ОТЗЫВ

официального оппонента на диссертационную работу

Шелманова Артема Олеговича

"Исследования 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"

Большие объемы поступающей текстовой информации требуют разработки все большего количества подходов, основанных на смысловой (семантической) обработке текстов. Проведение семантического анализа текстов обычно базируется на результатах синтаксического анализа. Оба вида автоматического анализа представляют собой сложные процедуры, и обычно проводятся отдельно, хотя во многих сложных случаях анализа для построения правильной синтаксической структуры предложения необходимы знания о его семантике. Поэтому актуальным является разработка интегрированных подходов к анализу предложения, совмещающих два вида анализа, так называемый семантико-синтаксический анализ.

В работе Шелманова Артема Олеговича представлен 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 исправлять ошибки в ролевой структуре высказывания. Метод позволяет 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.

Диссертационная работа А.О. Шелманова состоит из введения, четырех глав, заключения, списка сокращений и списка использованной литературы, включающего 178 названий.

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 обзор моделей представления синтаксической структуры и семантического представления предложений. Также в первой главе 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.

Вторая глава посвящена разработанным методам семантического и семантико-синтаксического анализа текстов на естественном языке. Алгоритм построения семантической структуры предложения на выходе порождает так называемую ролевую структуру предложения. Алгоритм работает на основе системы правил, и состоит из следующих пяти этапов: поиск предикатных слов; поиск аргументов; назначение семантических ролей найденным аргументам в соответствии с каждой найденной словарной статьей предиката; выбор наилучшего распределения семантических ролей по аргументам для заданной словарной статьи с учетом ограничения единственности ролей путем решения оптимизационной задачи о назначениях; выбор наилучшей словарной статьи предиката и соответствующего ей набора семантических ролей.

Предложенный подход к интегрированному семантико-семантическому анализу состоит в следующем. На первом шаге выполняется первоначальный синтаксический анализ предложения. Он может выполняться любым анализатором, который строит синтаксическое дерево зависимостей. Далее выполняется поиск семантических атрибутов предиката, в том числе и тех, которые не поддерживаются ранее полученной синтаксической структурой, т.н. дополнительные аргументы.

После того, как все дополнительные аргументы всех предикатных слов предложения найдены, для каждого дополнительного аргумента определяются исправления синтаксического дерева (добавление/удаление связей), которые необходимо совершить, чтобы синтаксическая структура соответствовала новой семантической структуре предложения, с добавленными семантическими

аргументами. Для определения набора дополнительных аргументов, для которых необходимо внести изменения в синтаксическую структуру предложения, используются два метода: частота совместной встречаемости признаков слов или ансамбль бинарных классификаторов. Исправленное синтаксическое дерево заново подается на семантический анализ.

В **третьей главе** представлены результаты экспериментальных исследований разработанных методов семантического и семантико-синтаксического анализа на размеченных корпусах русскоязычных текстов. Показано, что разработанный метод семантико-синтаксического анализа позволяет повысить полноту установления синтаксических связей. Приведены примеры удачного и ошибочного исправления синтаксической структуры предложения.

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. Такие задачи включают в себя: построение семантической сети предложения для реляционно-ситуационной модели текста, вопросно-ответный поиск в метапоисковой системе, автоматическое извлечение определений и авторских терминов из текстов научных публикаций.

В **заключении** описаны основные результаты диссертационной работы и направления дальнейших исследований. Основными новыми научными результатами, полученными в диссертации являются следующие:

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;

В качестве замечаний к тексту диссертации следует отметить следующее:

1. Тестирование подхода выполнено на неопубликованных данных (семантическая разметка), что затрудняет проведение сравнительного анализа в будущих работах других исследователей,

2. На странице 87 делается ссылка на «сильные» правила, описанные в разделе 2.2.3, но в самом разделе 2.2.3 перечисляемые правила не названы "сильными", что усложняет понимание текста,

3. Обнаружен ряд опечаток (стр. 88, 89)

Указанные недостатки не являются принципиальными и не умаляют достоинств диссертации.

Всего по теме диссертации 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. Опубликованные работы полностью отражают результаты диссертации.

Автореферат соответствует диссертации, отражает её содержание и дает представление об актуальности темы, целях, задачах, объекте и методах исследования, научной новизне, практической ценности, реализации, апробации, объеме, кратком содержании и результатах работы.

Методы проведения семантического и семантико-синтаксического анализа, метод построения семантической сети предложения, метод ранжирования сниппетов для вопросно-ответного поиска, а также метод извлечения определений и определяемых терминов реализованы и внедрены в поисковые и поисково-аналитические системы Exactus, Exactus Expert, а также в коммерческие электронные библиотечные системы РУКОНТ и Znanium.com.

Исходя из вышеизложенного, можно утверждать, что диссертация Шелманова А.О. на соискание ученой степени кандидата технических наук является законченной научно-квалификационной работой, в которой содержится описание новой модели интегрированного семантико-синтаксического анализа текстов на русском языке.

Считаю, что диссертационная работа Шелманова Артема Олеговича отвечает всем требованиям Положения ВАК о порядке присуждения ученых степеней, а её автор заслуживает присуждения ему учёной степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Кандидат физико-математических наук,
ведущий научный сотрудник
Научно-исследовательского
вычислительного центра
Московского государственного университета
им. М.В. Ломоносова

Лукашевич Наталья Валентиновна

21.09.2015

Контактные данные:
E-mail: louk_nat@mail.ru
Тел. +7(926)1446163
Адрес: Москва, Ленинские горы, 1, стр. 4

Подпись Лукашевич Н.В. заверяю

Директор НИВЦ МГУ,
проф. Тихонравов А.В.