

ОТЗЫВ

официального оппонента на диссертацию Большаковой Светланы Анатольевны на тему: «Совершенствование методов компьютерной обработки текстовой информации в аспекте задач, связанных с омонимией и синонимией», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки)

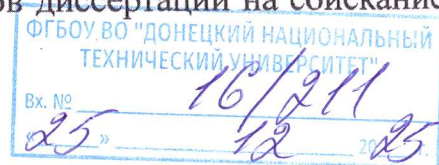
Актуальность темы исследования. Компьютерная обработка текстов сталкивается с фундаментальной проблемой – неоднозначностью языковых единиц, которая проявляется на всех уровнях языка и выражается в явлениях омонимии и синонимии. Рост объемов информации требует автоматизированной обработки текстовых данных, где разрешение многозначности является критически важным этапом.

Однако сложность русского языка создаёт значительные препятствия для полной автоматизации процессов, поскольку наблюдается широкое распространение различных видов омонимии, множество словоформ имеет неоднозначные грамматические характеристики, а сложные синтаксические конструкции требуют глубокого лингвистического анализа.

В связи с этим, тема диссертационного исследования, состоящая в совершенствовании методов компьютерной обработки текстовой информации является актуальной как с научной, так и с практической точек зрения.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации. Обоснованность и достоверность научных положений, выводов и рекомендаций подтверждается методологической базой исследований, основанной на системном подходе, корректностью поставленных задач, достаточным объемом и согласованностью результатов теоретических и экспериментальных исследований, их сопоставлением с данными других авторов, представленными в известных работах, посвященных рассматриваемой тематике.

По теме диссертационного исследования опубликовано 17 работ, в том числе 2 статьи в рецензируемых научных изданиях, рекомендованных ВАК РФ для опубликования основных научных результатов диссертаций на соискание



ученой степени кандидата и доктора наук. Результаты диссертационной работы докладывались и обсуждались на конференциях и научных семинарах. Имеется свидетельство о государственной регистрации программы для ЭВМ.

Содержание и структура работы. Диссертационная работа содержит 172 страниц машинописного текста и состоит из введения, четырех разделов, заключения, списка литературы из 121 источника на 15 страницах и 7 приложений.

В первом разделе представлен достаточно полный обзор существующих технологий обработки текста, освещены современные подходы к работе с синонимией и омонимией, отражены методы автоматизированной обработки текстовой информации.

Во втором разделе представлено описание структур словаря словоформ, описано представление элементов словаря в виде префиксного дерева и приведен алгоритм индексирования.

В третьем разделе содержится описание разработанных автором методов и алгоритмов снятия омонимии в русскоязычных текстах и их программной реализации. Автором описана разработанная база продукционных правил для снятия частеречной омонимии и алгоритм применения правил снятия омонимии для частеречной омонимии предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное.

В четвертом разделе в рамках разработки единой системы обработки и анализа текстовой информации предложены алгоритмы семантической обработки текста: метод лексической адаптации текста с помощью синонимических замен; метод автоматического разбиения сплошного текста на абзацы как семантически однородные фрагменты; автоматического создания элементов плана за счет использования словаря отглагольных существительных. Приведены результаты численных исследований эффективности предложенных методов и алгоритмов.

Диссертационная работа в целом достаточно хорошо структурирована, написана технически грамотным и доступным языком. Текст изложен последовательно.

Содержание автореферата отражает основные положения и выводы диссертационной работы.

Замечания. Выполненные исследования имеют достаточно высокий научный уровень, однако следует обратить внимание на некоторые замечания:

1. В целом работа базируется на нескольких авторских алгоритмах обработки текстов на естественном языке, построенных на базе различных эвристик. При этом автор, хотя и упоминает в обзорной части диссертации современные подходы к интеллектуальному анализу текстов (NLP) на базе моделей машинного обучения (МО), в частности – больших языковых моделей, однако в тексте диссертации не приводит оценку или аргументированное обоснование эффективности предложенных им методов и алгоритмов в сравнении с методами на базе МО, образующими мейнстрим в современных задачах NLP.

2. Большая часть предложенных автором методов и алгоритмов имеет эвристический характер, при этом в работе уделено недостаточно внимания обоснованию качества предложенных решений. В частности, не приводится доводов в пользу правильности предложенного алгоритма разбиения текста на семантически однородные фрагменты (абзацы) (в явном виде не указывается используемая метрика семантической однородности). Методы снятия омонимии проверяются на тестовых примерах, но из текста диссертации не вполне ясно каким образом отбираются эти примеры и чем обуславливается их количество и состав. В частности, в диссертации не указано – осуществлялась ли статистическая оценка результатов тестирования при использовании случайного отбора тестовых данных (разделы 3.4, 3.5, 3.6).

3. Блок-схемы алгоритмов и различные структурные схемы представлены в нестандартных нотациях, что затрудняет их понимание. В целом, для более точного, однозначного графического изображения рассматриваемых в работе структур автору следовало бы использовать стандартизованные графические нотации – IDEF, UML, DFD и т.п.

4. В тексте встречаются повторы идентичной мысли в одинаковых формулировках. Например, «в самом авторитетном из корпусов русского языка, НКРЯ, к подкорпусу с вручную снятой омонимией относится только 1% от общего числа вхождений» на сс.48 и 52., «Возможны случаи, когда предложение может нести более одного смысла.» на сс.52 и 53.

5. Множество знаков препинания Р, введенное на с.52, следует определять как упорядоченное, в противном случае говорить об «отрезке» между элементами множества некорректно.

6. В тексте встречается неаккуратное использование терминологии: например, в разделе 3.1 в одном контексте используются термины «массив» (очевидно, в смысле структуры данных) и «список». При этом последний термин, судя по всему, используется в общем смысле, т.е. не в смысле структуры данных с установленным отношением предшествования на множестве элементов.

7. Из раздела 2.1, описывающего структуру компонентов использованного в работе словаря, не вполне ясно – что было адаптировано соискателем относительно исходного словаря словоформ М. А. Хагена «Полная парадигма. Морфология». Текст организован таким образом, что описание структуры исходного словаря и дополнений/изменений, внесенных соискателем, выглядит идентичным образом.

8. В отдельных разделах диссертации смешаны описания алгоритмов (имеющих теоретическую ценность) и программная реализация оных (например, в разделе 3.1 приведено излишне подробное описание пользовательского интерфейса программного средства, реализующего авторский алгоритм снятия омонимии).

Перечисленные замечания не оказывают решающего влияния на значимость основных положений диссертационной работы.

Заключение. Диссертация является завершенной научной работой, в которой решена актуальная задача повышения эффективности обработки и анализа текстовой информации в контексте решения задач снятия омонимии и применения способов лексической адаптации. По области исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки) по пп. 3, 4, 5, а также удовлетворяет требованиям пп. 9, 10, 11, 13, 14 Положения о присуждении ученых степеней, утвержденного

постановлением Правительства РФ от 24 сентября 2013 г. № 842 с изменениями (редакция от 16.10.2024). Соискатель Большакова Светлана Анатольевна заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки).

Официальный оппонент:

Доктор технических наук по специальности 05.13.10 – «Управление в социальных и экономических системах» (технические науки), доцент, главный научный сотрудник Института информатики и математического моделирования им. В. А. Путилова – обособленного подразделения Федерального государственного бюджетного учреждения науки Федерального исследовательского центра  центр Российской академии наук»

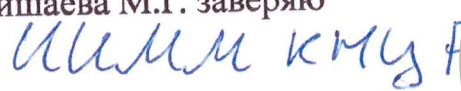
Шишаев Максим Геннадьевич

Я, Шишаев Максим Геннадьевич, даю согласие на автоматизированную обработку моих персональных данных, указанных в отзыве, и размещение их на сайте ФГБОУ ВО «ДонНТУ».

03.12.2025 г.

Шишаев Максим Геннадьевич

Подпись Шишаева М.Г. заверяю

уч. сек.   

Институт информатики и математического моделирования им. В. А. Путилова – обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Кольский научный центр Российской академии наук»

Адрес: 184209, Мурманская область, г. Апатиты, ул. Ферсмана, д. 24А

Тел.: 815-55-74-050

Эл. почта: iimm.administration@ksc.ru