На правах рукописи

Чистова Елена Викторовна

Методы анализа риторической структуры текстов на русском языке

Специальность 1.2.1 —
«Искусственный интеллект и машинное обучение»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2025

Работа выполнена в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН).

Научный руководитель: доктор технических наук, доцент
Смирнов Иван Валентинович

Официальные оппоненты: **Лукашевич Наталья Валентиновна,**
доктор технических наук,
Лаборатория анализа информационных ресурсов Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова,
ведущий научный сотрудник

Ильвовский Дмитрий Алексеевич,
кандидат технических наук,
Международная лаборатория интеллектуальных систем и структурного анализа Национального исследовательского университета «Высшая школа экономики»,
научный сотрудник

Ведущая организация: ФГАОУ ВО Московский физико-технический институт (национальный исследовательский университет)

Защита состоится 19 февраля 2026 г. в _____ часов на заседании диссертационного совета 24.1.224.03 при ФИЦ ИУ РАН по адресу: 119333, Россия, г. Москва, ул. Вавилова, 42.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН и на официальном сайте `https://www.frccsc.ru`.

Автореферат разослан « » _____ 2025 года.

Ученый секретарь
диссертационного совета
24.1.224.03,
к.т.н. И.А. Рейер

Общая характеристика работы

Актуальность темы. Дискурсивный анализ является одной из ключевых задач в области анализа текстов на естественном языке. В эпоху стремительно растущих объемов цифровой информации возникает острая необходимость в создании эффективных методов для автоматизированного анализа связных текстов. Теория риторических структур моделирует дискурс как связную структуру (*риторическую структуру*), включающую все высказывания внутри текста любого объёма. Анализ дискурсивной структуры в рамках Теории риторических структур открывает новые возможности для решения прикладных задач анализа текста, особенно в областях, связанных с когнитивными аспектами дискурса, таких как анализ аргументации, определение основной идеи и взаимосвязей между сложными высказываниями в пределах связного повествования. В развитие моделей и методов риторического дискурсивного анализа внесли вклад W. Mann, S. Thompson, D. Marcu, G. Hirst, C. Braud, A. A. Кибрик и другие.

Тексты играют фундаментальную роль в передаче информации и формировании восприятия у аудитории. Научные статьи, газетные публикации, юридические документы и рекламные материалы — все эти жанры письменного дискурса структурированы таким образом, чтобы максимально эффективно доносить информацию до целевой аудитории или воздействовать на мнение читателя. Особенности риторической структуры текста существенно влияют на его восприятие, помогая читателю лучше понимать логические связи и оценивать убедительность аргументов. Это делает анализ риторической структуры важнейшим инструментом для более глубокого понимания механизмов создания и интерпретации текстов.

Современные системы обработки текста нуждаются в повышении качества анализа связных и сложно структурированных текстов. Например, в задачах классификации мнений необходимо не только выделять ключевые высказывания, но и анализировать риторическую целостность излагаемого материала, что требует выделения в тексте дискурсивной структуры. Риторические отношения, выделяемые в рамках дискурсивного анализа, также играют ключевую роль в задачах анализа аргументации, таких как выделение аргументов, анализ их убедительности или обоснованности, построение структуры аргументации внутри или между документами. Анализ риторической структуры позволяет улучшить качество решения других задач дискурсивного анализа, в том числе разрешения анафоры и кореференции, где информация о природе дискурсивных связей между простыми высказываниями помогает более точно оценивать, к каким упомянутым ранее объектам и событиям отсылаются местоимения или другие кореференты. Это значительно повышает точность интерпретации текста, что критически важно для создания более совершенных систем обработки естественного языка.

На сегодняшний день значительная часть исследований в области риторического анализа посвящена дискурсивным феноменам в английском языке, из-за чего применимость связанных с риторическим анализом методов ограничена для других языков, включая русский. Однако синтаксические, стилистические и дискурсивные особенности русского языка требуют разработки специализированных методов и алгоритмов анализа текстов. Изучение риторических структур в русскоязычном дискурсе представляет собой важную научную задачу. Русскоязычные тексты многих жанров (научные, публицистические, художественные и др.) имеют свои особенности организации информации, которые могут существенно отличаться от англоязычных аналогов. Например, в некоторых жанрах могут использоваться более сложные синтаксические конструкции и более длинные предложения, что создает дополнительные сложности для анализа текста и построения его риторической структуры. Разработка методов, учитывающих эти особенности, является важным шагом для создания качественных систем анализа текстов на русском языке.

В условиях ограниченных ресурсов разметки данных на некотором языке применяются методы кросс-языкового переноса методов анализа текста, основанных на глубоком обучении. Кросс-языковой анализ риторических структур является недостаточно исследованной областью ввиду отсутствия достаточных для глубокого обучения объёмов параллельных данных полнотекстовой риторической разметки и разнородности интерпретации теории риторических структур при разработке размеченных корпусов текстов для разных языков. Однако кросс-языковые методы особенно перспективны для улучшения качества риторического анализа, поскольку дискурс является наименее специфической для разных языков языковой единицей.

Экспертная разметка риторических структур является задачей, требовательной к навыкам экспертов и жёсткости устанавливаемых формальных ограничений. В зависимости от жанрово-стилистических особенностей материалов, при разработке каждого корпуса эксперты устанавливают специфические для него отношения и их строгие определения, которые могут отличаться по детализированности или ограничениям на составляющие от отношений в других корпусах, что ведет также к различиям в определениях элементарных дискурсивных единиц. Ограниченность и теоретическая разнородность размеченных данных затрудняет построение универсальных систем риторического анализа не только для множества языков, но и для множества жанров внутри одного языка. Разработка универсального риторического анализатора требует создания методов, позволяющих эффективно использовать все доступные данные, вне зависимости от языка, жанра и набора риторических отношений, принятых для разметки конкретного корпуса. Такие методы могут позволить увеличить объем

материала для глубокого обучения и улучшить их обобщаемость на разные жанры дискурса.

Таким образом, разработка методов риторического анализа способствует более эффективной обработке связных текстов и решению прикладных задач. Важным направлением исследований в этой области является анализ обобщаемости автоматического разбора риторической структуры между языками и жанрами.

Целью данной работы является разработка методов дискурсивного анализа текстов на русском языке в рамках теории риторических структур на основе данных экспертной риторической разметки, а также применение риторического анализа для решения прикладных задач обработки естественного языка.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать методы риторического дискурсивного анализа текстов на естественном языке.
2. Разработать методы риторического анализа текстов на русском языке.
3. Исследовать возможности кросс-языкового обобщения полнотекстового риторического анализа.
4. Разработать методы дискурсивного анализа на основе разнородных данных риторической разметки разных жанров.
5. Разработать методы решения прикладных задач с использованием риторических структур. Оценить влияние разработанных методов риторического анализа на качество решения прикладных задач.

Научная новизна:

1. Впервые разработаны методы риторического анализа текстов на русском языке, включая анализ локальных и полнотекстовых дискурсивных структур.
2. Впервые исследованы возможности кросс-языковой адаптации дискурсивного анализа на материале большого параллельного корпуса дискурсивной разметки.
3. Предложен метод риторического дискурсивного анализа, позволяющий достичь качества полнотекстового анализа риторической структуры на русском и английском языках, превышающего качество предыдущих систем.
4. Впервые предложен метод реализации риторического анализа при помощи глубокого обучения на материалах разнородной риторической разметки.
5. Разработан метод классификации текстов с учетом риторических структур, показана его эффективность в задачах классификации тональности и аргументации.

6. Разработан метод разрешения кореференции с учётом риторической структуры.
7. Разработан метод построения структур аргументации на основе риторических структур. Экспериментально показано, что использование нескольких вариантов дискурсивной структуры при обучении анализатора аргументации улучшает качество построения структур аргументации в рассуждениях на русском и английском языках.

Практическая значимость. Разработанные в рамках диссертации методы риторического анализа текстов на русском языке реализованы в открытом дискурсивном анализаторе[1]. Разработанные методы риторического анализа текстов на русском языке являются основой для классификации, разрешения кореференции, построения структур аргументации и других прикладных задач обработки текстов на естественном языке. Разработанное программное обеспечение, реализующее методы анализа риторической структуры текстов на русском языке, внедрены в программные средства лингво-статистического и психолингвистического анализа текстов ООО «РИ ТЕХНОЛОГИИ», что подтверждается справкой об использовании.

Соответствие паспорту научной специальности. В соответствии с формулой специальности 1.2.1 «Искусственный интеллект и машинное обучение» в работе выполнены создание и исследование методов анализа риторических структур, разработаны программные средства автоматизации извлечения риторических структур из текстов на естественном языке. Работа соответствует следующим пунктам паспорта специальности: пункту 4 «Разработка методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных», пункту 5 в части «Методы и технологии поиска, приобретения и использования знаний и закономерностей, в том числе – эмпирических, в системах искусственного интеллекта», пункту 7 в части «Разработка специализированного математического, алгоритмического и программного обеспечения систем искусственного интеллекта и машинного обучения».

Методология и методы исследования. Для решения поставленных задач используются методы компьютерной лингвистики, машинного обучения, проверки статистической значимости полученных результатов, инженерии программного обеспечения.

Основные положения, выносимые на защиту:

1. Методы автоматического риторического дискурсивного анализа текстов на русском языке, позволяющие выделять поверхностные и полнотекстовые дискурсивные структуры.

[1] `https://github.com/tchewik/isanlp_rst`

2. Метод риторического дискурсивного анализа текстов на основе глубокого обучения, позволяющий использовать при обучении разнородные данные риторической разметки.
3. Методы кросс-языкового и кросс-жанрового риторического анализа текстов, отличающиеся от аналогов тем, что позволяют извлекать риторическую структуру текстов нескольких жанров и стилей на разных языках при помощи единой модели глубокого обучения.
4. Метод классификации текстов, использующий риторические отношения в дискурсивной структуре для определения основной идеи текста.
5. Метод разрешения кореференции в текстах на русском языке, отличающийся использованием автоматического анализа риторической структуры текста для более точного определения расстояния между кореферентными упоминаниями в соответствии с фокусом внимания автора.
6. Метод автоматического извлечения структуры аргументации в тексте, отличающийся построением структуры аргументации поверх риторической структуры, что позволяет улучшить качество анализа аргументации при снижении требований к объёму размеченных обучающих данных.

Достоверность результатов подтверждена экспериментальными исследованиями разработанных методов и апробацией результатов на тематических научных конференциях и внедрением в системы анализа текстов. Результаты работы согласуются с результатами, полученными другими исследователями.

Апробация работы. Основные результаты работы докладывались на следующих конференциях:

1. Международная конференция «Диалог 2019», (Россия, Москва, июнь 2019 г.).
2. Workshop on Discourse Relation Parsing and Treebanking 2019, (США, Миннеаполис, июнь 2019 г.).
3. Мультиконференция по проблемам управления МКПУ, (Россия, Геленджик, сентябрь 2019 г.).
4. Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, (Россия, Москва, октябрь 2020 г.).
5. Международная конференция «Диалог 2022», (Россия, Москва, июнь 2022 г.).
6. Международная конференция «Диалог 2023», (Россия, Москва, июнь 2023 г.).
7. The 61st Annual Meeting of the Association for Computational Linguistics, (Канада, Торонто, июль 2023 г.) (A* по рейтингу CORE).

8. The 62nd Annual Meeting of the Association for Computational Linguistics, (Таиланд, Бангкок, август 2024 г.) (A* по рейтингу CORE).

Личный вклад. Исследования, изложенные в работах [3; 4; 1; 2], выполнены соискателем самостоятельно; разработана программа для ЭВМ [10]. В коллективных публикациях [5—9] автором разработаны методы анализа риторических структур в текстах на русском языке а также их приложений в задачах классификации и разрешения кореференции; в том числе предложены идеи методов, реализованы экспериментальные исследования, результаты оформлены в виде публикаций и научных докладов. В работах [5; 6] автором разработаны и оценены методы классификации риторических отношений, а также проведён признаковый анализ риторических отношений в текстах на русском языке. В работе [7] соискателем разработаны и оценены методы анализа риторической структуры в текстах на русском языке, включая методы классификации риторических отношений, дискурсивной сегментации и построения риторической структуры. Работы [8; 9] описывают приложения анализа риторических структур в задачах классификации и разрешения кореференции в текстах на русском языке, предложенные, разработанные и экспериментально оцененные соискателем.

Грантовая поддержка. Научные исследования в рамках диссертационной работы поддержаны грантом РФФИ №17-29-07033 офи_м «Модели и методы дискурсивного и сюжетного анализа текстов для решения задач интеллектуальной обработки и понимания текстов, естественно-языковой коммуникации», проектом Министерства науки и высшего образования Российской Федерации №075-15-2020-799 «Методы построения и моделирования сложных систем на основе интеллектуальных и суперкомпьютерных технологий, направленные на преодоление больших вызовов», научной программой Национального центра физики и математики, направление № 9 «Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах». Результаты, представленные в Главе 5, были получены в рамках проекта Министерства науки и высшего образования Российской Федерации № 075-15-2024-544 «Математические модели и численные методы как основа для разработки робототехнических комплексов, новых материалов и интеллектуальных технологий конструирования».

Содержание работы

Во **введении** обоснована актуальность диссертационной работы, подчёркивается важность риторического анализа в анализе текстов на естественном языке и отсутствие методов риторического анализа для русского языка, определён предмет исследования. Сформулирована цель

исследования — разработка методов риторического анализа текстов на русском языке, а также применение риторического анализа для решения прикладных задач обработки естественного языка. Поставлены задачи работы, обоснована научная новизна и практическая значимость, приведены положения, выносимые на защиту.

Первая глава посвящена теоретическим основам анализа риторических структур в текстах на естественном языке. В первом разделе указаны основные сведения о дискурсе как языковой единице и моделях представления дискурса в обработке естественного языка. Рассмотрены различные подходы к формальному описанию дискурса, такие как локальные структуры, диктуемые коннекторами, графовые модели и иерархическая разметка в виде дерева составляющих.

Во втором разделе главы приведены сведения о Теории риторических структур (ТРС), предложенной Уильямом Манном и Сандрой Томпсон в 1980-х годах. ТРС рассматривает текст как иерархическое дерево дискурсивных единиц (ДЕ), связанных риторическими отношениями (рисунок 1). Эти отношения определяют логику организации текста, обеспечивая его связность и структурную целостность. В рамках теории выделяют различные классы риторических отношений между ДЕ, включая отношения связности, контраста, причинно-следственные связи, отношения времени и атрибуцию. Описана концепция *нуклеарности* — деления дискурсивных единиц на ядерные (ядро, N) и вспомогательные (сателлит, S) в зависимости от коммуникативной цели и отношения к основной мысли текста. Подчёркивается сложность и субъективность экспертной разметки риторических структур.

Третий раздел первой главы посвящен обзору методов автоматического риторического анализа. Рассматриваются ключевые этапы этого процесса, включая сегментацию текстов на элементарные дискурсивные единицы (ЭДЕ), установление риторических отношений между дискурсивными единицами и построение иерархической структуры текста.

В рамках разбора риторической структуры текста на естественном языке необходимо выделить в тексте элементарные дискурсивные единицы, установить риторические отношения между ними и сформировать иерархическую структуру составляющих, включающую информацию о значимости различных частей текста и типах дискурсивных отношений между ними.

Формально задача построения риторического дерева может быть описана следующим образом. Пусть дан текст D, состоящий из n токенов. Цель системы риторического анализа — построить риторическое дерево T_P, максимально точно соответствующее эталонному дереву T_G. Риторическое дерево можно задать как совокупность риторических отношений, каждое из которых имеет вид:

$$\langle \mathrm{du}_l, \mathrm{du}_r, rel, nuc \rangle,$$

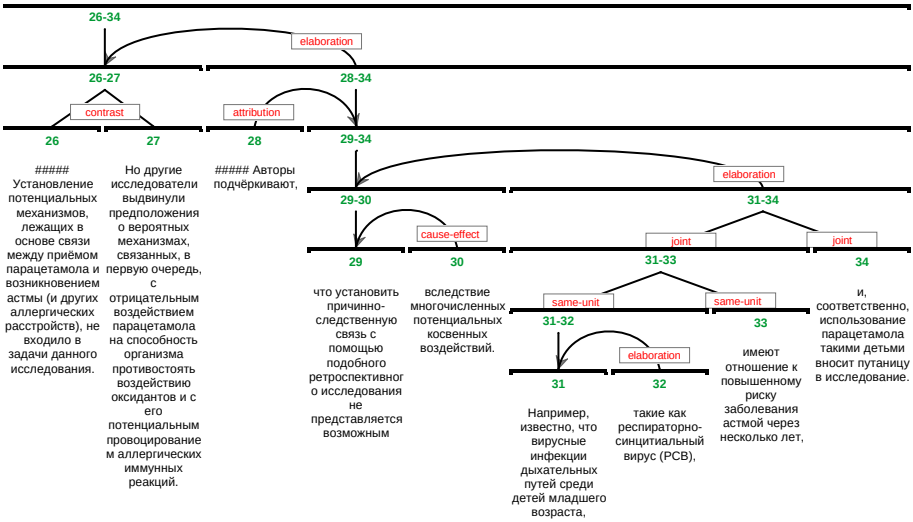

Рисунок 1 — Пример риторического разбора фрагмента текста на русском языке.

где $\mathrm{du}_l = \langle start_l, end_l \rangle$ — положение левой дискурсивной единицы в тексте D, $\mathrm{du}_r = \langle start_r, end_r \rangle$ — положение правой дискурсивной единицы, rel — риторическое отношение, выбранное из фиксированного набора $\mathcal{R}$, определённого экспертами (например, Конкатенация, Детализация, Контраст и т.д.), $nuc \in \{\mathrm{NS}, \mathrm{SN}, \mathrm{NN}\}$ — положение ядра (ядро N может находиться либо в левой, либо в правой единице, либо обе единицы могут быть равноправными).

Приведён обзор методов восходящего и нисходящего риторического анализа. Проанализированы существующие алгоритмы построения дерева, рассмотрены методы с применением классического машинного и глубокого обучения, а также подходы к описанию интерпретируемых признаков дискурсивных единиц. Особое внимание уделено ограничениям восходящих методов разбора, таким как отсутствие обратной связи между модулями и потеря информации о глобальном контексте при разборе локального дискурса. Приведён анализ подходов к нисходящему разбору риторической структуры текста, обеспечивающих не только учёт информации о высокоуровневой структуре дискурса, но и более высокую производительность.

Далее приведён обзор корпусов экспертной риторической разметки для различных языков и жанров, включая англоязычный RST-DT, русскоязычный RRT и другие. Сформулированы актуальные проблемы риторического анализа на основе существующих корпусов риторической разметки: проблема *низкой согласованности разметки внутри корпуса* и проблема *универсализации риторических корпусов*. Далее в разделе описаны ключевые параметры оценки качества риторического разбора:

точность, полнота и F1-мера сегментации, построения структуры, определения ядерности и типов риторических отношений.

В четвёртом разделе первой главы приведен анализ проблемы кросс-языковой обобщаемости методов риторического анализа. Проанализированы существующие подходы к кросс-языковым исследованиям, в том числе сравнительные исследования риторических структур в различных языках, таких как испанский, китайский и баскский, и ограничения, возникающие при адаптации методов для новых языков. Указано, что основной проблемой совмещения различных корпусов для обучения кросс-языковых и кросс-жанровых моделей являются различия в адаптации Теории риторических структур при разметке различных корпусов, включая разные наборы и детализацию риторических отношений и следующие из этого различия в ограничениях на сегментацию ЭДЕ. Оценка кросс-языковой обобщаемости существующих методов ограничена кросс-жанровым переносом моделей и гармонизацией корпусов с автоматическим приведением к общему набору отношений, которое искажает экспертную разметку. Сделан вывод о необходимости создания большого параллельного корпуса риторической разметки для исследования кросс-языковой обобщаемости дискурсивного анализа.

В заключении первой главы подводятся итоги теоретического обзора. В частности, подчёркивается отсутствие методов риторического анализа для русского языка. Далее сформулированы цель и задачи исследования, посвященного разработке методов анализа риторических структур для русского языка, проведению экспериментальных исследований и оценке кросс-языковой и кросс-жанровой обобщаемости разработанных методов, разработке методов дискурсивного анализа на основе разнородных данных риторической разметки множества жанров, разработке методов решения прикладных задач с использованием риторических структур.

Во второй главе описаны предложенные методы решения задачи анализа риторических отношений в русскоязычном письменном дискурсе. Основное внимание уделено проблеме определения типов риторических связей между дискурсивными единицами. Также исследуется задача классификации нуклеарности — определения ядерных и второстепенных элементов риторического дерева в локальном контексте. Для решения задач классификации типов риторических отношений и положения ядра предложены разнообразные лексические, морфологические, синтаксические и семантические признаки дискурсивных единиц.

В экспериментальной части главы представлены результаты применения различных алгоритмов машинного обучения, таких как логистическая регрессия, метод опорных векторов, полносвязные нейронные сети и градиентный бустинг над решающими деревьями, для классификации типов риторических отношений (таблица 1) и положения ядра (таблица 2).

Таблица 1 — Результаты оценки моделей классификации риторических отношений, %

Классификатор	Macro F_1		Micro F_1	
	mean	std	mean	std
NN	49,43	1,52	55,78	1,16
Logistic Regression	50,81	1,06	53,81	1,84
LGBM	51,39	2,18	59,91	1,32
Linear SVM	51,63	1,95	56,61	1,54
L_1 Feature selection + LGBM	51,64	2,22	60,29	1,74
CatBoost	53,32	0,96	60,71	0,81
L_1 Feature selection + CatBoost	53,45	2,19	61,09	1,96
voting((L_1 FS + LGBM), Linear SVM)	54,67	1,80	62,39	1,51
voting((L_1 FS + CatBoost), Linear SVM)	54,67	0,38	62,32	0,41

Предложены ансамбли моделей для классификации риторических отношений, а также «конвейерная» классификация, включающая ансамблирование классификаторов на автоматически отобранных признаках и линейных моделей на всех признаках. Такие методы позволяют более эффективно использовать поднаборы признаков, что необходимо при обучении моделей на данных небольшого объёма и высокой размерности. Для отбора признаков используется логистическая регрессия с L1-регуляризацией. Наилучшие показатели точности и полноты достигнуты при помощи моделей градиентного бустинга, в особенности в сочетании с отбором признаков методом L1-регуляризации и ансамблированием с SVM-классификаторами. Анализ важности признаков выявил доминирующую роль лексических признаков позиций дискурсивных коннекторов, что подтверждает их значимость для успешной классификации риторических связей.

Таблица 2 — Результаты оценки моделей классификации ядерности, %

Классификатор	Macro F_1		Micro F_1	
	mean	std	mean	std
Linear SVM	63,01	0,58	64,20	0,52
NN	63,32	0,88	64,59	0,75
Logistic Regression	63,66	0,37	65,02	0,26
L1 Feature selection + LGBM	67,82	0,86	69,17	0,73
CatBoost	68,03	0,45	69,37	0,36
LGBM	68,81	0,77	70,17	0,67
L1 Feature selection + CatBoost	68,82	0,84	70,31	0,76

Использованная в исследовании версия корпуса RRT 1.0 (первого крупного русскоязычного корпуса, размеченного согласно Теории риторических структур) включает разметку 179 текстов различных жанров. Эксперименты проведены с 11 наиболее репрезентативными классами

риторических отношений в корпусе. Сложности наблюдались при классификации некоторых классов с меньшим количеством обучающих примеров, таких как Сравнение, Свидетельство, Оценка и Фон. Результаты указывают на необходимость дальнейшего совершенствования моделей и увеличения объёма обучающих данных.

В заключении делаются выводы о высокой эффективности разработанных методов для риторического анализа русскоязычного дискурса и их потенциале для применения в дальнейших исследованиях. Полученные результаты использованы в последующих главах при создании риторических анализаторов для русского языка и адаптации большого корпуса риторической разметки GUM с английского на русский язык в целях исследования возможностей кросс-языкового переноса методов риторического анализа на основе глубокого обучения.

Третья глава посвящена разработке методов построения поверхностных риторических структур в текстах на русском языке, предложены ансамбли классических методов машинного обучения и нейросетевых подходов. В исследовании использован корпус RRT версии 2.0, включающий выверенную разметку риторических структур составляющих в новостных и блоговых текстах. Под поверхностным разбором подразумевается восстановление частичных низкоуровневых риторических структур, размеченных в корпусе RRT. Особое внимание уделено задачам дискурсивной сегментации и классификации риторических отношений между ДЕ, также предложен метод нисходящего разбора абзацев документа.

Автоматическая дискурсивная сегментация сводится к задаче разметки последовательности токенов, целью которой является корректная маркировка границ ЭДЕ в тексте. Для дискурсивной сегментации характерна низкая сбалансированность классов, а класс токена относительно принадлежности к границе ЭДЕ зависит от его контекста в предложении. Для решения задачи дискурсивной сегментации в рамках анализа риторических структур в текстах на русском языке предложена модель на основе архитектуры BiLSTM-CRF.

Поверхностный анализ риторической структуры текста требует решения двух задач классификации, а именно разработки *структурного классификатора* и *классификатора отношения и ядерности*. Процесс построения дискурсивной структуры текста из последовательности элементарных дискурсивных единиц можно разбить на два основных этапа. Сначала структурный классификатор оценивает вероятность наличия риторической связи между соседними дискурсивными единицами (с последующим сравнением с порогом τ) и выполняет их объединение при необходимости. Затем классификатор риторических отношений уточняет характер и организацию (ядра) каждой обнаруженной связи. В случае поверхностного анализа, когда итоговая структура документа представлена

в виде леса, использование порогового значения τ даёт возможность рекурсивно формировать риторические деревья внутри документа до тех пор, пока оценка наличия отношений остаётся выше порога. Это гарантирует моделирование только тех риторических отношений, для которых имеется достаточный объём экспертной разметки в обучающем корпусе.

Далее описаны предложенные подходы к классификации пар дискурсивных единиц. Методы на основе признакового описания основаны на описанном ранее исследовании, но с расширенным набором признаков. Методы классификации при помощи глубокого обучения основаны на нейронной архитектуре BiMPM с кодированием входных токенов предобученной языковой моделью; нейронный структурный классификатор также использует признаки нахождения двух ДЕ на одном уровне детализации (предложение, абзац). Для повышения точности классификации предложено ансамблировать модели двух типов с обучением параметров ансамблирования (рисунок 2).

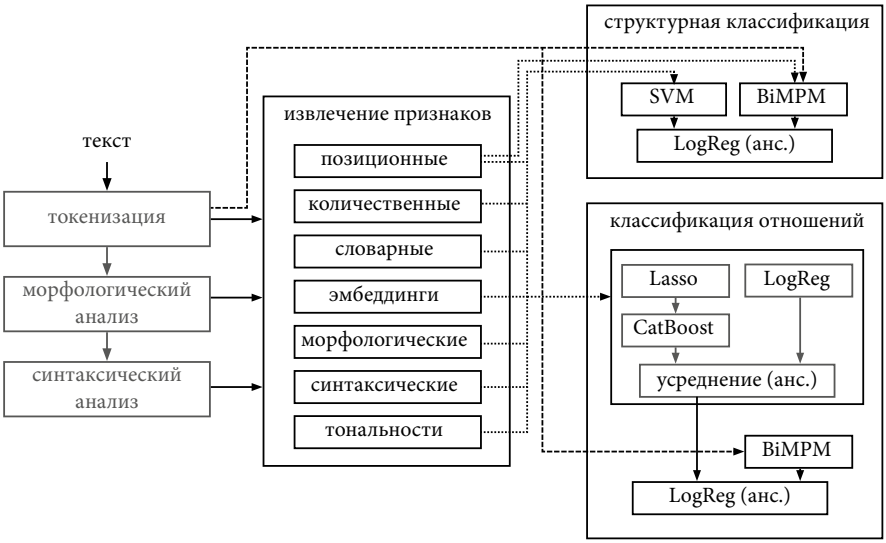

Рисунок 2 — Схема решения задач классификации в риторическом анализаторе.

Предложено два подхода к построению леса риторических структур: жадный восходящий разбором с порогом вероятностей (рис. 3) и его модификация с уточнением структуры абзаца при помощи локального нисходящего разбора с лучевым поиском, основанного на глубоком обучении (рис. 4).

Алгоритм 1: Построение леса риторических структур

Input: Список дискурсивных единиц $[e_1, e_2, ..., e_n]$.
Output: Дискурсивные деревья
$Trees \leftarrow [e_1, e_2, ..., e_n]$ ▷ Инициализация риторических структур
$P_{struct} \leftarrow \emptyset$ ▷ Оценки вероятностей наличия связей
for $i \leftarrow 1$ **to** $n-1$ **do**
  $P_{struct}[i] = f_{struct}(e_i, e_{i+1})$
end
while $|Trees| > 1$ **and** $\exists\ i : P_{struct}[i] > \tau_k$ **do**
  $j = argmax(P_{struct})$
  $DU^* = mergeNodes(j, j+1)$ ▷ Формирование новой ДЕ
  $DU^*.relation = f_{rel}(Trees[j], Trees[j+1])$ ▷ Назначение типа отношения и ядерности
  Заменить $Trees[j]$ и $Trees[j+1]$ на DU^*
  if $j \neq 0$ **then**
    $P_{struct}[j-1] = f_{struct}(Trees[j-1], DU^*)$ ▷ Оценки вероятностей связей с новой ДЕ
  end
  if $j \neq |\ P_{struct}\ |$ **then**
    $P_{struct}[j] = f_{struct}(DU^*, Trees[j+1])$
  end
end
return $Trees$

Рисунок 3 — Жадный алгоритм построения риторического леса

Алгоритм 2 Нисходящий разбор дерева с использованием лучевого поиска

Input: Список ЭДЕ $E = [e_1, e_2, \ldots, e_n]$; ширина луча b; Начальное состояние декодировщика s.
Output: Дискурсивное дерево
$Enc([e_1, e_2, \ldots, e_n]) \leftarrow [h_0, h_1, \ldots, h_m]$ ▷ Состояния кодировщика
$L_d \leftarrow |E| - 1$ ▷ Длина декодирования
$beam \leftarrow$ массив из L_d элементов
$init\_input\_span = [(0, |E|), (0, 0), ..., (0, 0)]$ ▷ Инициализация $L_d - 1$ пустых элементов
$init\_tree = [(0, 0, 0), (0, 0, 0), ..., (0, 0, 0)]$ ▷ Инициализация L_d элементов
$beam[0] = (0, s, init\_input\_span, init\_tree)$ ▷ Инициализация первого элемента луча (вероятность, состояние декодировщика, начальная ДЕ, дерево)
for $t \leftarrow 1$ **to** L_d **do**
  for $(logp, s, input\_span, tree) \in beam[t-1]$ **do**
    $(i, j) \leftarrow input\_span[t-1]$ ▷ Текущая ДЕ для разбиения
    $a, s' \leftarrow$ decoder-step$(s, h_{i,j})$ ▷ a — распред. вер. разбиения
    for $(k, p_k) \in$ top-B(a) **and** $i < k < j$ **do**
      $curr\_input\_span \leftarrow input\_span$
      $curr\_tree \leftarrow tree$
      $curr\_tree[t-1] \leftarrow (i, k, j)$
      if $k > i + 1$ **then**
        $curr\_input\_span[t] \leftarrow (i, k)$
      end if
      if $j > k + 1$ **then**
        $curr\_input\_span[t + j - k - 1] \leftarrow (k, j)$
      end if
      Добавить $(logp + \log(p_k), s', curr\_input\_span, curr\_tree)$ в луч $beam[t]$
    end for
  end for
  Обрезать $beam[t]$ ▷ Оставить топ-B поддеревьев
end for
$(logp^*, s^*, ip^*, S^*) \leftarrow \arg\max_{logp} beam[L_d]$ ▷ S^* – лучшая структура

Рисунок 4 — Алгоритм нисходящего разбора риторического дерева с использованием лучевого поиска

Описаны алгоритмы разбора, а также модель глубокого обучения для анализа риторической структуры составляющих внутри абзаца. Результаты экспериментов демонстрируют значительное повышение качества риторического анализа текстов на русском языке при использовании метода глубокого обучения с лучевым поиском для построения риторических структур на уровне абзаца (таблица 3).

Таблица 3 — Качество поверхностного риторического анализа с использованием разных методов построения локальной структуры; F1, в %.

Уровень дискурса		S	N	R	Full
Предложение	Базовый	58,0	38,9	27,8	27,1
	+ Нисх. разбор абзаца	**68,5**	**50,6**	**38,1**	**37,7**
Абзац	Базовый	49,4	31,0	20,4	20,3
	+ Нисх. разбор абзаца	**59,8**	**38,8**	**27,5**	**27,3**
Документ	Базовый	43,6	27,3	18,0	17,7
	+ Нисх. разбор абзаца	**52,5**	**34,2**	**24,2**	**23,9**

В заключении подчеркивается эффективность комбинации классических и нейросетевых методов машинного обучения для риторического анализа русскоязычных текстов. Резюмированы предложенные методы поверхностного риторического анализа. Сделан вывод, что нисходящий разбор при помощи глубокого обучения открывает новые возможности для автоматического дискурсивного анализа.

В **четвертой главе** приведено описание усовершенствованного метода DMRST для полнотекстового нисходящего разбора риторической структуры текста на естественном языке (рисунки 5–6). Метод основан на гибридном подходе, объединяющем сегментацию ЭДЕ и построение дерева риторических структур с использованием единой архитектуры глубокого обучения. Во втором разделе главы подробно описаны внесённые улучшения в базовую модель DMRST, такие как разработка сегментатора на основе BiLSTM-CRF, использование BiLSTM для локального кодирования токенов в контексте ЭДЕ и модификация механизма динамического взвешивания функций потерь (DWA) для подбора числа предыдущих значений функции потерь при оценке веса текущего значения функции потерь для оптимизации параметров нейронного модуля анализатора.

Третий раздел главы посвящен кросс-языковому анализу. В разделе подробно описывается актуальность анализа кросс-языковой обобщаемости и создания большого параллельного корпуса риторической разметки. Приводятся сведения о создании первого большого параллельного корпуса риторической разметки RRG на основе корпуса GUM разметки 213 текстов

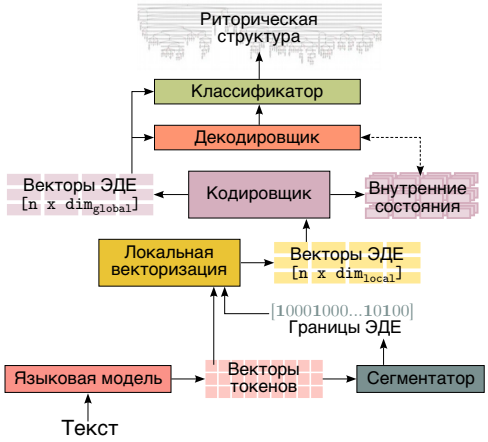

Рисунок 5 — Архитектура DMRST

12 жанров на английском языке: академическая статья (`academic`), биография из Википедии (`bio`), диалог (`conversation`), художественная проза (`fiction`), интервью (`interview`), новости (`news`), форум (`reddit`), официальная речь (`speech`), учебник (`textbook`), видеоблог (`vlog`), путеводитель (`voyage`), инструкция (`whow`). Описаны основные этапы создания корпуса: перевод, сопоставление риторических структур, повышение согласованности и устранение ошибок в разметке текстов на русском языке при помощи описанного в главе 4 классификатора типа и ядерности риторических отношений. В таблице 4 приведены статистики корпусов риторической разметки для русского и английского языков, включая разработанный корпус RRG.

Таблица 4 — Статистики корпусов

Корпус	Жанров	Источ.	Док.	Классов	Токенов на дерево			ЭДЕ	ЭДЕ на дерево	Пар ДЕ
					мин.	макс.	мед.			
RST-DT	1	1	385	41	30	2624	396	21789	56,6	21404
GUM	12	12+	213	27	167	1879	989	26319	123,6	26106
RRT	2	17+	233	24	2	1148	89	28372	11,7	25957
RRG	12	12+	213	27	137	1629	833	25223	118,4	25010

В четвертом разделе рассматривается проблема кросс-жанрового переноса риторического анализа. Предложен метод обучения полнотекстового анализатора на смешении данных из различных корпусов с разными интерпретациями теории риторических структур, позволяющий создавать универсальные риторические анализаторы, адаптированные к разнообразию жанров и стилей. Экспериментальные результаты подтверждают

Алгоритм 3 Гибридный нисходящий риторический разбор

Input: Токены документа $T = \{t_1, t_2, \ldots, t_n\}$, параметры модели θ
Output: Дискурсивное дерево $Tree$.

$S = \arg\max_S P(S \mid T, \theta)$, где $S = \{s_1, s_2, \ldots, s_m\},\ s_0 = 0$
$E = \{e_1, e_2, \ldots, e_m\}$, где $e_i = \{t_{s_{i-1}+1}, \ldots, t_{s_i}\}$ ▷ Сегментация текста на ЭДЕ
$\mathbf{v}_{local}(e_i) = f_{local}\left(emb(t_{s_{i-1}+1}), emb(t_{s_{i-1}+2}), \ldots, emb(t_{s_i})\right)$ ▷ Локальные представления ЭДЕ
$\{\mathbf{v}_{global}(e_i)\}_{i=1}^{m} = f_{global}\left(\{\mathbf{v}_{local}(e_i)\}_{i=1}^{m}\right)$ ▷ Глобальные представления ЭДЕ

Инициализировать стек: $stack \leftarrow [\, e_{1:m} \,]$ и дерево: $Tree \leftarrow \emptyset$
while $stack \neq \emptyset$ **do**
    Извлечь $e_{i:j} \leftarrow pop(stack)$
    if $i = j$ **then**
        $Tree \leftarrow Tree \cup \{e_{i:j}\}$ ▷ Достигнут лист дерева
    else
        $h_t \leftarrow \mathrm{GRU}\big(f_{agg}(\mathbf{v}_{global}(e_i), \ldots, \mathbf{v}_{global}(e_i)), h_{t\text{-}1}\big)$ ▷ Обновление параметров декодировщика
        for $u = i + 1$ **to** $j - 1$ **do**
            $s_{t,u} = \sigma(\, h_t^{\top} W\, \mathbf{v}_{global}(e_u) + b)$ ▷ Оценка точек разбиения
        end
        $k^* = \underset{u \in \{i+1, \ldots, j-1\}}{\operatorname{argmax}} s_{t,u}$
        $Tree \leftarrow Tree \cup \{e_{i:k^*}, e_{k^*+1:j}\}$ ▷ Обновление дерева
        $stack \leftarrow push(stack, e_{i:k^*}, e_{k^*+1:j})$ ▷ Обновление стека
    end
end
for all $e_{i:j} \in Tree : i < j$ **do**
    Определить пару дочерних узлов $(e_{i:k}, e_{k+1:j})$, где k — точка разбиения
    $rel\_nuc(e_{i:k}, e_{k+1:j}) = \operatorname{argmax} P\left(rel\_nuc \mid \mathbf{v}_{global}(e_{i:k}), \mathbf{v}_{global}(e_{k+1:j}), \theta\right)$ ▷ Классификация типа отношения и положения ядра
end

Рисунок 6 — Алгоритм гибридного нисходящего риторического разбора

эффективность метода на двух крупных русскоязычных корпусах, демонстрируя увеличение точности разбора и улучшение обобщаемости моделей.

В пятом разделе главы описаны результаты экспериментальных исследований предложенных методов. Эффективность предложенного метода полнотекстового анализа оценивалась на различных корпусах риторической разметки: RST-DT (английский), GUM (английский) и RRT (русский), а также разработанном в рамках диссертационного исследования корпусе RRG (русский). В таблице 5 приведены оценки качества предложенного метода полнотекстового риторического анализа, обученного на каждом корпусе отдельно. Результаты исследования демонстрируют, что предложенный метод разбора с использованием кодирующей языковой модели `xlm-roberta-large` позволяет повысить

Таблица 5 — Качество риторического анализа, F1, в %

Корпус	Метод	Segm.	S	N	R	Full
RST-DT	SegBot & Top-Down	92,2	62,3	50,1	40,7	39,6
	RPFS	96,3	68,4	59,1	47,8	46,6
	DMRST	96,4	69,8	59,4	49,4	48,6
	+ Cross-translation	96,5	70,4	60,6	51,6	50,1
	DMRST (this work)	97,3 ± 0,1	74,3 ± 0,6	64,1 ± 0,7	53,9 ± 0,5	52,4 ± 0,5
	+ ToNy	**97,9 ± 0,1**	**75,1 ± 0,7**	**64,8 ± 0,7**	**54,5 ± 0,9**	**53,0 ± 0,9**
	+ ToNy + E-BiLSTM	97,8 ± 0,1	74,8 ± 0,5	64,5 ± 0,8	**54,5 ± 0,7**	**53,0 ± 0,7**
GUM v9.1	DMRST (this work)	94,7 ± 0,4	65,0 ± 0,5	54,2 ± 0,5	47,3 ± 0,5	46,4 ± 0,4
	+ ToNy	95,4 ± 0,1	66,4 ± 0,3	55,8 ± 0,5	48,5 ± 0,5	47,6 ± 0,6
	+ ToNy + E-BiLSTM	**95,5 ± 0,1**	**66,9 ± 0,5**	**56,1 ± 0,3**	**48,8 ± 0,4**	**47,9 ± 0,4**
RRT	DMRST (this work)	**92,4 ± 0,3**	**66,5 ± 1,0**	**52,4 ± 1,2**	**45,3 ± 1,0**	**45,3 ± 1,0**
	+ ToNy	**92,4 ± 0,2**	65,4 ± 1,1	51,3 ± 0,6	44,6 ± 0,5	44,5 ± 0,5
	+ ToNy + E-BiLSTM	92,2 ± 0,2	65,9 ± 0,5	51,0 ± 0,7	43,9 ± 1,0	43,8 ± 1,0
RRG	DMRST (this work)	96,3 ± 0,1	65,6 ± 0,3	52,8 ± 0,3	45,1 ± 0,2	44,0 ± 0,3
	+ ToNy	96,7 ± 0,2	**66,6 ± 0,9**	53,0 ± 1,7	45,3 ± 1,7	44,3 ± 1,5
	+ ToNy + E-BiLSTM	**96,9 ± 0,2**	66,5 ± 0,4	**53,3 ± 0,6**	**45,8 ± 0,5**	**44,6 ± 0,4**

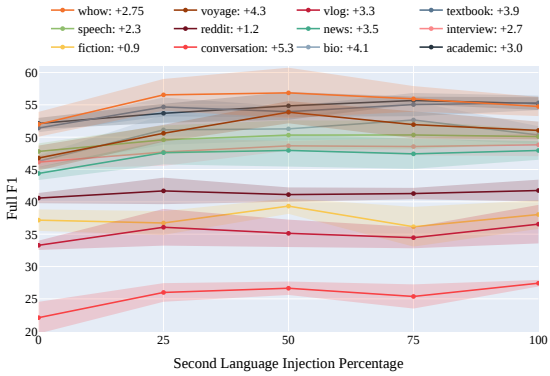

Рисунок 7 — Влияние добавления русского языка в обучающие данные на качество риторического анализа по жанрам

качество дискурсивной сегментации (97.9% Seg F1) и полного разбора риторической структуры (53,0% Full F1) относительно предыдущих методов на бенчмарке риторического анализа для английского языка RST-DT. Качество сегментации текстов на русском языке (корпус RRT) достигло 92,4% Seg F1, качество полного разбора размеченных структур в тестовой части RRT: 45,3% Full F1. На основе предложенного метода полнотекстового анализа разработаны двуязычные модели риторического анализа, демонстрирующие возможности переноса знаний между языками. Исследования кросс-языковой обобщаемости на параллельных данных продемонстрировали перспективность переноса риторического анализа

в условиях ограниченной экспертной разметки на целевом языке (рисунок 7). Двуязычный риторический анализатор показал наилучшее качество анализа на тестовой выборке корпуса RRG (96,8% Seg F1, 45,4% Full F1). В рамках экспериментов с предложенным методом обучения анализатора на смешении непосредственно несовместимых TPC-корпусов проведен анализ совместимости двух больших корпусов риторической разметки для русского языка RRT и RRG. Предложенный метод обучения полнотекстового риторического анализатора на смешанных данных позволил получить жанрово-универсальный полнотекстовый риторический анализатор для русского языка.

В заключении подчеркивается перспективность предложенных методов полнотекстового нисходящего анализа риторических структур для кросс-языкового и кросс-жанрового риторического анализа, а также дискурсивного анализа в условиях ограниченной разметки.

Пятая глава посвящена применению методов анализа риторических структур для решения прикладных задач обработки естественного языка. Во введении указывается важность автоматического анализа риторической структуры текста для решения задач обработки текстов со сложным дискурсом, таких как классификация, разрешение кореференции и анализ аргументации.

В первом разделе, посвящённом классификации текстов, предлагается двухэтапный метод классификации, в котором риторическая структура используется для улучшения классификации текстов современными языковыми моделями. Метод включает классификацию отдельных дискурсивных единиц любым методом классификации текста как последовательности токенов, позволяющим генерировать оценки вероятностей классов, и последующую агрегацию результатов с помощью архитектуры TreeLSTM. Экспериментальные результаты демонстрируют улучшение показателей F1 на наборах данных различных тематик в задачах анализа мнений и анализа аргументации по сравнению с базовым BERT-классификатором (таблица 6). Показано, что учёт дискурсивной структуры позволяет лучше выявлять полярность сложных высказываний, представляющих сложность для базового классификатора.

Второй раздел посвящён задаче разрешения кореференции. В нем описан разработанный метод на основе глубокого обучения, учитывающий риторические признаки, такие как линейное и риторическое расстояния между упоминаниями, а также расстояние до наименьшего общего предка в риторическом дереве. Экспериментальные исследования метода на больших русскоязычных корпусах AnCor и RuCoCo демонстрируют повышение точности разрешения кореференции при использовании риторических признаков по сравнению с базовыми моделями (таблица 7). Признак

Таблица 6 — Оценки качества на кросс-валидации. F1, в %.

Метод	Тип текстов	Маски	Вакцины	Карантин	Mean
		Позиция автора			
BERT	Неэлементарные	59,8 ± 2,7	62,4 ± 3,4	54,5 ± 3,4	58,9 ± 2,3
	Все	60,6 ± 2,6	64,4 ± 2,2	56,4 ± 2,8	60,5 ± 1,9
+ RST-LSTM	Неэлементарные	**61,3 ± 2,7**	**63,4 ± 4,2**	**55,6 ± 2,7**	**60,1 ± 2,3**
	Все	**61,7 ± 2,6**	**65,1 ± 3,0**	**57,5 ± 2,4**	**61,4 ± 1,8**
		Аргументация			
BERT	Неэлементарные	66,4 ± 2,9	**61,7 ± 4,3**	56,4 ± 2,8	61,5 ± 2,2
	Все	66,0 ± 2,4	**62,6 ± 2,7**	57,0 ± 2,3	61,9 ± 1,6
+ RST-LSTM	Неэлементарные	**68,1 ± 2,1**	60,4 ± 3,3	**57,6 ± 2,0**	**62,0 ± 1,3**
	Все	**67,5 ± 1,9**	61,5 ± 2,3	**58,3 ± 2,1**	**62,4 ± 0,9**

расстояния до наименьшего общего предка с антецедентом в полнотекстовом риторическом дереве позволил достичь лучшего качества разрешения кореференции на двух датасетах.

Таблица 7 — Оценка качества разрешения кореференции; F1, в %.

		AnCor					RuCoCo				
		MUC	B3	CEAF	CoNLL	LEA	MUC	B3	CEAF	CoNLL	LEA
Базовый метод		66,0	54,8	52,9	57,9	51,3	83,8	77,3	70,1	77,1	75,0
RRT	Lin	64,4	52,9	51,3	56,2	49,2	**83,9**	**77,5**	70,1	**77,2**	**75,1**
	Rh	**66,4**	**55,2**	**53,8**	**58,5**	**51,7**	**84,1**	**77,8**	**70,6**	**77,5**	**75,6**
	LCA	**66,2**	**55,4**	**53,2**	**58,3**	**51,8**	**84,0**	**77,6**	**70,5**	**77,4**	**75,4**
RRG	Lin	63,5	52,1	51,0	55,5	48,5	83,8	**77,5**	**70,4**	**77,2**	**75,2**
	Rh	65,9	**55,1**	**53,0**	**58,0**	**51,4**	**84,0**	**77,8**	**70,5**	**77,4**	**75,5**
	LCA	65,1	54,2	52,6	57,3	50,4	**83,9**	**77,6**	**70,6**	**77,4**	**75,3**
RRT/G	Lin	65,2	54,0	52,9	57,4	50,3	83,8	**77,4**	70,1	77,1	**75,1**
	Rh	66,0	54,4	52,9	57,8	50,9	**83,9**	**77,6**	**70,5**	**77,3**	**75,4**
	LCA	64,9	53,7	51,8	56,8	50,1	**84,3**	**78,3**	**70,9**	**77,8**	**76,1**

В третьем разделе пятой главы рассматривается использование анализа риторических структур для построения аргументативных структур в текстах-рассуждениях. Предлагается биафинный анализатор аргументации (BAP) и его модификация (DBAP), использующая при построении структуры аргументации зависимости из риторической структуры. Также предложено использовать перефразирования текстов для получения нескольких вариантов риторического разбора. Эксперименты подтверждают эффективность предложенных методов в улучшении качества анализа структуры аргументации (таблица 8). Предложенный метод построения структуры аргументации на основе риторической структуры также позволил эффективно обучить анализатор структур аргументации в неразмеченном тексте на небольшом корпусе текстов-рассуждений на русском и английском языках. Показано, что использование нескольких вариантов разбора риторической структуры позволяет лучше оценивать связи между двумя моделями описания структуры текста.

Таблица 8 — Оценки качества методов BAP и DBAP на корпусе с экспертной сегментацией АДЕ

Метод	Доп. стр.	cc	ro	fu	UAS	LAS
			Английский язык			
BAP	Нет	88,3 ± 4,9	**71,1 ± 5,7**	77,1 ± 4,6	59,1 ± 6,8	52,9 ± 6,3
	Да	88,9 ± 4,7	69,2 ± 3,9	**78,3 ± 4,9**	61,2 ± 5,8	55,1 ± 5,9
DBAP	Нет	**90,3 ± 3,3**	68,8 ± 6,9	77,3 ± 3,2	64,5 ± 6,6*	56,2 ± 5,3*
	Да	89,5 ± 4,3	68,8 ± 7,6	76,5 ± 3,1	**64,6 ± 4,1**	**56,6 ± 3,2**
			Русский язык			
BAP	Нет	**90,5 ± 5,7**	69,3 ± 7,8	78,9 ± 4,2	61,7 ± 6,6	55,2 ± 6,7
	Да	90,3 ± 2,8	66,9 ± 6,9	77,5 ± 4,3	61,6 ± 4,7	53,9 ± 5,7
DBAP	Нет	90,3 ± 5,7	68,9 ± 2,5	**79,8 ± 3,6**	**64,6 ± 5,8**	**58,0 ± 3,6**
	Да	88,3 ± 6,4	**69,9 ± 5,4**	77,2 ± 6,1	**64,6 ± 5,8**	57,0 ± 5,8

В заключении к пятой главе сделан вывод, что анализ риторических структур позволяет повысить эффективность решения прикладных задач обработки естественного языка, таких как классификация текстов, разрешение кореференции и анализ аргументации.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработаны и реализованы методы анализа риторических структур в текстах на русском языке.
2. Проведены экспериментальные исследования методов анализа риторических отношений, поверхностного и полнотекстового риторического анализа текстов на русском языке.
3. Разработаны и реализованы методы кросс-языкового анализа риторических структур, исследованы возможности кросс-языкового обобщения полнотекстового риторического анализа.
4. Разработаны методы реализации дискурсивного анализа на основе непосредственно несовместимых данных риторической разметки разных жанров.
5. Экспериментально показано, что использование риторической структуры текста повышает эффективность решения практических задач анализа естественного языка: классификации текстов, разрешения кореференции, анализа структуры аргументации.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

1. *Чистова Е. В.* Методы анализа риторических структур в текстах на русском языке // Искусственный интеллект и принятие решений — № 4. — 2024. — С. 79–92. **К1**.

2. *Чистова Е. В.* Влияние признаков иерархического дискурса на разрешение кореференции в русском языке // Искусственный интеллект и принятие решений — № 1. — 2025. — С. 95–102. **К1**.

В сборниках трудов конференций

3. *Chistova E.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023. — Association for Computational Linguistics. Toronto, Canada — 2023. — С. 3376–3391. **Scopus, CORE A***.
4. *Chistova E.* Bilingual Rhetorical Structure Parsing with Large Parallel Annotations // Findings of the Association for Computational Linguistics: ACL 2024. — Association for Computational Linguistics. Bangkok, Thailand — 2024. — С. 9689–9706. **Scopus, CORE A***.
5. *Chistova E., Shelmanov A., Kobozeva M., Pisarevskaya D., Smirnov I., Toldova S.* Classification models for RST discourse parsing of texts in Russian // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — 2019. — С. 163–176. **Scopus**.
6. *Chistova E., Kobozeva M., Pisarevskaya D., Shelmanov A., Smirnov I., Toldova S.* Towards the Data-driven System for Rhetorical Parsing of Russian Texts // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. — Association for Computational Linguistics. Minneapolis, MN — 2019. — С. 82–87.
7. *Chistova E., Shelmanov A., Pisarevskaya D., Kobozeva M., Isakov V., Panchenko A., Toldova S., Smirnov I.* RST discourse parser for Russian: an experimental study of deep learning models // Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020. Moscow, Russia — 2021. — С. 105—119. **Scopus**.
8. *Chistova E., Smirnov I.* Discourse-aware text classification for argument mining // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — 2022. — С. 93–105. **Scopus**.
9. *Chistova E., Smirnov I.* Light Coreference Resolution for Russian with Hierarchical Discourse Features // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — 2023. — С. 34–41. **Scopus**.

Свидетельства о регистрации программ для ЭВМ

10. *Чистова Е. В.* Программа для анализа дискурсивной риторической структуры текстов // Свидетельство о государственной регистрации программы для ЭВМ № 2024618391. — 2024.

Чистова Елена Викторовна

Методы анализа риторической структуры
текстов на русском языке

Автореф. дис. на соискание ученой степени канд. тех. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____