«**УТВЕРЖДАЮ**»
Директор ФГБУН «Институт проблем управления им. В. А. Трапезникова РАН
академик РАН, д.ф.-м.н.

_____ С.Н. Васильев.

«_04_» сентября 2015 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного учреждения науки Института проблем управления им. В. А. Трапезникова Российской академии наук
на диссертацию
Артема Олеговича Шелманова
«ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ И РАЗРАБОТКА ИНТЕГРИРОВАННОЙ СИСТЕМЫ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗА»,
представленную на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики»

Актуальность. Решение многих задач информационного поиска и обработки текстов на естественном языке осуществляется с использованием автоматического синтаксического и семантического анализа. Среди этих задач можно назвать вопросно-ответный поиск, извлечение информации и знаний из текстов, автоматическое реферирование.

В традиционных подходах к автоматическому анализу текстов синтаксический и семантический анализ выполняются раздельно: сначала строится синтаксическое дерево предложения, на основе которого затем строится его ролевая структура. Однако без знания о смысле предложения не всегда возможно однозначно построить правильное синтаксическое дерево. Правила грамматики естественного языка часто допускают множество вариантов разбора, но не все из них являются семантически верными. Ошибки при построении синтаксического дерева негативно отражаются на дальнейших этапах обработки текста, в том числе и на этапе семантического анализа. Существует ряд подходов к разрешению синтаксической неоднозначности путем анализа семантики, однако задача повышения их точности по-прежнему остается *актуальной.* Именно этой задаче посвящена диссертация А.О. Шелманова. В ней предлагается оригинальный метод семантико-синтаксического анализа, который за счет интеграции этапов построения

синтаксических деревьев зависимостей и определения ролевых структур высказываний позволяет повысить точность и полноту как синтаксического, так и семантического анализа. Это в свою очередь позволяет повысить качество решения прикладных задач информационного поиска и обработки текстов на естественном языке. Поэтому работа А.О. Шелманова является актуальной.

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

Состав и содержание диссертации. Диссертация А.О. Шелманова 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.

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 и методов синтаксического и семантического видов анализа текстов на естественном языке. Рассмотрены существующие модели семантики текста (формальная семантика Монтегю, модель ролевых структур и реляционно-ситуационная модель) и методы интеграции синтаксического и семантического анализа.

Во второй главе рассмотрены разработанные автором методы семантического и семантико-синтаксического анализа. Предложенный метод семантического анализа для определения ролевых структур высказываний опирается на семантический словарь, разработанный в ИСА РАН и основанный на теории коммуникативной грамматики русского языка Г.А. Золотовой. Сформулирована возникающая в ходе работы метода задача назначения семантических ролей фрагментам предложения. Предложено решение этой задачи на основе методов целочисленного программирования. В методе семантико-синтаксического анализа информация, полученная на этапе семантического анализа, используется для корректировки синтаксического дерева предложения, что в свою очередь помогает исправить ошибки в его ролевой структуре. В методе применяются подходы, основанные как на правилах, так и на машинном обучении.

В третьей главе представлены экспериментальные исследования разработанных методов. Описаны данные, на которых проводились исследования, методики оценки, представлены результаты экспериментов.

Проведено экспериментальное сравнение с известными методами и показана более высокая эффективность предложенных методов. Приведены примеры.

В четвертой главе представлены методы решения прикладных задач обработки текстов на естественном языке, опирающиеся на структуры, полученные с помощью семантического или семантико-синтаксического анализа. Рассмотрены задачи построения семантической сети предложения, вопросно-ответного поиска, извлечения определений и авторских терминов из текстов научных публикаций. Представлены экспериментальные исследования методов вопросно-ответного поиска и ранжирования ответов вопросно-ответных систем. Предложен метод извлечения определений и авторских (т.е. определяемых в тексте) терминов из научных публикаций. Экспериментально показано, что использование разработанных в диссертации методов семантического и семантико-синтаксического анализа повышает качество решения этих задач.

Заключение содержит основные результаты, полученные в работе.

В приложении 1 представлены характеристики реализованного в ходе работы программного обеспечения для семантического и семантико-синтаксического анализа, представлены требования к системам анализа текстов на естественном языке, описана реализованная архитектура, которая позволяет выполнить эти требования.

В приложении 2 представлена структура семантического словаря, используемого в разработанном методе определения ролевых структур высказываний.

В приложении 3 описаны эксперименты с обучаемым синтаксическим анализатором.

В приложении 4 представлен ряд примеров, на которых демонстрируется работа системы семантико-синтаксического анализа текстов.

Научная новизна работы заключается в следующем:

- Разработан новый метод автоматического определения ролевых структур высказываний, основанный семантическом словаре и 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.

Теоретическая значимость диссертации состоит в создании и экспериментальном исследовании новых методов интеграции и взаимодействия

синтаксического и семантического видов анализа текстов на естественном языке.

Практическая ценность диссертации состоит в том, что разработанные в ней методы семантического и семантико-синтаксического анализа являются основой для решения многих прикладных задач информационного поиска и обработки текстов на естественном языке: вопросно-ответного поиска, извлечения информации и знаний из текстов, автоматического реферирования и др.

Замечания к диссертации:

1. При описании алгоритмов семантического и семантико-синтаксического анализа следовало воспользоваться теоретико-множественными методами, что позволило бы представить алгоритмы более формально и упростить их восприятие. Следовало ввести отдельную нумерацию для алгоритмов.
2. В формуле (6) на странице 83 допущена опечатка, изменяющая смысл формулы: вместо символа $\notin$ используется символ $\in$.
3. В задаче распределения ролей (гл.2, п.2.2.5) вместо термина «функция стоимости» при постановке задачи оптимизации следовало использовать другой термин, например, «функция веса», поскольку в диссертации ставится задача не минимизации, а максимизации совокупного веса варианта распределения ролей.
4. В диссертации недостаточно описаны разработанные программные средства. Следовало более подробно описать архитектуру, входные-выходные данные, модули и компоненты разработанных систем.

Приведенные замечания не снижают общей положительной оценки диссертационной работы.

Научные результаты диссертации, выносимые на защиту, являются новыми и получены диссертантом лично. Диссертация А.О. Шелманова является законченной научно-исследовательской работой, выполнена на высоком научном уровне. Сформулированные положения подкрепляются экспериментальными исследованиями. Автореферат достаточно полно отражает содержание диссертации.

Результаты диссертационной работы обоснованы и докладывались на четырех российских и международных конференциях. По теме исследования А.О. Шелмановым 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.

Полученные в диссертации результаты были внедрены в четырех организациях:

- «ООО Национальный цифровой ресурс «Руконт».
- «ООО Научно-издательский центр ИНФРА-М».
- «ЗАО РосИнтернет технологии».
- «Федеральное государственное бюджетное учреждение науки Институт системного анализа РАН».

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

Результаты могут быть также использованы в ФИЦ ИУ РАН, ИПУ РАН, ИПС РАН, МИЭМ-ВШЭ, ИСП РАН, а также в других научных учреждениях и коммерческих организациях, где ведутся исследования и разработки в области информационного поиска и анализа текстов на естественном языке.

Диссертация соответствует критериям, установленным Положением о присуждении ученых степеней: в ней содержится решение задачи, имеющей существенное значение для теоретических основ информатики, она написана автором самостоятельно, обладает внутренним единством и содержит новые научные результаты, полученные автором лично. В диссертации приведены рекомендации по использованию полученных результатов. Работа полностью соответствует требованиям ВАК, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а ее автор Шелманов Артем Олегович заслуживает присуждения ему искомой ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Отзыв обсужден и одобрен на семинаре лаборатории №11 «Методов интеллектуализации дискретных процессов и систем управления» Федерального государственного бюджетного учреждения науки Института проблем управления им. В.А. Трапезникова Российской академии наук 2 сентября 2015г., протокол № 7/2015.

Председатель семинара доктор технических наук, профессор

(О.П.Кузнецов)

Адрес организации: 117997, Москва, ул. Профсоюзная, д. 65
Телефон: +7 495 334-89-10
Электронная почта: snv@ipu.ru