые играют участники в ситуации.

Задачи синтаксического и семантического анализа, как правило, решаются раздельно: сначала строится синтаксическое дерево предложения, на основе которого затем строится его семантическая структура. Для этого существует ряд методов, основанных как на правилах, так и на машинном обучении. Хотя современные методы позволяют добиваться достаточно хорошего качества решения этих задач, остается значительное пространство для их улучшения.

Анализ ошибок синтаксических и семантических анализаторов показывает, что для построения правильной синтаксической структуры предложения необходимы знания о его семантике, но при этом ошибки в синтаксическом дереве негативно отражаются на качестве семантического анализа. Существует гипотеза о том, что совмещение синтаксического и семантического видов анализа может повысить их качество. Такой совмещённый подход к решению задачи глубокого лингвистического анализа будем называть семантико-синтаксическим анализом.

В настоящей диссертации разработан новый метод семантико-синтаксического анализа, в котором интегрированы методы построения синтаксических деревьев зависимостей и определения ролевых структур высказываний. За счет информации, полученной на этапе семантического анализа предложения, корректируется синтаксическое дерево, что в свою очередь помогает исправить ошибки в ролевой структуре высказывания. Метод позволяет значительно повысить качество как синтаксического, так и семантического анализа, что подтверждается проведенными экспериментами на размеченных русскоязычных корпусах текстов, а также улучшением качества решения прикладных задач обработки текстов.

Несмотря на то, что работы по созданию подходов, интегрирующих методы построения синтаксических деревьев зависимостей и методы определения ролевых структур высказываний, ведутся довольно давно, ранее не было предложено эффективного подхода, который позволил бы повысить как качество синтаксического, так и качество семантического анализа. Поэтому исследования в области семантико-синтаксического анализа, проведенные в настоящей диссертационной работе, являются актуальными.

Предмет исследования – методы и алгоритмы определения ролевых структур высказываний, а также методы и алгоритмы семантико-синтаксического анализа.

Целью исследования является повышение качества автоматического анализа текстов на естественном языке на основе интеграции методов синтаксического и семантического анализа.

Задачи исследования:

1. Провести исследование методов синтаксического и семантического анализа текстов на естественном языке.
2. Разработать метод определения ролевых структур высказываний в текстах на русском языке.

3. Разработать эффективный метод семантико-синтаксического анализа, в котором интегрированы методы построения синтаксических деревьев зависимостей и определения ролевых структур высказываний.
4. Реализовать методы семантического и семантико-синтаксического анализа. Разработать интегрированную систему семантико-синтаксического анализа.
5. Провести экспериментальные исследования методов семантического и семантико-синтаксического анализа.
6. Разработать методы решения прикладных задач, в которых используются результаты семантического и семантико-синтаксического анализа.
7. Провести экспериментальное исследование методов решения прикладных задач. Оценить влияние разработанных методов семантического и семантико-синтаксического анализа на качество решения этих задач.

Для решения поставленных задач применены следующие **методы исследования:** методы оптимизации, методы машинного обучения, методы компьютерной лингвистики, методы оценки качества алгоритмов машинного обучения, методы проверки статистической значимости полученных результатов, методы исследования качества синтаксического и семантического анализа, методы объектно-ориентированного проектирования программного обеспечения.

Научная новизна и результаты, выносимые на защиту:

1. Разработан новый метод автоматического определения ролевых структур высказываний, основанный на коммуникативной грамматике русского языка.
2. Разработан новый метод компьютерного семантико-синтаксического анализа текстов, в котором интегрированы методы построения синтаксических деревьев зависимостей и определения ролевых структур высказываний, позволяющий повысить точность и полноту синтаксического и семантического анализа по сравнению с реализацией, в которой эти виды анализа выполняются раздельно.
3. Разработана и реализована интегрированная система семантико-синтаксического анализа. Система применена для решения задач вопросно-ответного поиска, извлечения определений и авторских терминов из текстов научных публикаций.

4. Экспериментально показано, что при использовании интегрированной системы семантико-синтаксического анализа существенно повышается точность вопросно-ответного поиска по сравнению с раздельным применением методов синтаксического и семантического анализа.
5. Экспериментально показано, что использование ролевой структуры предложения повышает эффективность построения правил для извлечения определений и авторских терминов из текстов научных публикаций.

Теоретическая значимость работы состоит в создании и экспериментальном исследовании новых методов интеграции и взаимодействия синтаксического и семантического видов анализа текстов на естественном языке.

Практическая значимость: разработанные методы семантического и семантико-синтаксического анализа являются основой для извлечения информации и знаний из текстов, вопросно-ответного поиска, автоматического реферирования и для решения других прикладных задач обработки текстов на естественном языке и информационного поиска.

Разработанное программное обеспечение, включающее реализацию методов семантического и семантико-синтаксического анализа текстов на естественном языке, **внедрено** в следующих системах:

1. Информационно-поисковые сервисы портала «Руконт», «ООО Национальный цифровой ресурс «Руконт».
2. Электронно-библиотечная система «Znanium.com», «ООО Научно-издательский центр ИНФРА-М».
3. Информационно-аналитическая система «Exactus Expert», «ЗАО РосИнтернет технологии».
4. Метапоисковая машина «Exactus», «Федеральное государственное бюджетное учреждение науки Институт системного анализа РАН».

Результаты исследований по теме диссертационной работы **использованы** при выполнении научно-исследовательских работ по следующим проектам Минобрнауки РФ, программам ОНИТ РАН и грантам РФФИ:

1. «Создание программного комплекса информационно-аналитической поддержки научно-технической деятельности на основе вычислительного семантического поиска и анализа неструктурированной текстовой информации» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № 07.551.11.4003, 2011–2013 гг.).

2. «Исследование и разработка программного обеспечения понимания неструктурированной текстовой информации на русском и английском языках на базе создания методов компьютерного полного лингвистического анализа» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № 07.514.11.4134, 2012 – 2013 гг.).
3. «Развитие методов и технологии семантического поиска и анализа научных публикаций Exactus Expert» (в рамках проекта 2.9 ОНИТ РАН 2012 – 2013 гг.).
4. «Исследование и разработка новых методов автоматического семантико-синтаксического анализа текстов, основанных на коммуникативной грамматике, реляционно-ситуационной модели текста и теории неоднородных семантических сетей» (в рамках проекта 12-07-33068 мол_а_вед РФФИ 2012 – 2013 гг.).
5. «Исследование и разработка методов извлечения целевой информации из первичных научных публикаций на основе реляционно-ситуационного анализа текстов и активного машинного обучения с использованием индуктивных и статистических моделей» (в рамках проекта 14-29-05023 офи_м РФФИ 2014 – 2016 гг.).

Достоверность результатов подтверждена экспериментальными исследованиями разработанных методов и алгоритмов.

Апробация результатов исследования. Основные положения диссертации докладывались и обсуждались на следующих конференциях и семинарах:

1. XIII национальная конференция по искусственному интеллекту с международным участием (КИИ: Россия, Белгород, Белгородский государственный технологический университет, октябрь 2012 г.).
2. Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13) (Россия, Москва, март 2013 г.).
3. Международная конференция «Диалог 2014» (Россия, Бекасово, июнь 2014 г.).
4. Шестая международная конференция «Системный анализ и информационные технологии», (Россия, Калининградская обл., г. Светлогорск, июнь 2015 г.).

Публикации. Всего по теме исследования опубликовано 7 работ: 4 из них в рецензируемых изданиях из списка ВАК РФ и приравненных к ним, 2

публикации – в материалах международных и российских конференций, 1 – зарегистрированная программа для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка сокращений и условных обозначений, списка использованной литературы, а также четырех приложений. Полный объем диссертации составляет 210 страниц с 38 рисунками, 11 таблицами и 4 приложениями. Список литературы содержит 178 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, приведены методы исследования, изложены основные результаты и их научная новизна, обоснована теоретическая и практическая значимость полученных результатов, а также приведены данные о структуре и объеме диссертации.

В **первой главе** представлен аналитический обзор моделей синтаксической структуры предложения и семантики текста на естественном языке, проанализированы проблемы, возникающие при синтаксическом и семантическом анализе текстов, рассмотрены современные методы синтаксического и семантического анализа, а также подходы к их интеграции в системах семантико-синтаксического анализа. Обзор состоит из трех разделов.

Первый раздел главы посвящен синтаксическому анализу текстов на естественном языке. В нем рассмотрены задачи синтаксического анализа и две основные модели синтаксической структуры предложения: деревья составляющих и деревья зависимостей. Проанализированы проблемы, возникающие при проведении синтаксического анализа текстов, главной из которых является высокая степень неоднозначности естественных языков. Эта проблема приводит к тому, что классические алгоритмы, используемые для синтаксического анализа формальных языков, при анализе естественных языков становятся неэффективными или вовсе неприменимыми. Рассмотрены современные методы построения каждой модели.

Второй раздел главы посвящен семантическому анализу текстов на естественном языке. В нем рассмотрены следующие модели семантики текста:

- формальная семантика, которая описывает смысл естественно-языковых выражений на языке логики[1];

[1] Montague R. The proper treatment of quantification in ordinary English // Approaches to Natural Language / Ed. by K. J. J. Hintikka, J. Moravcsic, P. Suppes. — Springer, 1973. — P. 221–242.

- модель, основанная на ролевой структуре предложения, семантика предложения в которой представлена с помощью предикатных слов (глаголы, отглагольные существительные, причастия, деепричастия и др.), обозначающих в тексте ситуации, их аргументов – синтаксических конструкций, обозначающих участников ситуаций, и значений аргументов – семантических ролей, которые играют участники в ситуации[2];
- реляционно-ситуационная модель[3,4], в которой глубинные семантические структуры текста описываются с помощью аппарата неоднородных семантических сетей и которая опирается на теорию коммуникативной грамматики Г.А. Золотовой[5].

Представлены современные методы определения ролевых структур высказываний, основанные на машинном обучении с учителем и без учителя.

В третьем разделе главы проведен обзор подходов, в которых интегрируются синтаксический и семантический виды анализа текстов на естественном языке. Детально проанализированы подходы, в которых предлагаются способы интеграции методов построения синтаксических деревьев зависимостей и определения ролевых структур высказываний. В заключительной части главы сделаны выводы, обосновывающие актуальность исследования, поставлена цель и сформулированы задачи исследования.

Вторая глава посвящена разработанным методам семантического и семантико-синтаксического анализа текстов на естественном языке.

В первом разделе главы детализирована задача определения ролевых структур высказываний в текстах на русском языке. Она сформулирована следующим образом: получая на вход информацию о морфологической и синтаксической структуре (в виде синтаксического дерева зависимостей) предложения, а также категориально-семантические классы (КСК) слов (обобщенные значения слов), построить ролевую структуру предложения, в которой определены: предикатные слова (ПС), семантические аргументы ПС,

---

[2] Gildea D., Jurafsky D. Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.

[3] Osipov G. Methods for extracting semantic types of natural language statements from texts // 10th IEEE International Symposium on Intelligent Control. — Monterey, California, USA, 1995.

[4] Relational–situational method for intelligent search and analysis of scientific publications / Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov // Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13). — Vol. 968. — 2013.

[5] Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. — М.: Институт русского языка РАН им. В. В. Виноградова, 2004. — 544 с.

значения аргументов при заданных ПС (т.е. аргументам назначены семантические роли).

На рисунке 1 представлен пример синтаксической структуры предложения из корпуса СинТагРус[6] и пример его семантической структуры, которая должна быть построена семантическим анализатором.

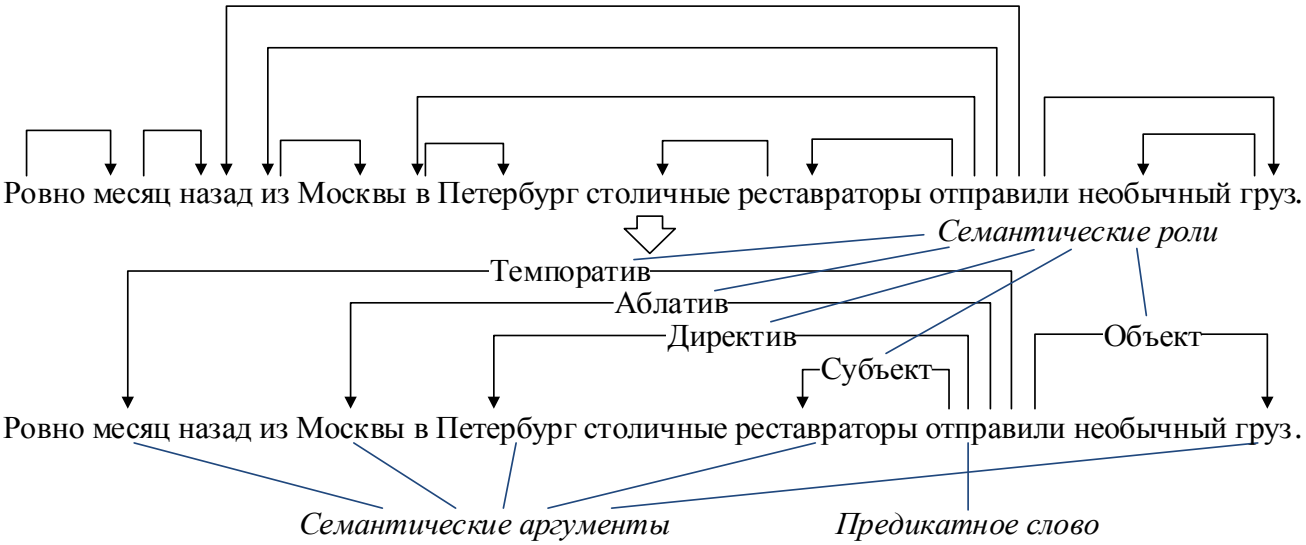


Рисунок 1 – Пример синтаксической структуры, подаваемой на вход семантического анализатора, и пример семантической структуры, которая должна быть построена в результате его работы

Постановка и решение задачи определения ролевых структур высказываний опирается на семантический словарь, разработанный в Институте системного анализа Российской академии наук экспертами в области теоретической лингвистики совместно со специалистами в области информатики[7,8]. Словарь основан на теории коммуникативной грамматики Г.А. Золотовой. В нем содержатся знания о предикатных словах, их ролевых структурах и признаках семантических ролей.

Во втором разделе главы представлен разработанный метод определения ролевых структур высказываний в текстах на русском языке, в котором используется семантический словарь. Дано описание общей схемы метода и

[6] Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю. Д. Апресян, И. М. Богуславский, Б. Л. Иомдин и др. // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. — С. 193–214.
[7] Осипов Г. С. Методы искусственного интеллекта. — М.: Физматлит, 2011. — 296 с.
[8] Завьялова О. С. О принципах построения словаря глаголов для задач автоматического анализа текста // Труды ежегодной международной конференции «Диалог» (2004). — 2004. — С. 198 – 201.

общего алгоритма определения ролевых структур высказываний, блок-схема которого представлена на рисунке 2.

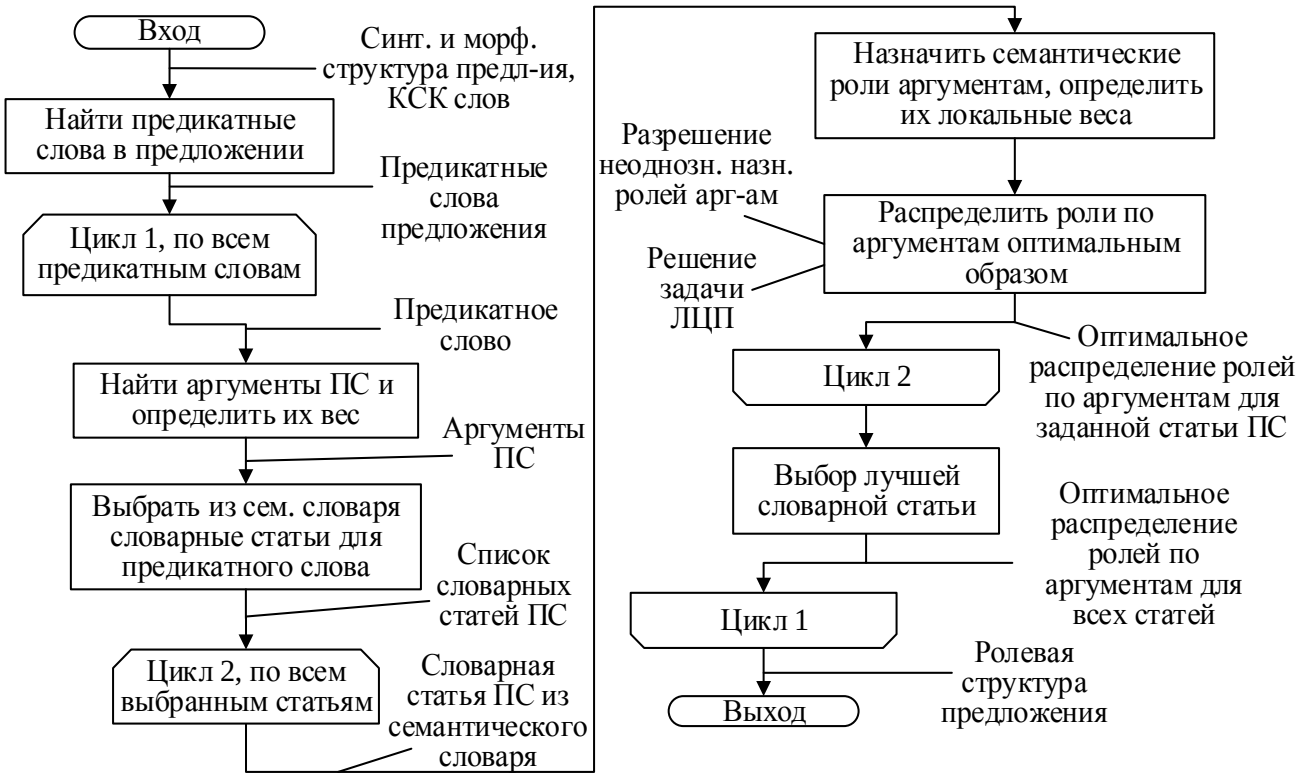


Рисунок 2 – Общий алгоритм определения ролевой структуры предложения

Рассмотрен метод поиска предикатных слов в предложении, использующий набор морфосинтаксических шаблонов. Приведены примеры шаблонов и конструкций, соответствующих предикатным словам.

Представлен метод поиска семантических аргументов, основанный на правилах, в которых анализируются лексические, морфологические, синтаксические признаки слов. На основе этих признаков и набора эвристик вычисляется вес аргумента или «степень уверенности» в том, что семантический аргумент был выделен правильно.

Описан метод назначения семантических ролей аргументам предикатных слов и определения весов ролей. Рассмотрим его более подробно.

Сначала аргументам назначаются «основные» роли. «Основные» («обязательные» или «центральные») роли – это те роли ПС, которые чаще всего встречаются при нем, и без которых описание значения предикатного слова является семантически неполным. Назначение «основных» ролей аргументам осуществляется главным образом на основе информации, содержащейся в разделе ролевых структур предикатных слов семантического словаря. Для этого используется три признака: предлог, падеж, категориально-семантический класс

слова. Предлагается метод нечеткого сравнения признаков аргументов с признаками ролей в словарных статьях семантического словаря и вычисления оценки соответствия аргумента семантической роли (вес роли для заданного аргумента). Сопоставление является нечетким, поскольку аргументы могут иметь несколько значений признаков (например, несколько КСК или падежей из-за неснятой омонимии), совпадения одних признаков оцениваются весомее других, кроме того, вес роли может понижаться за счет эвристик, учитывающих признаки семантических ролей в словарных статьях, а также признаки предикатных слов и аргументов в тексте.

Аргументам, которые являются вопросительными словами, роли назначаются с помощью отдельного раздела семантического словаря. Аргументам, которым не была назначена семантическая роль из набора «основных» ролей, может быть назначена одна из «периферийных» или «необязательных» ролей (локатив, темпоратив, инструментив, директив, каузатив, и др.) с помощью раздела периферийных ролей семантического словаря.

При назначении семантических ролей аргументам, как правило, нарушается основное ограничение ролевой структуры предикатного слова: каждый аргумент может иметь не более одной роли, и не может быть двух аргументов с одинаковой ролью при заданном предикатном слове (за исключением случаев с симметричными глаголами). Возникает задача распределения ролей по аргументам. Представлен метод снятия неоднозначности распределения семантических ролей по аргументам с помощью решения задачи целочисленного программирования. Поскольку каждый семантический аргумент и каждая семантическая роль обладают некоторым весом, можно определить наилучшую ролевую структуру для заданного предикатного слова, решив оптимизационную задачу. Рассмотрим постановку этой задачи.

Пусть A – упорядоченное множество семантических аргументов; R – упорядоченное множество всех уникальных ролей, которые можно назначить аргументам A: $R = \bigcup_{i=1}^{|A|} R_i$, где R_i – упорядоченное множество ролей, которые были назначены аргументу $a_i \in A$ $(i = 1, \dots |A|)$; r_j – семантические роли $r_j \in R$ $(j = 1, \dots |R|)$. Поскольку в общем случае $|A| \neq |R|$, положим $n = \max\{|A|, |R|\}$. Введем переменные $x_{ij} \in \{0,1\}$ $(i, j = 1, \dots n)$ такие, что для $i \leq |A|$ и $j \leq |R|$ $x_{ij} = 1$, если i-ому аргументу назначена j-ая роль, иначе $x_{ij} = 0$; для $i > |A|$ или $j > |R|$ x_{ij} – служебные переменные, которые не соответствуют никаким ролям и аргументам в тексте. Обозначим: $w_i^{arg} > 0 -$ вес i-ого

аргумента, полученный на этапе выявления семантических аргументов; $w_{ij}^{role} > 0 -$ вес j-ой роли для i-ого аргумента, полученный на этапе назначения ролей аргументам. Пусть $C(i,j)$ – функция стоимости назначения i-ому аргументу j-ую роль:

$$C(i,j) = \begin{cases} -w_i^{arg} \cdot w_{ij}^{role}, & \text{если } i \leq |A| \text{ и } j \leq |R| \text{ и } r_j \in R_i \\ 0, & \text{иначе} \end{cases}.$$

Тогда задачу разрешения неоднозначности распределения семантических ролей по аргументам можно свести к оптимизационной задаче о назначениях, которая решается с помощью венгерского метода. Решение задачи используется для назначения аргументам ролей из заданной словарной статьи ПС, а значение целевой функции определяет насколько хорошо словарная статья подходит для заданного предикатного слова и его аргументов. Оно используется для выбора оптимальной словарной статьи и соответствующей ей ролевой структуры предикатного слова.

Третий раздел главы посвящен разработанному методу семантико-синтаксического анализа. Он начинается с исследования того, как информация о ролевой структуре предикатных слов может помочь при проведении синтаксического анализа предложений.

Далее приведено описание разработанного метода семантико-синтаксического анализа. В нем информация, полученная на этапе определения ролевых структур высказываний, используется для корректировки синтаксического дерева. Вначале проводится синтаксический анализ, в результате которого строится исходная синтаксическая структура. Затем в строгом соответствии с построенной синтаксической структурой выявляются предикатные слова и их семантические аргументы, которые назовем «основными». «Основным» аргументам назначаются семантические роли. Далее находятся «возможные» аргументы, которые потенциально не выявлены из-за ошибок в синтаксическом дереве предложения, и проверяется возможность включения их в ролевую структуру предикатных слов. Если ролевая структура предикатного слова допускает новый аргумент, он вносится в список «дополнительных» аргументов. Затем синтаксическое дерево корректируется так, чтобы оно соответствовало новой ролевой структуре предикатного слова с включенными в нее «дополнительными» аргументами. В итоге на исправленном синтаксическом дереве заново проводится семантический анализ. На рисунке 3

представлена блок-схема общего алгоритма семантико-синтаксического анализа.

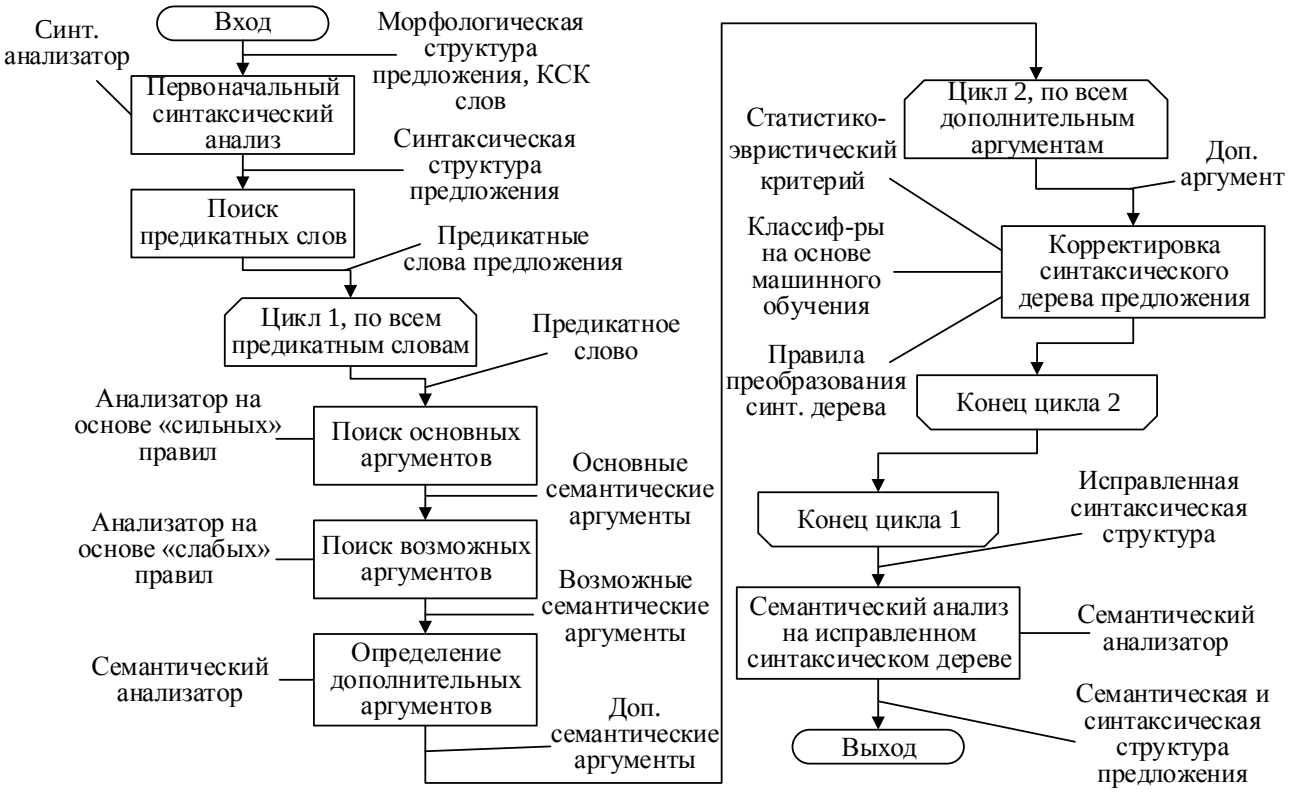


Рисунок 3 – Общий алгоритм семантико-синтаксического анализа

Рассмотрен метод поиска «возможных» аргументов. В нем накладываются менее жесткие ограничения на синтаксическую структуру предложения, чем в методе поиска «основных» семантических аргументов, и не требуется, чтобы «возможные» аргументы были синтаксически связаны с предикатными словами. Однако для того, чтобы отфильтровать наименее подходящие варианты, в нем вводится ряд дополнительных ограничений на аргументы. Проверяется, что «возможный» аргумент: находится на одном уровне скобочной структуры вместе с ПС, является синтаксической вершиной составной именованной сущности, не входит в устойчивое выражение, не является «основным» аргументом другого предикатного слова и др.

Представлен метод выявления «дополнительных» аргументов. Сначала анализатор назначает семантические роли «основным» аргументам заданного предикатного слова. Затем каждый «возможный» аргумент поочередно добавляется к множеству «основных» аргументов, и анализатор пытается назначить роли расширенному набору аргументов. Если добавленному «возможному» аргументу назначить роль не удалось, то он отвергается. Если роль новому аргументу назначена, то проверяется, принадлежит ли эта роль аргументу из «основного» нерасширенного набора в исходной ролевой

структуре. Если нет, «возможный» аргумент добавляется к множеству «дополнительных» аргументов, иначе возможный аргумент отвергается.

После того как все «возможные» аргументы проанализированы, отобранные «дополнительные» аргументы добавляются к множеству «основных» аргументов, и формируется общий набор семантических аргументов. Процедура семантического анализа запускается на этом наборе заново. Те «дополнительные» аргументы, которые семантическую роль не получили, отвергаются.

Описан метод корректировки синтаксического дерева: представлены основные варианты корректировки, приведены их примеры. Кроме того, рассмотрены основные проверки новой синтаксической структуры: проверка наличия циклов (а также метод разрыва циклов), проверка проективности добавленных связей, проверка преобразования дерева либо с помощью метода на основе статистико-эвристического критерия, учитывающего совместную встречаемость признаков слов, либо с помощью метода на основе машинного обучения.

В последнем методе применяется ансамбль из трех бинарных классификаторов. Первый классификатор C_d оценивает вес удаленной связи между дополнительным аргументом и его родителем в исходном синтаксическом дереве. Второй классификатор C_a оценивает вес добавленной связи между аргументом и предикатным словом. Третий классификатор C_f получает на вход значения целевых функций от первых двух и принимает окончательное решение фиксировать или нет изменения в синтаксическом дереве. В качестве метода машинного обучения используется метод опорных векторов. В первых двух классификаторах используется линейное ядро, поскольку их признаковое пространство велико (порядка нескольких сотен тысяч признаков), в последнем классификаторе используется радиальное ядро.

В **третьей главе** представлены результаты экспериментальных исследований разработанных методов семантического и семантико-синтаксического анализа на размеченных корпусах русскоязычных текстов.

В первом разделе главы дано описание тестовых данных, на которых проводились эксперименты, и метрик оценки качества синтаксического и семантического анализа текстов на естественном языке. Кратко охарактеризован синтаксически размеченный корпус русского языка (СинТагРус), который использовался для машинного обучения и оценки качества синтаксического анализа (более 50 000 предложений и более 770 000 токенов без учета пунктуации). Кроме того, дано описание корпуса для оценки качества

определения ролевых структур высказываний (более 1 700 предложений и около 29 000 токенов без учета пунктуации; размечено около 3 000 предикатных слов и 4 000 ролей), а также корпуса для оценки качества метода поиска семантических аргументов (200 предложений, около 1 700 токенов без учета пунктуации; размечено около 800 семантических аргументов для более чем 460 предикатных слов).

Во втором разделе главы приведены результаты оценки качества синтаксического и морфологического анализаторов, а также анализатора для определения категориально-семантических классов слов.

В третьем разделе главы описаны экспериментальные исследования метода поиска семантических аргументов на эталонных синтаксических деревьях (Gold_Morph_Synt) и на деревьях, сгенерированных автоматически (Morph_Synt). В таблице 1 представлены полученные значения точности p, полноты r и F_1-меры.

Таблица 1 – Результаты экспериментальной оценки метода выявления семантических аргументов

Конфигурация	$\boldsymbol{p}$, %	$\boldsymbol{r}$, %	$\boldsymbol{F_1}$, %
Gold_Morph_Synt	94,4	95,0	94,6
Morph_Synt	86,1	83,2	84,6

Сделаны выводы о применимости метода для нахождения основных аргументов в системе семантико-синтаксического анализа и о важности качественной синтаксической разметки как для поиска семантических аргументов, так и для решения задачи определения ролевых структур высказываний в целом. Проведен анализ ошибок метода.

В четвертом разделе главы описаны экспериментальные исследования метода определения ролевых структур высказываний. Качество метода оценивалось при работе на разметке, взятой из «золотого стандарта», и на разметке, сгенерированной автоматически:

- Gold_Morph_CSC_Synt – морфологические признаки, синтаксическая разметка и КСК из «золотого стандарта»;
- CSC_Gold_Morph_Synt – морфологические признаки и синтаксическая разметка из «золотого стандарта», а КСК определяются автоматически;
- Synt_Gold_Morph_CSC – морфологические признаки и КСК из «золотого стандарта», а синтаксическая разметка генерируется автоматически;

- CSC_Synt_Gold_Morph – морфологические признаки из «золотого стандарта», а КСК и синтаксическая разметка генерируются автоматически;
- Morph_CSC_Synt – вся разметка генерируется автоматически.

Результаты представлены в таблице 2.

Таблица 2 – Результаты экспериментальной оценки метода определения ролевых структур высказываний

Конфигурация	***p*, %**	***r*, %**	***F_1*, %**
Gold_Morph_CSC_Synt	92,3	80,8	86,2
CSC_Gold_Morph_Synt	89,4	73,4	80,6
Synt_Gold_Morph_CSC	92,1	74,3	82,3
CSC_Synt_Gold_Morph	89,2	67,3	76,7
Morph_CSC_Synt	89,6	61,0	72,6

Полученные результаты свидетельствуют о том, что разработанный семантический анализатор для текстов на русском языке демонстрирует качество решения задачи определения ролевых структур высказываний на уровне современных анализаторов для других языков с малым количеством языковых ресурсов. Сделан вывод о значимости тех или иных признаков для решения задачи определения ролевых структур высказываний, проведен анализ ошибок метода.

В пятом разделе главы описаны экспериментальные исследования метода семантико-синтаксического анализа. Проведена оценка влияния метода семантико-синтаксического анализа на качество синтаксического анализа путем сравнения его с синтаксическим анализатором MaltParser[9], который в настоящее время является одним из лучших синтаксических анализаторов на основе машинного обучения для русскоязычных текстов[10,11,12].

[9] MaltParser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95-135.

[10] Nivre J., Boguslavsky I. M., Iomdin L. L. Parsing the SynTagRus treebank of Russian // Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008). — 2008. — P. 641–648.

[11] Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Papers from the Annual International Conference "Dialogue" (2011). — 2011. — P. 591 – 604.

[12] Казенников А. О. Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей // Труды международной конференции «Диалог» (2010). — 2010. — С. 157 – 162.

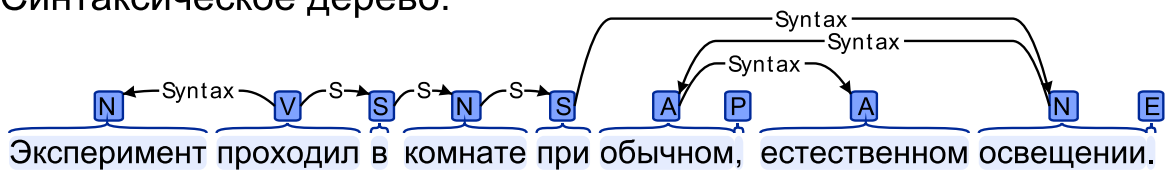
Описаны методика оценки, корпуса для машинного обучения и для тестирования. Сравнивались три анализатора: MaltParser; система семантико-синтаксического анализа, в которой реализован статистико-эвристический критерий проверки исправлений («Сем.-син. Стат.»); система семантико-синтаксического анализа, в которой реализована проверка исправлений с помощью ансамбля классификаторов на основе машинного обучения («Сем.-син. МО»). Результаты представлены в таблице 3.

Таблица 3 – Качество построения синтаксических деревьев зависимостей с помощью анализатора MaltParser и с помощью системы семантико-синтаксического анализа

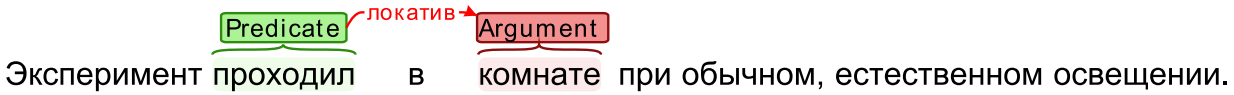
Анализатор	$\boldsymbol{p}$, %	$\boldsymbol{r}$, %	$\boldsymbol{F_1}$, %
MaltParser	**90,6**	89,0	89,8
Сем.-син. Стат.	90,5	**90,2**	90,3
Сем.-син. МО	**90,6**	90,1	**90,4**

Полученные результаты показывают, что разработанный метод семантико-синтаксического анализа позволяет повысить полноту установления синтаксических связей между предикатными словами и аргументами более чем на 1,1 %, тем самым ему удается установить более 10 % всех наиболее сложных синтаксических связей между предикатными словами и аргументами, которые не были найдены MaltParser. При этом точность семантико-синтаксического анализатора не изменяется. Сделан вывод об эффективности методов проверки исправлений синтаксического дерева. Точность внесения исправлений в синтаксическое дерево с помощью системы семантико-синтаксического анализа составляет 82-85% в зависимости от метода проверки исправлений, параметров классификаторов, а также порога решающего правила. Приведены примеры, в которых система семантико-синтаксического анализа исправила синтаксическое дерево (рисунок 4), проведен анализ ошибок метода.

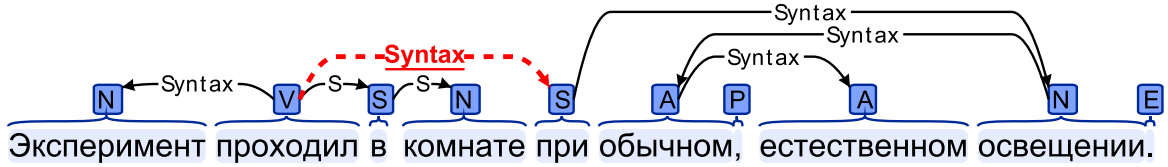
А) Синтаксическое дерево:



Ролевая структура предложения:



Б) Синтаксическое дерево:



Ролевая структура предложения:

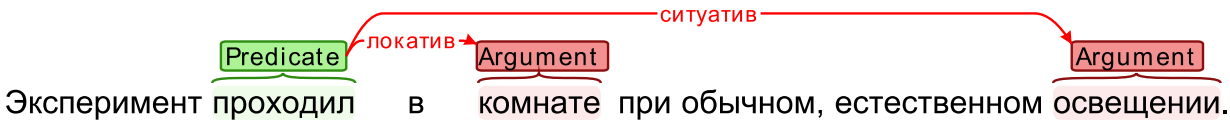


Рисунок 4 – Пример, в котором система семантико-синтаксического анализа откорректировала синтаксическое дерево. На рисунке А) – пример работы MaltParser, на рисунке Б) – пример работы системы семантико-синтаксического анализа (добавленная в синтаксическое дерево связь выделена пунктиром)

Далее описаны эксперименты по оценке влияния метода семантико-синтаксического анализа на качество определения ролевых структур высказываний. Сравнивались две системы: в первой синтаксический и семантический анализ выполняются раздельно («Базовая»), вторая – система семантико-синтаксического анализа, в которой реализована проверка исправлений с помощью ансамбля классификаторов на основе машинного обучения («Сем.-син.»). Результаты приведены в таблице 4.

Таблица 4 – Качество семантического анализа с помощью системы семантико-синтаксического анализа и с помощью системы, в которой синтаксический и семантический анализ выполняются раздельно

Система	$\boldsymbol{p}$, %	$\boldsymbol{r}$, %	$\boldsymbol{F_1}$, %
Базовая	**89,6**	61,0	72,6
Сем.-син.	**89,6**	**62,7**	**73,8**

Результаты показывают, что применение разработанного метода семантико-синтаксического анализа ведет к повышению качества определения ролевых

структур высказываний. В частности, система семантико-синтаксического анализа позволяет повысить полноту на 1,7%, что ведет к повышению F_1-меры на 1,2%. При этом точность не уменьшается. За счет корректировки синтаксической структуры повышается доля правильно выявленных семантических аргументов, что в итоге позволяет поднять качество семантического анализа в целом.

В **четвертой главе** предложены методы решения прикладных задач обработки текстов на естественном языке, в которых используется информация, получаемая в результате семантического и семантико-синтаксического анализа.

В первом разделе главы рассмотрена задача построения семантической сети предложения для реляционно-ситуационной модели текста. Описан метод построения сети на основе информации о ролевой структуре предложения. Обоснована важность построения семантической сети над ролевой структурой: она позволяет нормализовать отношения между концептами предметной области, имеющими отражение в тексте. На рисунке 5 представлен пример, в котором аргументы имеют разные семантические роли, но между ними установлено одно и тоже семантическое отношение причины-следствия (CAUS).

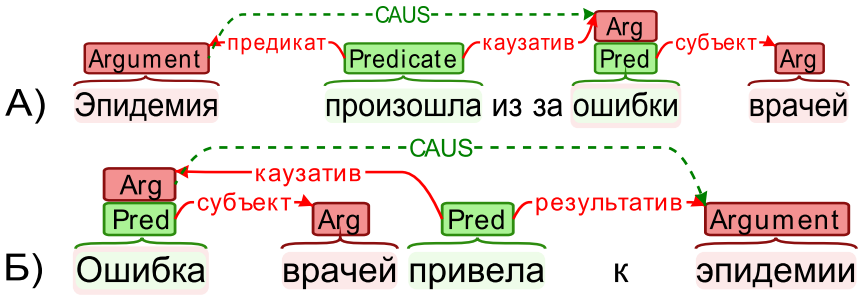


Рисунок 5 – Пример, в котором аргументы имеют разные семантические роли, но между ними установлено одно и то же семантическое отношение

Во втором разделе главы рассмотрена задача вопросно-ответного поиска в метапоисковой системе. Приведена схема работы метапоисковой системы: она перенаправляет запросы пользователей к другим поисковым машинам в Интернет (Yandex, Google, Bing, Yahoo), агрегирует полученные от них ссылки и сниппеты – фрагменты текста, которые выводятся рядом со ссылкой в поисковой выдаче, и ранжирует их специальным образом. Далее описан метод ранжирования сниппетов для вопросно-ответного поиска в метапоисковой системе, в котором используется информация, полученная в результате реляционно-ситуационного анализа, включающего в себя семантический или семантико-синтаксический анализ, а также построение семантической сети. В

методе используется сочетание лексических и семантических критериев сходства запроса и сниппета.

Вначале дано общее описание разработанного алгоритма определения релевантности сниппетов запросу, а затем детально описан алгоритм определения лексико-семантической оценки релевантности предложения сниппета запросу. Лексико-семантическая оценка релевантности предложения сниппета вопросительному запросу $r_{ls}^{s} \in [0, 1]$ вычисляется как линейная свертка трех компонент:

- лексическая оценка предложения – оценивает близость запроса и предложения сниппета по лексике;
- оценка семантических ролей – оценивает близость запроса и сниппета по предикатно-аргументным структурам, используя их семантические роли;
- оценка семантических отношений – оценивает близость запроса и сниппета по семантическим отношениям семантической сети.

Детально рассмотрены разработанные алгоритмы вычисления всех оценок.

Далее приведены экспериментальные исследования метода ранжирования сниппетов для вопросно-ответного поиска в метапоисковой системе. Описан процесс формирования тестовых данных, приведены содержащиеся в них примеры вопросительных запросов и ответов, представлена методика оценки. Для оценки качества вопросно-ответного поиска вычислялась точность системы как отношение количества вопросов, ответы на которые система дала на уровне выдачи не больше чем $d > 0$ ко всему количеству запросов. Точность оценивалась при $d = 1$, $d = 2$, $d = 3$, $d = 4$. Рассчитывалась оценка метрики mean reciprocal rank (MRR)[13] – средний обратный ранг ответа. Эта метрика наряду с правильными ответами системы, которые оказались первыми, учитывает также ответы, которые находятся на небольшой глубине выдачи, при этом штрафует их в зависимости от глубины, снижая их вклад в итоговую оценку.

Далее приведены результаты экспериментальной оценки вопросно-ответных систем, в которых реализованы различные способы ранжирования: система, которая ранжирует ответы случайным образом («Случ. ранж.»); система, которая ранжирует ответы на основе только лексического критерия («Лексич. ранж.»); система, которая вычисляет релевантность с учетом семантической информации, полученной от системы, в которой синтаксический и семантический анализ выполняются раздельно («Сем. ранж. Сем. ан.»);

[13] Voorhees E. M., Tice D. M. The TREC-8 question answering track report // Proceedings of the Second International Conference on Language Resources and Evaluation, (LREC 2000). — 2000.

система, которая вычисляет релевантность с учетом семантической информации, полученной от системы семантико-синтаксического анализа («Сем. ранж. Сем.-син. ан.»). В таблице 5 и на рисунке 6 представлены значения двух, чаще всего используемых метрик для оценки качества вопросно-ответного поиска: точности на уровне $d = 1$ и MRR.

Таблица 5 – Результаты оценки точности и MRR для различных способов ранжирования

Способ ранжирования	**Точность, d = 1, %**	**MRR, %**
Сем. ранж. Сем. ан.	53,9	67,2
Сем. ранж. Сем.-син. ан.	**58,0**	**69,6**
Лексич. ранж.	46,1	59,0
Случ. ранж.	26,0	-

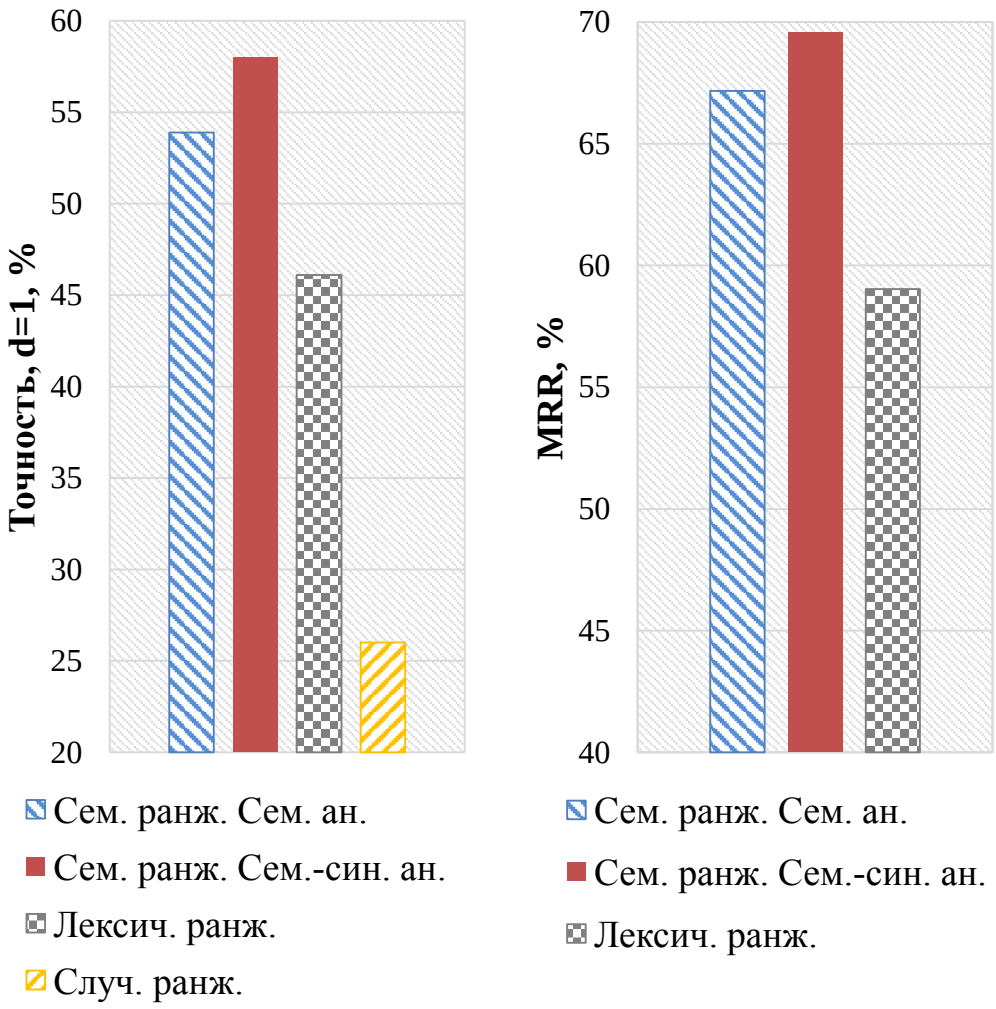


Рисунок 6 – Точность и MRR ранжирования

На основе результатов экспериментов сделаны следующие выводы:

- Разработанный критерий значительно повышает точность ранжирования ответов при решении задачи вопросно-ответного поиска. Точность на 32% выше по сравнению со случайным ранжированием.
- Учет семантической структуры предложения при оценке релевантности ответов в вопросно-ответном поиске позволяет повысить точность на 12% и MRR на более чем 10% по сравнению с лексическим критерием ранжирования и позволяет также извлекать из сниппета сам ответ на вопрос.
- Качество ранжирования при использовании системы семантико-синтаксического анализа существенно выше, чем при использовании системы, в которой синтаксический и семантический анализ выполняются раздельно: MRR выше почти на 2,5%, точность на уровне $d = 1$ выше на 4%, что составляет треть прироста точности ранжирования при использовании семантической информации по сравнению с лексическим критерием.

В третьем разделе главы рассмотрена задача автоматического извлечения определений и авторских терминов из текстов научных публикаций. Авторским термином называется термин, которому в тексте дается определение. Предложен метод, основанный на сравнении лексико-синтаксической и семантической структуры предложения со списком фреймов (шаблонов). Приведены примеры подобных фреймов и алгоритм сопоставления фреймов с текстом.

Далее описаны экспериментальные исследования метода, тестовые данные и методика оценки. Приведены результаты экспериментов для трех систем извлечения определений и авторских терминов:

- система, в которой реализованы фреймы, использующие семантические роли, полученные от анализатора, в котором синтаксический и семантический анализ выполняются раздельно («Sem»);
- система, в которой реализованы фреймы, использующие семантические роли, полученные от интегрированной системы семантико-синтаксического анализа («SemSyn»);
- система, в которой отсутствуют фреймы, использующие семантические роли («NoSem»).

Результаты экспериментальных исследований представлены в таблице 6.

Таблица 6 – Результаты оценки метода извлечения определений и авторских терминов из текстов научных публикаций

Система	***p*,%**	***r*,%**	***F*$_1$,%**
Sem	82,6	68,3	74,8
SemSyn	**82,7**	**68,7**	**75,0**
NoSem	81,6	58,3	68,0

Полученные результаты демонстрируют эффективность разработанного метода и значимость вклада фреймов, учитывающих семантические роли, в решение этой задачи. Использование ролевой структуры предложения упрощает построение фреймов для извлечения определений и авторских терминов. Применение системы семантико-синтаксического анализа не дало значимого прироста качества по сравнению с применением системы, в которой синтаксический и семантический анализ выполняются раздельно. Причина этого заключается в том, что большую роль в решении этой задачи играют лексико-морфологические фреймы, не учитывающие семантику и синтаксис предложений. Далее проведен анализ ошибок метода.

В **заключении** приведены основные результаты и выводы диссертационной работы.

В **первом приложении** описаны особенности реализации программного обеспечения для семантического и семантико-синтаксического анализа текстов на естественном языке. Приведены его характеристики и примененные технологии (таблица 7).

Таблица 7 – Особенности реализации программных компонент для лингвистического анализа текстов на ЕЯ и для проведения экспериментальных исследований

Тип программного компонента	**Язык реализации**	**Объем кода**
Модули лингвистического анализа	C++	более 40 000 строк кода
Сценарии для проведения экспериментальных исследований	Python 2.7	более 3 000 строк кода

Определены требования, которым должен удовлетворять лингвистический процессор, описана реализованная в ходе работы программная архитектура, позволяющая выполнить все эти требования.

Во **втором приложении** представлена структура семантического словаря. В **третьем приложении** описаны эксперименты с обучаемым синтаксическим анализатором MaltParser. В **четвертом приложении** приведены примеры, в которых разработанная система семантико-синтаксического анализа исправила синтаксическую и семантическую структуру предложения.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработан новый метод автоматического определения ролевых структур высказываний, основанный на коммуникативной грамматике русского языка.
2. Разработан новый метод компьютерного семантико-синтаксического анализа текстов, в котором интегрированы методы построения синтаксических деревьев зависимостей и определения ролевых структур высказываний, позволяющий повысить точность и полноту синтаксического и семантического анализа по сравнению с реализацией, в которой эти виды анализа выполняются раздельно.
3. Разработана и реализована интегрированная система семантико-синтаксического анализа. Система применена для решения задач вопросно-ответного поиска, извлечения определений и авторских терминов из текстов научных публикаций.
4. Экспериментально показано, что при использовании интегрированной системы семантико-синтаксического анализа существенно повышается точность вопросно-ответного поиска по сравнению с раздельным применением методов синтаксического и семантического анализа.
5. Экспериментально показано, что использование ролевой структуры предложения повышает эффективность построения правил для извлечения определений и авторских терминов из текстов научных публикаций.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ИССЛЕДОВАНИЯ

Публикации в изданиях, входящих в перечень ВАК и приравненных к ним:

1. Смирнов И.В., Шелманов А.О. Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов. // Искусственный интеллект и принятие решений. — М: ИСА РАН. — 2013. — №1. — С. 41–54.
2. Семантико-синтаксический анализ естественных языков Часть II. Метод семантико-синтаксического анализа текстов / И.В. Смирнов, А.О. Шелманов, Е.С. Кузнецова, И.В. Храмоин // Искусственный интеллект и принятие решений. — М: ИСА РАН. — 2014. — № 1. — С. 11–24.
3. Relational–situational method for intelligent search and analysis of scientific publications / Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov // In Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13). — Vol. 968. — 2013. — P. 57–64.
4. Shelmanov A.O., Smirnov I.V. Methods for Semantic Role Labeling of Russian Texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014). — Issue 13 (20). — 2014. — P. 607–619.

Прочие публикации:

5. Шелманов А.О. Метод автоматического выделения многословных терминов из текстов научных публикаций // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. — Белгород: БГТУ. — 2012. — т. 1. — С. 268–274.
6. Осипов Г.С., Шелманов А.О. Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой международной конференции «Системный анализ и информационные технологии» (САИТ). — 2015. — т. 1. — С. 229–240.
7. Шелманов А. О., Смирнов И. В. «Программа лингвистического анализа неструктурированной текстовой информации на русском и английском языках» // Свидетельство о государственной регистрации программы для ЭВМ. № 2013613430. — 2013.

Личный вклад соискателя: опубликованные в этих работах результаты, относящиеся к методам семантического и семантико-синтаксического анализа, к методам вопросно-ответного поиска и методам извлечения определений и авторских терминов из текстов научных публикаций, получены лично автором.

Шелманов Артем Олегович (Россия)

ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ И РАЗРАБОТКА ИНТЕГРИРОВАННОЙ СИСТЕМЫ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗА

Разработан новый метод автоматического определения ролевых структур высказываний, основанный на коммуникативной грамматике русского языка.

Разработан новый метод компьютерного семантико-синтаксического анализа текстов, в котором интегрированы методы построения синтаксических деревьев зависимостей и определения ролевых структур высказываний, позволяющий повысить точность и полноту синтаксического и семантического анализа по сравнению с реализацией, в которой эти виды анализа выполняются раздельно.

Разработана и реализована интегрированная система семантико-синтаксического анализа. Система применена для решения задач вопросно-ответного поиска, извлечения определений и авторских терминов из текстов научных публикаций.

Экспериментально показано, что при использовании интегрированной системы семантико-синтаксического анализа существенно повышается точность вопросно-ответного поиска по сравнению с раздельным применением методов синтаксического и семантического анализа.

Экспериментально показано, что использование ролевой структуры предложения повышает эффективность построения правил для извлечения определений и авторских терминов из текстов научных публикаций.

Shelmanov Artem (Russia)

RESEARCH OF METHODS FOR NATURAL LANGUAGE PROCESSING AND DEVELOPMENT OF THE INTEGRATED SYSTEM FOR SEMANTIC-SYNTACTIC PARSING

The new method for semantic role labeling based on communicative grammar of the Russian language was developed.

The new method for semantic-syntactic parsing of natural language texts was developed. It integrates methods for dependency parsing and semantic role labeling. It has the better precision and recall of syntactic and semantic parsing than the implementation, in which these types of parsing are performed separately.

The integrated system for semantic-syntactic parsing was developed and implemented. The system was applied to the tasks of question answering, definition and author’s term extraction from scientific publications.

Experiments show that the usage of the integrated system for semantic-syntactic parsing significantly increases the accuracy of question answering compared with the usage of separate methods for syntactic and semantic parsing.

Experiments show that the usage of role structure of sentences improves the efficiency of the construction of rules for extracting definitions and author’s terms from scientific publications.