Федеральное государственное учреждение
«Федеральный исследовательский центр «Информатика и управление»
Российской академии наук»
(ФИЦ ИУ РАН)

На правах рукописи

Чистова Елена Викторовна

Методы анализа риторической структуры текстов на русском языке

Специальность 1.2.1 —
«Искусственный интеллект и машинное обучение»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
доктор технических наук
Смирнов Иван Валентинович

Москва — 2025

Оглавление

Стр.

Стр.

Стр.

Стр.

Введение

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, А. А. 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 приклад-

ных 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 автоматиза-

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

ции 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 риторического дискурсивного анализа текстов на русском языке, позволяющие выделять поверхностные и полнотекстовые дискурсивные структуры.
2. Метод риторического дискурсивного анализа текстов на основе глубокого обучения, позволяющий использовать при обучении разнородные данные риторической разметки.
3. Подходы, улучшающие кросс-языковую и кросс-жанровую обобщаемость риторического анализатора.
4. Метод классификации текстов, использующий риторические отношения в дискурсивной структуре для определения основной идеи текста.
5. Метод разрешения кореференции, в котором используются метрики расстояний между упоминаниями с учётом ядерности отношений в дискурсивной структуре.

6. Метод построения структуры аргументации в тексте поверх дискурсивной структуры.

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 [1–4], выполнены соискателем самостоятельно; разработана программа для ЭВМ [5]. В коллективных публикациях [6–10] 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 [6;7] автором разработаны и оценены методы классификации риторических отношений, а также проведён признаковый анализ риторических отношений в текстах на русском языке. В работе [8] 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 [9;10] 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».

Публикации. Основные результаты по теме диссертации изложены в 9 научных публикациях, из них 2 — в журналах из Перечня ВАК категории К1 [3; 4], 7 — в сборниках трудов международных конференций [1; 2; 6–10], из них 6 публикаций в изданиях, индексируемых в Scopus [1; 2; 6; 8–10], и 2 публикации в трудах конференций категории A* по рейтингу CORE [1; 2]. Зарегистрирована 1 программа для ЭВМ [5].

Объем и структура работы. Диссертация состоит из введения, пяти глав и заключения. Полный объём диссертации составляет 217 страниц, включая 25 рисунков и 36 таблиц. Список литературы содержит 201 наименование.

Глава 1. Теоретические основы анализа риторических структур в текстах на естественном языке

1.1. Введение

Современные приложения анализа естественного языка преимущественно опираются на моделирование морфологического и синтаксического компонентов языковой системы. С распространением методов глубокого обучения активно исследуются методы векторизации меньших языковых единиц — морфем и слов [11–14], а также их семантики в контексте предложения [15–17]. Однако смысл текста не сводится к простой сумме смыслов случайной последовательности предложений. В зависимости от коммуникативных целей автора и жанровых ограничений различные части текста могут выполнять разнообразные функции: экспозиция, развитие основной мысли, её обоснование, подтверждение и контраргументацию; развитие побочных идей; оценочные суждения; логические выводы, перефразировки и обобщения.

Языковой единицей наибольшего, потенциально неограниченного объёма является дискурс. В контексте обработки естественного языка под дискурсивным анализом понимается изучение структуры дискурса как целостной языковой единицы. Этот анализ включает в себя идентификацию элементарных дискурсивных единиц, определение связей между ними, а также построение локальной и глобальной структуры текста. Важным аспектом дискурсивного анализа также является изучение поведения меньших языковых единиц внутри общей структуры дискурса, включая, например, разрешение референциальных связей между упоминаниями сущностей или событий.

Существует несколько подходов к формальному описанию дискурсивной структуры связного текста на естественном языке. Среди них выделяют:

– Разметку локальных дискурсивных отношений (ЛДО) на основе явных и неявных *коннекторов* [18–23], а также пунктуации [24]. Коннекторы дискурса указывают на наличие дискурсивных отношений между предложениями или более крупными связными фрагментами текста[1]. Пример разметки предложения на основе коннектора: [Если]$_{\text{кон}}$ [Тимур выполнил всю работу,]$_{\text{арг1}}$ [он может пойти домой.]$_{\text{арг2}}$. Помимо явных (эксплицитных) коннекторов и пунктуации, представленных в тексте, могут анализироваться и неявные (имплицитные) маркеры, когда связь между дискурсивными единицами подразумевается, но явный коннектор опущен. Автоматический анализ локальных дискурсивных отношений применяется преимущественно в приложениях, где полезно выделение коннекторов и низкоуровневых отношений: в классификации отношения аргументации между двумя высказываниями [25], оценке коммуникативной сложности текста [26], извлечении информации [27; 28]. Необходимо заметить, что

– Графовые модели для представления дискурсивных структур. Так, в графовой модели Вольфа и Гибсона [29] к элементарным единицам дискурса для простоты сегментации приравнены любые клаузы, при этом клауза может вступать в любое количество отношений, а дискурсивные отношения между клаузами непроективные. Предложенная модель позволяет более полно отразить сложные взаимосвязи между высказываниями, однако отсутствие множества характерных для других моделей ограничений, таких как привязка к локальным коннекторам или укладка структуры в дерево

[1] Примеры дискурсивных коннекторов в русском и других языках см. в словаре http://connective-lex.info.

проективных связей, делает её наиболее сложной в имплементации и использовании. Перспективным направлением исследований в этой области является дополнение вторичными дискурсивными связями стандартных риторических стуктур [30].

– Иерархическую разметку дискурса в виде дерева составляющих. Теория риторических структур (TPC [31]), предложенная в 1980-х годах Вильямом Манном и Сандрой Томпсон, рассматривает текст как совокупность дискурсивных единиц, связанных между собой различными отношениями. Это наиболее широко используемая в анализе текстов теория описания связного дискурса, позволяющая анализировать его структуру на всех уровнях, используя единый набор отношений, а также выделять основную (ядра) и побочную (сателлиты) информацию относительно основной идеи текста из каждого риторического отношения. Анализ риторической структуры связного текста активно применяется в прикладных задачах анализа естественного языка: классификации текстов [9; 32; 33], реферировании [34–36], машинном переводе [37–39], вопросно-ответном поиске [40; 41], анализе структуры аргументации в тексте-рассуждении [42–44] и структуры нарратива [45; 46].

В данной главе рассматриваются основные задачи дискурсивного анализа в рамках теории риторических структур, а также обсуждаются модели представления риторических структур для различных языков и методы автоматизации их анализа.

1.2. Теория риторических структур

Теория риторических структур описывает дискурс в виде дерева составляющих, где *элементарные дискурсивные единицы* (ЭДЕ) являются листьями. Составляющие на каждом уровне дерева также являются *дискурсивными единицами* (ДЕ), связанными между собой риторическими отношениями. ТРС допускает вариативность интерпретации ограничений на ЭДЕ и риторические отношения в зависимости от языка, модуса (устный, жестовый [47], письменный), типа текста и жанра. Наиболее часто выделяют следующие общие классы риторических отношений: отношения связности (например, Детализация/Elaboration, Фон/Background, Подготовка/Preparation), контраста (Контраст/Contrast, Уступка/Concession, Антитезис/Anithesis), причинно-следственные (Причина/Cause, Волитивный-результат/Volitional-Effect, Неволитивный-результат/Non-volitional-Effect, Цель/Purpose и прочие), временны́е (Конкатенация/Joint, Последовательность/Sequence), Атрибуция (Attribution). Поскольку структура представляется строго в виде дерева составляющих, в случаях, когда одна дискурсивная единица вклинивается в другую, используется служебный тип связи Разрыв/Same-unit. Отношения бывают симметричными (несколько ядер[2]) и асимметричными (одно ядро и один сателлит, обозначается стрелкой в направлении ядра) в зависимости от фокуса внимания входящих в дискурсивное отношние единиц: ядра более центральны по отношению к основной идее текста, в то время как сателлит сообщает второстепенную информацию.

[2]В современном автоматическом разборе дискурса риторические структуры искусственно приводят к виду бинарных деревьев, однако естественная разметка предусматривает неограниченное число ядер во многих симметричных отношениях, например, временны́х: *[Катится колобок по дороге, а навстречу ему заяц: ...] [Катится колобок, а навстречу ему волк: ...] [Катится колобок, а навстречу ему медведь: ...] [Катится колобок, а навстречу ему лиса: ...].*

Риторическая структура текста свидетельствует об убеждении, аргументации и других манипуляциях информацией, посредством которых автор реализует свою коммуникативную цель. Определения ЭДЕ, риторических отношений и их ядер отражают этот факт, основываясь на оценке знаний и убеждений о мире говорящего/пишущего (автора) и адресата (читателя). К примеру, к критериям выделения элементарной ДЕ относятся *описание одной ситуации* и *отражение одного фокуса сознания*. Так, предложение [«Я подошёл к мальчику, который держал в руках слонёнка»]$_1$ соответствует элементарной дискурсивной единице с фокусом в первой предикации. В то же время предложение [«Я подошёл к Антону,] [который протягивал мне слонёнка»] содержит две самостоятельные ЭДЕ с известными актантами, между которыми можно выделить причинно-следственное отношение. Пример определения отношения Волитивная Причина по Манну и Томпсон [31] приведен в таблице 1.

Таблица 1 — Определение отношения Волитивная причина в Теории риторических структур

Ограничения на ядро (N)	Описывает осознанное действие либо ситуацию, которая могла бы возникнуть в результате осознанного действия.
Ограничения на сателлит (S)	—
Ограничения на их сочетание	– Сателлит может побуждать агента ядра к совершению осознанного действия. – В отсутствие сателлита действие может не восприниматься как мотивированное или его мотивация может быть неясна. – При описании сочетания ядро играет более значимую роль в воплощении авторского замысла, чем сателлит.
Эффект	Читатель признает описанное в сателлите причиной описанного в ядре.
Расположение эффекта	Ядро и сателлит.

Рисунок 1.1 иллюстрирует один из вариантов разбора риторической структуры небольшого текста на русском языке в корпусе RRT [48].

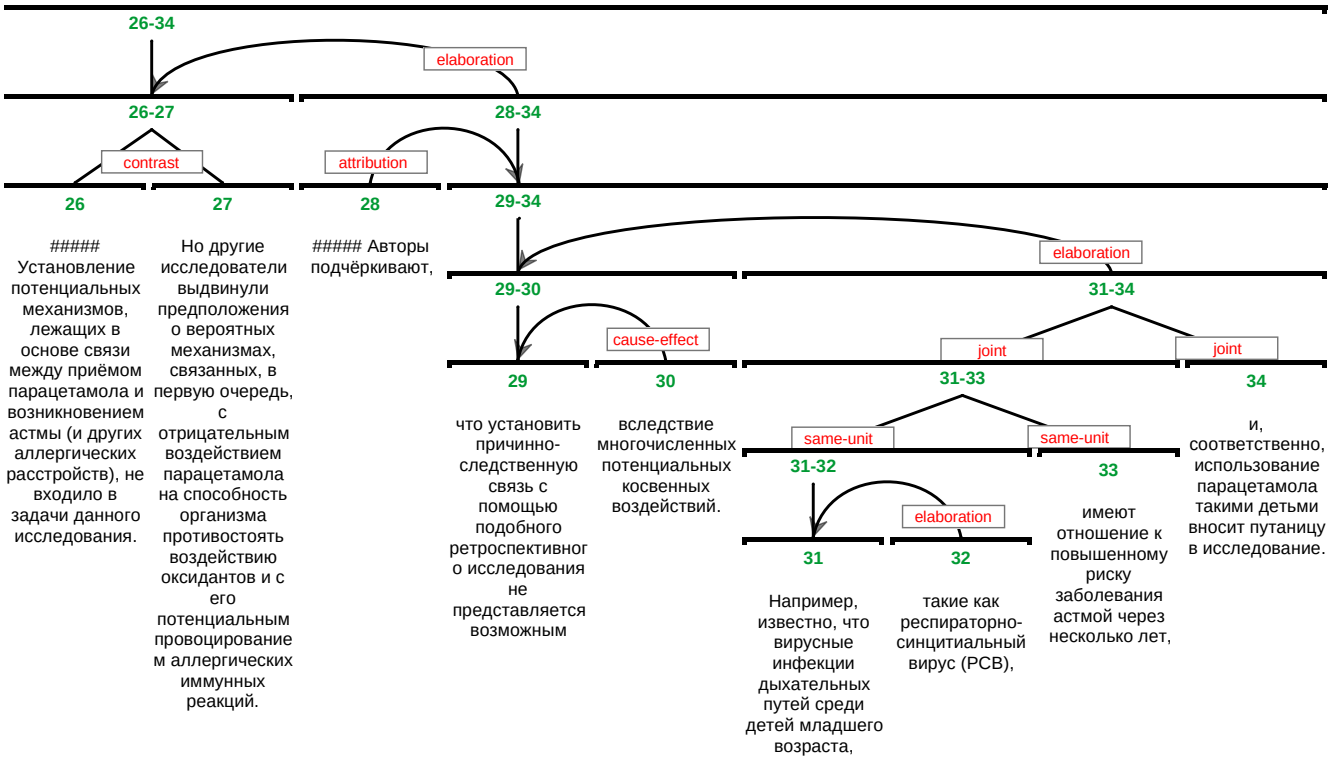


Рисунок 1.1 — Пример риторического разбора фрагмента текста на русском языке (`RRT:news1_24`).

1.3. Анализ риторических структур

В этом разделе приведены основные сведения о задаче анализа риторических структур. Рассмотрены основные подходы к решению задачи, описаны существующие корпусы риторической разметки для различных языков и стандартные способы оценки качества риторического анализа.

1.3.1. Постановка задачи

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:

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

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).

Задачу риторического анализа можно свести к поиску оптимального набора риторических отношений, который должен удовлетворять следующим критериям:

1. Корректное определение границ элементарных дискурсивных единиц. Эти единицы образуют терминальные узлы дерева, и внутри них не должно определяться никаких дискурсивных отношений. Последовательность ЭДЕ формирует исходный текст $D = \{\mathrm{edu}_1, \mathrm{edu}_2, \ldots, \mathrm{edu}_m\}$, где каждая элементарная дискурсивная

единица edu_i определяется интервалом $\text{edu}_i = \langle start_i, end_i \rangle, \quad 1 \leqslant start_i \leqslant end_i \leqslant n$.

2. Корректное определение риторических связей между соседними дискурсивными единицами на всех уровнях; в случае полнотекстового риторического анализа набор риторических отношений формирует единое дерево составляющих текста.
3. Корректное установление положения ядра в каждой риторической связи.
4. Правильное назначение типа риторического отношения между дискурсивными единицами.

Формальная цель задачи заключается в минимизации расхождений между предсказанным деревом T_P и эталонным деревом T_G по всем вышеуказанным критериям.

1.3.2. Обзор методов

Основополагающей работой в области риторического анализа является труд Дэниэла Марку «Теория и практика дискурсивного разбора и реферирования» [49]. Эта книга посвящена первым экспериментам по разметке риторических структур и оценке методов, основанных как на правилах, так и на машинном обучении. Для упрощения задачи Марку бинаризует деревья и устанавливает формальные ограничения на их детализацию, определяя абзацы и предложения как самостоятельные дискурсивные единицы. В ходе исследования было размечено 90 текстов, включающих 30 новостей, 30 публицистических заметок и 30 научных статей. Метод, основанный на правилах, связывает риторические отношения с явными лексическими маркерами — коннекторами. Обучаемый сегментатор помимо словаря маркеров использует признаки частей речи

слов текста и нахождения в контексте аббревиатур (для различения точки как конца предложения). Цель классификационной модели сегментации заключается в определении, является ли слово правой границей элементарной дискурсивной единицы. Для построения дерева используется алгоритм переноса-свертки; среди признаков для классификации действий на каждом шаге используются лексические, морфологические, структурные, сходства «мешков слов», а также отношения между сущностями в словаре WordNet [50]. Сегментация и построение дерева реализованы обучением деревьев решений на основе алгоритма C4.5. На основе опыта экспертной риторической разметки позже был разработан и опубликован крупный корпус для обучения и разметки RST-DT (RST Discourse Treebank) [51]. Он включает разметку 385 новостных статей и заметок из газеты “Wall Street Journal”, входящих в корпус морфосинтаксической разметки Penn Treebank [52]. Этот англоязычный корпус стал эталоном в области автоматического риторического анализа.

В последующие два десятилетия ключевые принципы построения риторического анализатора оставались неизменными:

1. *Сегментатор* предсказывает границы ЭДЕ в последовательности токенов.
2. *Построитель дерева*, зачастую на основе алгоритма переноса-свёртки, собирает из ЭДЕ дерево составляющих.
3. *Классификатор риторических отношений* определяет тип и ядерность связи между дискурсивными единицами. Может быть частью построителя дерева как, например, расширение набора действий свёртки, или реализован двумя отдельными классификаторами ядерности и отношения.

Риторические анализаторы, основанные на этих принципах, известны как *восходящие*. Их суть заключается в предсказании ЭДЕ и последовательном переходе ко все более крупным фрагментам текста, вплоть до его

полного покрытия. Основным направлением развития методов этого типа стало совершенствование составляющих их обучаемых моделей благодаря прогрессу в области традиционного машинного и глубокого обучения, а также развитию методов представления и анализа текстов. В восходящем разборе с успехом применяются дистрибутивная семантика [8; 53], контекстная векторизация с использованием предобученных языковых моделей [53–56], морфосинтаксические признаки [57–60], а также методы разрешения кореференции [54; 61] и анализа тональности [8; 61]. Для обработки признаков, классификации слов и дискурсивных единиц, а также для построения деревьев используются методы традиционного машинного обучения, такие как условные случайные поля [59; 60; 62], метод опорных векторов [58; 63; 64], градиентный бустинг [8], и методы глубокого обучения, включая полносвязные [55;57] и рекуррентные [53;61;65–68] нейронные сети. Помимо переноса-свёртки [53–57; 63; 64; 67; 68] построение структуры составляющих из последовательности ЭДЕ может быть реализовано линейной свёрткой с ранжированием вероятностей [7; 58; 60; 65] или генерацией дерева в строку со скобочным синтаксисом [61]. Для глобальной оптимизации риторического дерева помимо непосредственного «жадного» разбора применяют алгоритмы динамического программирования: Кока–Янгера–Касами [59; 62], Витерби [63]. С развитием «больших» языковых моделей-декодировщиков (LLM, Large Language Model) активно исследуются возможности классификации дискурсивных отношений [69] и текстовой генерации риторической структуры. Например, классификация действий переноса-свёртки, ядерности и типа отношения может осуществляться посредством текстовых запросов к генерирующей предобученной модели [70]. Несмотря на достижение наилучшего качества построения структур из заданной последовательности ЭДЕ и обобщающей способности, такие методы построения риторических структур остаются экспериментальными и не применимы на практике ввиду чрезмерных требований к вычислительным ресурсам и крайне низкой производительности.

Современные восходящие методы риторического разбора, состоящие как минимум из двух модулей (сегментатора и построителя дерева), обладают двумя основными недостатками: громоздкой архитектурой и отсутствием обратной связи между модулями. Необходимость обучать и использовать отдельные модули, каждый из которых может состоять из нескольких моделей, снижает производительность системы. Независимое обучение модулей приводит к следующим проблемам:

1. Методы автоматической сегментации преимущественно опираются на лексические маркеры, а также морфосинтаксические признаки; их точность ограничена этими признаками. Важно отметить, что ЭДЕ не является синтаксическим конструктом, хотя часто оформляется предикацией. Определяющей особенностью ЭДЕ является неспособность выделить внутри неё риторическое отношение из принятого в данном корпусе набора. Опыт корпусной разметки [49] показывает, что при предварительном разделении текста на клаузы эксперт-разметчик корректирует сегментацию в процессе построения дерева в соответствии с определениями отношений в руководстве по разметке, объединяя или разделяя клаузы.
2. Методы построения дерева обучаются на основе экспертной сегментации текста и склонны к переобучению на категориях ЭДЕ, которые автоматический сегментатор не способен адекватно предсказать. Это приводит к тому, что качество построения дерева на экспертной последовательности ЭДЕ не всегда соотносится с качеством анализа текста с нуля.
3. Методы сегментации и построения структуры развиваются и оцениваются раздельно. Во многих исследованиях сценарий полного разбора неразмеченного текста вовсе не рассматривается, в редких случаях оценивается в связке с одним сторонним методом сегментации, что затрудняет оценку фактического качества полных систем разбора.

Ограничения восходящих методов стимулировали интерес к методам *нисходящего разбора*, которые предлагают альтернативное решение проблемы. В отличие от восходящих, в нисходящих методах риторическое дерево строится, начиная с глобальной дискурсивной структуры и постепенно уточняя его до уровня элементарных дискурсивных единиц. Это позволяет избежать некоторых проблем, присущих восходящему разбору, таких как накопление ошибок на каждом этапе и ограниченный контекст векторных представлений дискурсивных единиц на каждом шаге. Исследователи предлагают в том числе и методы нисходящего разбора структуры в отрыве от сегментации и полного анализа, опирающиеся на экспертную сегментацию [71–74]. Однако главным преимуществом нисходящего разбора является возможность реализации анализа текста с нуля одной моделью глубокого обучения. Для этого строят гибридное дерево с листьями-токенами [75] или дообучают единую предобученную языковую модель для кодирования токенов текста с выходами для сегментации элементарных или рекуррентной сегментации неэлементарных дискурсивных единиц, функционально связанными только на этапе предсказания дерева [76]. На данный момент такие методы обеспечивают лучшие качество и производительность риторического разбора текста с нуля.

1.3.3. Корпусы риторической разметки

Наряду с эталонным англоязычным корпусом RST-DT, содержащим разметку 385 новостных текстов и 19778 дискурсивных отношений 17 основных типов, было создано множество корпусов риторической разметки для различных языков и жанров:

– Английский: Мультижанровый многослойный корпус Джорджтаунского университета GUM v9.1 [77] содержит разметку 213 доку-

ментов 12 жанров, включая письменный и устный дискурс; 26106 примеров отношений 17 типов.

– Баскский: Basque RST DT [78] (88 аннотаций, 31 тип отношений, 1292 примера отношений, согласованность — 0,61 R).
– Бразильский вариант португальского языка: CST-News [79] (140 новостей, 31 тип отношений, 5216 примеров отношений; согласованность разметчиков: 0,78 N, 0,66 R); Summ-it [80] (50 научно-популярных статей, 1677 примеров отношений); Rhetalho [81] (20 новостных текстов, 20 статей из информатики, 23 типа отношения); CorpusTCC [82] (100 введений из диссертаций по информатике, 31 тип отношения, 3946 примеров отношений).
– Испанский: Spanish RST-DT [83] (267 текстов научных тематик, 29 типов отношений, 3115 примеров).
– Китайский: Небольшой корпус параллельной испано-китайской разметки RST SC TB [84] (∼15000 токенов, 26 типов отношений, 692 примера отношений для китайского языка); мультижанровый корпус GCDT [85] (50 текстов, 5 жанров, набор отношений GUM RST, 8413 примеров отношений).
– Немецкий: Postdam Commentary Corpus [86] (175 коротких новостных заметок, 30 типов отношений, 2665 примеров).
– Нидерландский: Корпус DDA [87] (80 текстов из онлайн-энциклопедий и научных новостей, 31 тип отношений, 2264 примера отношений, согласованность — 0,77 N, 0,79 R,).
– Персидский: Persian RST [88] (150 новостных текстов, 19 типов отношений, 5191 пример отношения).
– Русский: RuRSTreebank (RRT, [48]) (233 новостных и блоговых текстов из различных интернет-ресурсов). Ранний подкорпус разметки научных статей содержит непроективные деревья и не может быть использован в разработке методов разбора риторических структур. Важной особенностью корпуса является частичная разметка доку-

ментов. Текущая версия содержит разметку 25957 отношений 17 типов.

На материалах экспертной разметки, включающей не менее 100 структур, были обучены риторические анализаторы для английского языка (на основе RST-DT и GUM), португальского (на объединённых и гармонизированных корпусах CST-News, Summ-it, Rhetalho и CorpusTCC), испанского, немецкого, персидского языков [61; 76; 88].

Проблема согласованности разметки внутри корпуса. Анализ риторической структуры текста произвольной длины не является однозначной задачей [89]. Выбор определённого варианта сегментации, группировки дискурсивных единиц, определения ядерных единиц и типов риторических отношений зачастую обусловлен интерпретацией текста и наличием определённых в инструкции формальных ограничений. Так, для пяти новостных заметок из эталонного корпуса RST-DT, полученных из одного источника и размеченных более чем дюжиной экспертов в течение года [51], была обнаружена согласованность 87% для построения структуры, 81% для назначения ядер и 72% для классификации типа отношения [51]. Для других корпусов, включающих тексты различной длины, жанров и языков, показатели согласованности по назначению ядер и определению типов отношений обычно не превышают 80% [78; 79; 87]. Следует также отметить, что метод расчёта согласованности между дискурсивными деревьями зачастую даёт завышенные значения, что связано с учётом обоих вариантов сегментации [77]. Таким образом, можно сделать вывод о сравнительно низкой согласованности данных, что создаёт дополнительные трудности при построении на их основе автоматических анализаторов.

Проблема универсализации риторических корпусов. Помимо разнообразия языков и жанров, в каждом корпусе принят собственный набор риторических отношений с различными определениями и детализацией.

Для решения задачи кросс-языкового риторического анализа разметку из нескольких корпусов часто *гармонизируют*, приводя её к более общему набору отношений по некоторому словарю [61; 76; 90; 91]. Такой подход позволяет исследовать кросс-языковые возможности автоматического риторического разбора, однако имеет следующие важные недостатки:

1. Снижается информативность отношений.
2. Затрудняется сравнение кросс-языковых методов анализа с методами для одного языка, оперирующими экспертным набором отношений.

Это подчёркивает необходимость разработки методов риторического анализа, позволяющих при обучении моделей эффективно использовать множество несовместимых между собой корпусов без искажения содержащейся в них информации.

1.3.4. Оценка качества риторического разбора

Риторические деревья T_G (размеченный экспертом эталон) и T_P (дерево, предсказанное системой) можно описать множествами содержащихся в них отношений вида $\langle du_L, du_R, rel, nuc\rangle$, где $du_L = \langle start_L, end_L\rangle$ — позиции левой дискурсивной единицы в исходном тексте, $du_R = \langle start_R, end_R\rangle$ — позиции правой ДЕ, $nuc \in \mathrm{NS}, \mathrm{SN}, \mathrm{NN}$ — положение ядра (ядро N слева, справа или дискурсивные единицы равноправны), $rel \in \{\textsc{Конкатенация}, \textsc{Детализация}, \textsc{Контраст}, \ldots\}$ — отношение из заданного экспертами набора.

Качество риторического анализа оценивается по следующим пяти параметрам:

1. Сегментация (Seg): Оценивается точность определения границ элементарных дискурсивных единиц (ЭДЕ), основанная на разметке

токенов предложения (находится ли данный токен на границе ЭДЕ).

2. Построение неразмеченной структуры составляющих (Span): Оценивается на основе позиций дискурсивных единиц в тексте: сравниваются пары вида $\langle du_L, du_R \rangle$.
3. Структура с учётом ядерности (Nuclearity, N): Включает также положения ядер (nuc), что позволяет оценить корректность распределения значимости между частями текста.
4. Структура и отношения (Relation, R): Учитываются только отношения (rel) без учёта ядерности, что позволяет оценить правильность идентификации типов риторических связей: $\langle du_L, du_R, rel \rangle$.
5. Полный разбор (Full): Учитываются все параметры разметки, включая позиции дискурсивных единиц, типы отношений и положение ядер: $\langle du_L, du_R, rel, nuc \rangle$. Этот параметр предоставляет наиболее полную оценку качества риторического анализа.

В современных исследованиях основным критерием оценки является микроусредненная F1-мера. Наборы пар дискурсивных единиц, выделенные экспертами и предсказанные системой, сравниваются по стандартной процедуре Parseval [92]:

– Точность:

$$\text{Precision} = \frac{|T_P \cap T_P|}{|T_P|}, \tag{1.1}$$

где $|\mathrm{T}_P \cap \mathrm{T}_P|$ — количество корректно предсказанных пар дискурсивных единиц; $|\mathrm{T}_P|$ — общее количество предсказанных системой пар дискурсивных единиц.

– Полнота:

$$\text{Recall} = \frac{|T_P \cap T_G|}{|T_G|}, \tag{1.2}$$

где $|T_P \cap T_G|$ — количество правильно предсказанных пар дискурсивных единиц; $|T_G|$ — общее количество пар дискурсивных единиц в эталонном дереве.

– F1-мера:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1.3}$$

F1-мера является гармоническим средним между точностью и полнотой, предоставляя сбалансированную оценку, учитывающую как полноту, так и точность предсказания.

1.4. Кросс-языковая обобщаемость анализа риторических структур

Качественное экспертное сравнение риторических структур в работе [93] заложило основу для исследования возможностей кросс-языкового анализа риторических структур. Авторами был проведён сравнительный анализ небольшого параллельного корпуса на английском, испанском и баскском языках (318 ЭДЕ на каждый язык). Анализ выявил значительные сходства в риторических структурах между языками. Различия в основном проявились в дискурсивной сегментации (на синтаксическом уровне) и следующей из сегментации дискурсивной структуре предложений. В дальнейшем был разработан испано-китайский двуязычный корпус, включающий разметку 50 текстов разной сложности дискурса (111–1774 слов) [84]. На материалах корпуса был проведен сравнительный анализ испанского и китайского языков. Как ключевые факторы различия дискурсивной структуры были выделены различия в использовании дискурсивных коннекторов и пунктуации, вариативность порядка ЭДЕ, а также вставки и удаления отдельных ЭДЕ.

Первые эксперименты по автоматическому кросс-языковому риторическому анализу включали обязательный этап гармонизации риторических структур в корпусах, различающихся в интерпретации риторических структур и наборе типов дискурсивных отношений. В частности, в работе [94] были выделены 18 унифицированных типов отношений, к которым по

словарю наименований отношений приводились схожие отношения из корпусов на различных языках. В последующей работе [90] исследовались возможности адаптации различных моно- и мультиязыковых дискурсивных анализаторов к анализу структуры в малоресурсном баскском языке при помощи мультиязыковых предобученных моделей статической векторизации слов. Для кросс-языкового анализа предложено использовать машинный перевод на уровне ЭДЕ [76; 91]. Несмотря на то, что этот метод позволяет компенсировать нехватку данных для отдельных типов отношений, независимый автоматический перевод отдельных ЭДЕ искажает дискурсивную естественность текстов на втором языке, калькируя дискурсивную сегментацию из исходного языка.

В дискурсивном корпусе Джорджтауновского университета для китайского языка (GCDT) [85] предложена разметка риторических структур 50 текстов (9710 ЭДЕ), охватывающих 5 из 10 жанров, размеченных в англоязычном корпусе GUM [77], согласно тому же набору отношений и принятым в GUM формальным ограничениям на структуры. Примечательно, что 19 документов, взятых из многоязычных источников, таких как Wikipedia, Wikinews и wikiHow, имеют англоязычные аналоги в GUM, однако содержание и подача информации в них могут значительно различаться для разных языков, что затрудняет объективную оценку кросс-языковой обобщаемости существующих методов.

Следует отметить, что до сих пор исследования кросс-языковой обобщаемости автоматического дискурсивного анализа на параллельных данных ограничивались преимущественно задачами выделения дискурсивных коннекторов и классификации типа дискурсивных отношений на небольшом параллельном корпусе TED-MDB [95]. Данный корпус содержит разметку шести текстов на каждом из шести языков (английский, русский, польский, португальский, немецкий, турецкий) и включает от 560 до 661 примера дискурсивных отношений для каждого языка. Такого количества данных недостаточно для обучения моделей разбора целостной структуры

текста. Для сравнения, корпус риторической разметки для русского языка RRT содержит 25957 примеров дискурсивных отношений.

Таким образом, создание большого параллельного корпуса дискурсивной разметки и исследование на его основе кросс-языковой обобщаемости методов дискурсивного анализа остаётся актуальной задачей.

1.5. Выводы

Теория риторических структур описывает дискурс в виде дерева составляющих, где элементарные дискурсивные единицы связаны между собой риторическими отношениями. Это позволяет выявлять иерархические связи между частями текста, что имеет важное значение для анализа его глобальной структуры и коммуникативной цели. В первой главе изложены основные сведения о теории риторических структур, а также приведен обзор методов автоматического разбора риторических структур. Рассмотрены этапы развития методов риторического анализа, начиная от ранних подходов на основе правил и классического машинного обучения, до современных методов глубокого обучения, включая как восходящие, так и нисходящие стратегии построения дискурсивных деревьев. В частности, проанализированы существующие ограничения восходящих методов, такие как необходимость обучения и использования отдельных модулей, отсутствие обратной связи между ними, а также проблемы, связанные с качеством сегментации и построения деревьев. В обзоре также описаны ключевые параметры оценки качества риторического анализа: оценка качества сегментации, построения структуры, определения ядерности и назначения типов риторических отношений.

В главе также приведены сведения о существующих корпусах текстов с риторической разметкой для разных языков. Сформулированы основные

проблемы анализа риторических структур на основе корпусов: проблема низкой согласованности разметки внутри корпуса, продиктованная неоднозначностью интерпретации текстов, и проблема универсализации разметки между корпусами, приводящая к сложностям с обобщением автоматического анализа на различные жанры и языки. Эти проблемы подчёркивают сложность создания универсальных анализаторов, использующих разметку нескольких корпусов, а также важность создания большого параллельного корпуса дискурсивной разметки и исследования кросс-языковой обобщаемости риторического анализа.

Методы риторического дискурсивного анализа разработаны для множества языков, однако риторический анализ для русского языка изучен недостаточно. Разработанный ранее корпус RRT содержит поверхностную риторическую разметку текстов на русском языке, необходимую для исследования особенностей типов риторических отношений и построения моделей риторического анализа для русского языка. Анализ исследований обнаруживает необходимость в создании методов риторического анализа текстов на русском языке; разработке первого большого параллельного корпуса дискурсивной разметки и исследовании на его основе возможностей кросс-языкового обобщения методов риторического анализа; разработке кросс-жанровых методов риторического анализа.

1.6. Цель и задачи исследования

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.

Глава 2. Анализ риторических отношений в русскоязычном письменном дискурсе

2.1. Введение

В этой главе предложены подходы к решению задачи классификации риторических отношений. Риторический анализ включает две задачи классификации риторических отношений: выделение ядерных и второстепенных элементов риторического дерева (*классификация нуклеарности*) и определение типа риторической связи. Описано исследование типов риторических отношений в русскоязычном письменном дискурсе на основе машинного обучения и анализа текстовых признаков.

Анализ риторических отношений ранее был предметом исследования преимущественно для англоязычных текстов. Для классификации риторических связей применялись различные признаки: параметры синтаксической структуры, дискурсивные маркеры, оценки семантической близости. С развитием технологий машинного обучения особую популярность приобрели нейросетевые модели, которые позволяют извлекать неявные связи в паре дискурсивных единиц. Однако такие подходы требуют значительных объёмов данных, а результаты могут быть недостаточно интерпретируемы.

В этой главе описано исследование, включающее разработку новых методов классификации риторических отношений и признаков, основанных на лексическом, морфологическом, синтаксическом и семантическом анализе текста. Представлены результаты применения различных методов машинного обучения, таких как логистическая регрессия, метод опорных векторов и градиентный бустинг, к задачам классификации типов риторических отношений и нуклеарности. Анализируется значимость различных

групп признаков дискурсивных единиц для определения типа риторического отношения. В заключении приводятся выводы о применимости разработанных методов для риторического анализа русскоязычного дискурса и их значении для дальнейших исследований в области анализа риторических структур в текстах на русском языке.

2.2. Признаковое описание дискурсивных единиц в задаче классификации риторических отношений

Классификация риторических отношений является ключевой задачей анализа дискурсивной структуры текста. Большинство работ в этом направлении посвящено анализу риторических отношений в письменном англоязычном дискурсе. В ранних исследованиях активно исследовались синтаксические признаки. Например, был предложен метод построения дискурсивных предложений на основе лексикализованных синтаксических деревьев [96]. В дальнейшем подход получил развитие за счёт использования тегов частей речи и синтаксических признаков в анализаторе на основе алгоритма переноса-свёртки [97]. Со временем набор признаков для классификации риторических отношений был значительно расширен. В частности, был предложен анализатор, включавший такие признаки, как коннекторы, пунктуация и словесные N-граммы [58]. В других исследованиях рассматривалось использование синтаксических правил для дискурсивных шаблонов, таких как отношения подчинения между дискурсивными единицами в синтаксическом дереве, последовательности частей речи, расстояние от элементарных дискурсивных единиц до ближайшего общего предка [98–100]. Эти работы подтвердили значимость синтаксических признаков для улучшения качества классификации риторических отношений в английском языке.

В последующие годы внимание стало уделяться семантическим и лексическим признакам. Рассматривались такие аспекты, как семантическое сходство между глаголами или существительными в паре дискурсивных единиц [99], а также особенности расположения токенов и частей речи в начале и конце каждой дискурсивной единицы [101]. Стали также активно исследоваться методы глубокого обучения. Например, нейронная тензорная сеть с интерактивным вниманием использовалась для выделения значимых пар слов, что позволило улучшить точность классификации [102]. Также изучалась возможность применения признаков, связанных с сущностями, для извлечения неявных[1] дискурсивных отношений между предложениями в абзаце [103], что повысило эффективность классификации отношений типов Expansion и Comparison. На материале англоязычного корпуса было установлено, что для некоторых типов дискурсивных отношений, таких как Comparison, Contingency и Expansion, характерны определённые семантические и лексические свойства, которые могут быть использованы для их классификации [104]. Более того, было выявлено, что задачи классификации риторических и локальных дискурсивных отношений, а также задачи классификации коннекторов могут быть успешно решены с помощью единой модели глубокого обучения. При этом получаемые внутренние представления отношений обеспечивают более точное отражение неявных связей, временных последовательностей и отношений детализации по сравнению с моделями, которые обучаются только классификации риторических отношений в корпусе RST-DT [105]. Однако важно отметить, что неявные отношения в риторических структурах и поверхностная разметка дискурсивных зависимостей часто имеют низкую согласованность [106].

Несмотря на успехи, достигнутые с помощью глубоких нейросетевых моделей, эффективное обучение часто требует больших объёмов дан-

[1] Дискурсивное отношение называется *неявным*, когда оно не определяется присутствующим коннектором однозначно или коннектор опущен. Сравните примеры отношения Причина: «[Грызуны не станут приближаться,] [**потому что** запах слишком сильный]» и «[Запах слишком сильный,] [грызуны приближаться не станут.]»

ных, а результаты могут быть недостаточно интерпретируемы. В связи с этим, в данном исследовании для классификации риторических отношений в письменном русскоязычном дискурсе был выбран подход, основанный на разработке и изучении текстовых признаков что позволяет глубже исследовать особенности отдельных типов риторических отношений.

2.3. Методы классификации риторических отношений в текстах на русском языке на основе машинного обучения

Объектами классификации являются пары дискурсивных единиц, объединённых риторическим отношением в размеченном корпусе. Исследуются две задачи многоклассовой классификации в рамках риторического анализа. Первая задача заключается в классификации пар дискурсивных единиц по 11 типам риторических отношений. Вторая задача состоит в классификации ядерности. В РРС различают три типа ядерности в зависимости от положения ядра: асимметричные «Сателлит–Ядро» (SN), «Ядро–Сателлит» (NS) и равноправный «Ядро–Ядро» (NN).

В рамках обеих задач рассматриваются комбинации лексических, морфологических и семантических признаков. В качестве лексических признаков используется список маркерных фраз (коннекторов), включающий около 450 элементов. Этот список был составлен вручную на основе трёх источников: выражений, извлечённых экспертами из аннотированных текстов, словаря союзов в сложных предложениях, описанных в РусГраме[2], и списка функциональных многословных единиц, предложенных в Национальном корпусе русского языка[3].

[2] http://rusgram.ru

[3] http://ruscorpora.ru/obgrams.html

Перед извлечением признаков выполняются следующие этапы предобработки текста: токенизация, лемматизация, разметка частей речи и морфологический анализ.

Набор признаков включает различные числовые характеристики каждой дискурсивной единицы в паре:

1. Количество слов.
2. Средняя длина слов.
3. Количество полностью прописных слов.
4. Количество слов, начинающихся с заглавной буквы.
5. Количество различных морфологических признаков (например, для глаголов — лицо и число).
6. Частеречные теги для первых и последних пар слов.
7. Количество вхождений стоп-слов.
8. Количество вхождений каждой маркерной фразы.
9. Наличие маркерной фразы в начале и в конце ДЕ.
10. TF-IDF [107] вектор дискурсивной единицы.
11. Усреднённые векторы слов. Модели векторизации были обучены с использованием word2vec [108].

Также учитываются признаки пары ДЕ, оценивающие лексическое, синтаксическое и семантическое сходство двух высказываний:

1. Признаки, указывающие на сходство между векторами морфологических признаков двух ДЕ, с использованием различных мер сходства: косинусное сходство, расстояние Хэмминга, расстояние Канберра и мера сходства для бинаризованных векторов.
2. Косинусное сходство между TF-IDF векторами.
3. Индекс Жаккара между лемматизированными ДЕ.
4. Мера сходства BLEU.

Образцы примеров наименее частотных классов использовались для обучения регрессора, который оценивает вероятность появления асиммет-

ричных (одноядерных) связей между ДЕ. Эта оценка вероятности также используется как признак в задаче маркировки отношений.

Для классификации используются основные широко используемые алгоритмы обучения с учителем: логистическая регрессия, полносвязная нейронная сеть (NN), метод опорных векторов (SVM) с различными ядрами [109], градиентный бустинг на решающих деревьях (GBT), реализованный в пакетах LightGBM [110] и CatBoost [111].

Полносвязная нейронная сеть представляет собой двухслойный перцептрон. Функцией активации первого слоя является ReLU, используется Dropout-регуляризация. К выходам первого слоя применяется пакетная нормализация, а в выходном слое используется функция активации softmax. Поскольку несбалансированность данных сильно влияет на производительность модели нейронной сети, для аугментации всех классов, кроме наиболее представленного, применяется техника SMOTE [112].

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[4]. Для отбора признаков используется логистическая регрессия с L1-регуляризацией.

[4]От 320 до 1200 примеров каждого класса при более 3200 признаков в описываемых экспериментах.

2.4. Экспериментальное исследование риторических отношений в русскоязычном письменном дискурсе

2.4.1. Описание данных

Для исследования признаков в русскоязычном письменном дискурсе использована ранняя версия корпуса RuRSTreebank ($RRT_{v1.0}$) [48], включающая 179 текстов различных жанров письменного дискурса, таких как новостные статьи, научно-популярные тексты и исследовательские статьи по лингвистике и компьютерным наукам, с общим объёмом в 203287 токенов. Подкорпус академических текстов из $RRT_{v1.0}$ использовался исключительно для исследования классификации отношений, поскольку содержит непроективные деревья, которые непригодны для построения риторических структур. В целях исследования исключены наиболее частые и наименее информативные классы Конкатенация, Детализация, а также служебный тип связи Same-unit. В результате эксперименты были проведены с 11 наиболее представленными в корпусе классами: Причина (cause), Подготовка (preparation), Условие (condition), Цель (purpose), Контраст (contrast), Атрибуция (attribution), Свидетельство (evidence), Последовательность (sequence), Оценка (evaluation), Фон (background), Сравнение (comparison). Распределение классов в наборе данных изображено на рисунке 2.1.

2.4.2. Детали экспериментов

Лучшее значение параметра регуляризации для фильтрации признаков было найдено с помощью жадного поиска при 5-кратной кросс-

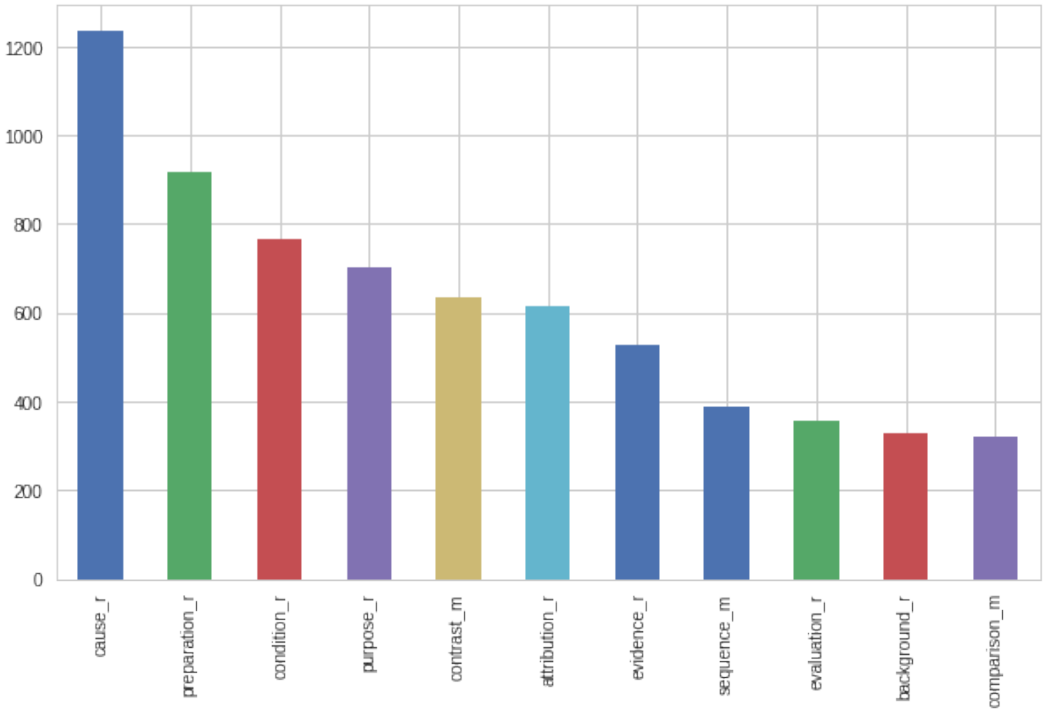


Рисунок 2.1 — Распределение классов в наборе данных для анализа типов риторических отношений

валидации. Для поиска оптимальных параметров логистической регрессии и SVM, таких как коэффициент регуляризации C и тип штрафа (L1, L2), а также параметров нейронной сети (количество слоёв, функция активации, коэффициент отсева) использовался алгоритм случайного поиска. Для выбора лучших гиперпараметров для моделей градиентного бустинга, таких как количество деревьев, число листьев, скорость обучения, коэффициент выборки признаков и коэффициенты регуляризации, применялся аналогичный подход. Для определения оптимального числа итераций в модели CatBoost использовался встроенный детектор переобучения на валидации.

2.4.3. Оценка качества

Для оценки результатов использовались стандартные метрики: точность, полнота и F_1-мера. Использовалось макроусреднение метрики, тогда как точность не рассматривалась из-за несбалансированности классов. Все

эксперименты проводились с использованием 5-кратной перекрёстной проверки со стратифицированным случайным разделением данных: 90% для обучения и 10% для тестирования.

2.4.4. Анализ результатов

В таблице 2 представлены результаты экспериментов с классификацией риторических отношений на основе интерпретируемых признаков. Результаты показывают, что ансамбли моделей градиентного бустинга с отбором признаков и линейными классификаторами SVM достигают наилучшего качества на тестовой выборке. Наиболее высокое качество достигается при классификации при помощи мягкого ансамбля модели градиентного бустинга CatBoost с отфильтрованными признаками и линейной SVM-модели.

Таблица 2 — 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

Результаты классификации положения ядра в отношении представлены в таблице 3. Эксперименты с полным набором признаков показали,

что модели градиентного бустинга значительно превосходят полносвязную нейронную сеть, SVM и логистическую регрессию.

Таблица 3 — 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

Проведена оценка значимости лексических признаков риторических отношений. Признаки дискурсивных маркеров включают два типа: позиционные (наличие коннектора в начале или конце дискурсивных единиц) и количественные (количество коннекторов в каждой единице). В таблице 4 можно заметить снижение качества классификации типов риторических отношений после исключения позиционных признаков. В то же время количественные признаки не оказывают существенного влияния на значение F_1-меры. Соответствующий риторическому отношению коннектор ожидаемо чаще находится в начале или конце дискурсивной единицы. Эти результаты демонстрируют целесообразность разработанного словаря коннекторов для дискурсивного анализа на русском языке.

Таблица 4 — F_1, % для задачи классификации риторических отношений с различными стратегиями учёта вхождения в ДЕ коннекторов

Набор признаков	Macro F_1		
	LogReg	SVM	CatBoost
Все	51,5	50,6	52,4
Без количественных	-0,3	+0,1	-0,1
Без позиционных	-4,0	-4,0	-2,8

Для ранжирования информативности признаков, используемых в лучшем методе классификации риторических отношений (CatBoost после отбора признаков логистической регрессией с L1-регуляризацией), оценивается влияние изменения значения каждого признака на предсказание модели. При помощи этого подхода из всего набора признаков (3,624 признака) выделено 2,014 наиболее информативных признаков. Анализ этих признаков представлен в таблице 5. Можно заключить, что модель градиентного бустинга после фильтрации признаков также преимущественно опирается на лексические признаки вхождения явных коннекторов дискурса. После исключения информации о 1887 признаках, связанных с дискурсивными маркерами, качество классификации снижается на 2,49% макроусредненной F_1-меры.

Детальная оценка классификации лучшим методом представлена в таблице 6. В дополнение, на рисунке 2.2 приведена матрица ошибок этого метода. Относительно лучшее качество классификации достигается для ассиметричных типов отношений, наиболее часто сигнализируемых коннекторами; например, для класса Атрибуция (attribution) достигнута F_1-мера в 74.36%. Это показывает, что выражение некоторых дискурсивных отношений в русском языке характеризуется формальными особенностями, не требующими глубокого семантического и прагматического анализа, такими как соотношение длин дискурсивных единиц, вхождение определённых частей речи, синтаксический параллелизм (таблица 5).

Более низкие оценки качества с F_1-мерой менее 50% получены для четырёх классов, имеющих наименьшее количество обучающих экземпляров: Сравнение (comparison, 320 примеров), Свидетельство (evidence, 529 примеров), Оценка (evaluation, 356 примеров) и Фон (background, 328 примеров). Специфика определений этих классов требует более тщательного анализа помимо морфологического и синтаксического анализа высказываний. Например, классы Свидетельство, Оценка и Фон наиболее ошибочно классифицируются как Причина (рисунок 2.2), что объясняет-

Таблица 5 — Информативные признаки, выделенные из обученной модели градиентного бустинга после фильтрации признаков

Тип	Признаки	Кол-во	%	-F1, %
Представления	4 элемента векторов TF-IDF первой дискурсивной единицы; 4 элемента векторов TF-IDF второй дискурсивной единицы;	8	0,4	0,11
Морфосинтаксические	Комбинации пунктуации, существительных, глаголов, наречий, союзов, прилагательных, предлогов, местоимений, числительных, частиц в начале первой ДЕ; Комбинации пунктуации, глаголов, наречий, существительных, местоимений, прилагательных, союзов, предлогов, частиц, числительных в конце первой ДЕ; Количество существительных в творительном падеже, местоимений, наречий в первой ДЕ; Различные комбинации глаголов, местоимений, существительных, наречий, союзов, пунктуации, частиц в начале второй ДЕ; Различные комбинации пунктуации, существительных, глаголов, местоимений, наречий, прилагательных, предлогов, союзов, частиц в конце второй ДЕ; Количество союзов, наречий, прилагательных, местоимений, предлогов во второй ДЕ; Количество пассивных глаголов, деепричастий и инфинитивов во второй ДЕ; Корреляция между морфологическими векторами пары ДЕ.	119	5.9	0.45
Лексические	Количество вхождений 355 маркеров в первой ДЕ (18%) Количество вхождений 331 маркера во второй ДЕ (17%) Вхождения 298 маркеров в начале первой ДЕ (16%) Вхождения 326 маркеров в конце первой ДЕ (17%) Вхождения 335 маркеров в начале второй ДЕ (19%) Вхождения 242 маркеров в конце второй ДЕ (13%)	1887	93,69	2,49

ся значительным числом примеров для последнего класса (1235 примеров) при наиболее схожем оформлении высказываний в этих классах. С низкой полнотой определяется отношение Фон, часто классифицируемой как Подготовка. Следует отметить, что в автоматических анализаторах для английского языка эти отношения объединяют ввиду близких определений. В других случаях, например, в парах Причина – Подготовка или

Таблица 6 — Качество классификации по классам риторических отношений, %

Класс	Precision	Recall	F1
attribution	73,11	75,77	74,36
purpose	71,87	73,71	72,70
condition	73,60	65,75	69,36
preparation	57,82	81,09	67,49
cause	51,73	69,96	59,46
contrast	68,43	56,69	56,69
sequence	54,46	54,55	54,22
evidence	44,75	34,53	38,95
comparison	50,43	31,25	38,49
evaluation	31,89	17,46	22,56
background	24,09	5,15	8,41

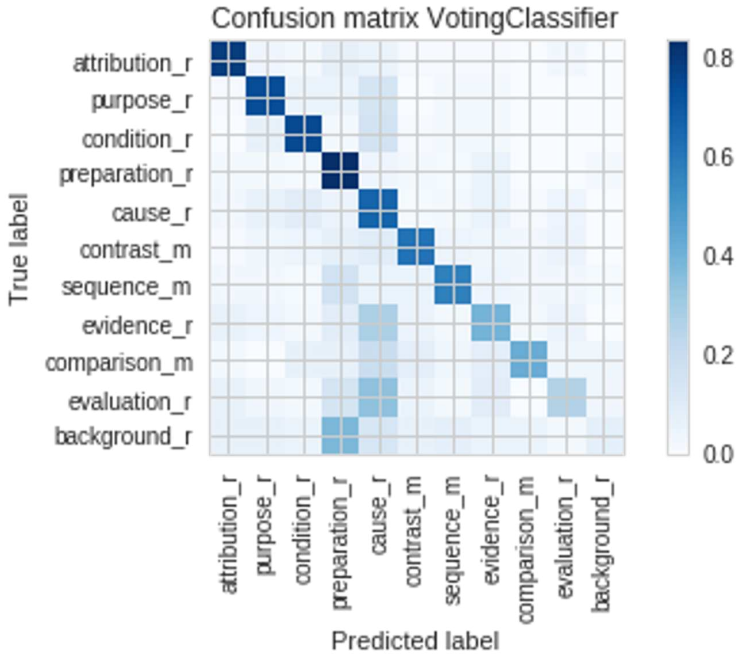


Рисунок 2.2 — Матрица ошибок лучшего классификатора

Подготовка – Атрибуция, ошибки могут быть вызваны стилистическими различиями в новостных текстах и научных статьях, входящих в набор данных.

2.5. Выводы

В главе 2 рассмотрена задача анализа риторических отношений в русскоязычном дискурсе. Предложен подход к представлению дискурсивных единиц, предусматривающий анализ лексических, синтаксических и семантических признаков. Предложенные методы классификации риторических отношений в русскоязычном дискурсе продемонстрировали важность широкого спектра лексических, морфологических и семантических признаков для решения задачи классификации риторических отношений. Экспериментальные исследования показали, что наилучшие результаты достигаются с использованием моделей градиентного бустинга и ансамблей классификаторов. Фильтрация признаков повысила эффективность классификации на основе небольшого размеченного корпуса с признаковым пространством высокой размерности. Результаты исследования опубликованы автором в работах “Classification models for RST discourse parsing of texts in Russian” [6] и “Towards the Data-driven System for Rhetorical Parsing of Russian Texts” [7].

Результаты анализа отношений в русскоязычном письменном дискурсе и предложенные методы классификации применены при разработке риторических анализаторов для русского языка, описанных в главе 3, а также при адаптации большого корпуса полнотекстовой дискурсивной разметки с английского на русский язык с целью исследования возможностей кросс-языкового переноса риторического разбора (глава 4).

Глава 3. Методы поверхностного разбора риторической структуры текста на русском языке

3.1. Введение

В данной главе представлены методы построения поверхностных риторических деревьев для текстов на русском языке с использованием восходящего и нисходящего подходов к риторическому разбору. Корпус RuRSTreebank [48], на котором основано исследование, представляет собой первый крупный корпус текстов на русском языке, размеченный в соответствии с Теорией риторических структур. Важной особенностью этого корпуса является частичная разметка текстов: в документах выделены риторические поддеревья с произвольными границами (до 42 поддеревьев на документ), при среднем значении в 11,7 поддеревьев на документ. Такая частичная («поверхностная») разметка усложняет задачу дискурсивного анализа и требует разработки специальных методов, позволяющих реконструировать лес риторических деревьев из каждого текста, согласующихся с имеющейся разметкой риторических отношений

В рамках разработки подходов к поверхностному разбору риторической структуры русскоязычных текстов в диссертационном исследовании предложены методы сегментации текста на элементарные дискурсивные единицы, а также методы классификации риторических отношений между дискурсивными единицами. Предложенные методы классификации риторических отношений в рамках риторического анализа включают как методы классического машинного обучения с широким набором признаков дискурсивных единиц, так и подходы на основе глубокого обучения, а также ансамблирование моделей различных типов.

В этой главе предлагаются два подхода к поверхностному разбору риторической структуры текста. Первый подход — жадный восходящий разбор, основанный на линейной свёртке пар дискурсивных единиц в соответствии с локальной оценкой вероятности формирования ими дискурсивного отношения; он позволяет формировать лес риторических деревьев документа в соответствии со статистиками отношений в размеченном корпусе. Второй подход — нисходящий разбор локальных дискурсивных структур с лучевым поиском, основанный на глубоком обучении; показано, что данный подход позволяет наиболее эффективно формировать риторическую структуру в пределах абзаца.

3.2. Дискурсивная сегментация

Дискурсивная сегментация представляет собой задачу разбиения текста на элементарные дискурсивные единицы (ЭДЕ) — минимальные фрагменты, обладающие самостоятельным значением, внутри которых нельзя выделить дополнительных дискурсивных отношений. В рамках теории риторических структур ЭДЕ выступают в роли терминальных узлов риторического дерева. В корпусе риторической разметки для русского языка RRT к ЭДЕ относятся [113]:

- Клаузы — простые предложения и равноправные части сложносочинённых предложений;
- Придаточные обстоятельственные конструкции;
- Придаточные дополнительные, если они входят в отношение атрибуции;
- Описательные придаточные, определительные обороты и причастные конструкции, если они вносят дополнительную (обособляемую) информацию;

– Деепричастные обороты, имеющие причинно-следственное или уточняющее значение;
– Уточняющие конструкции с союзом «то есть» и конструкции, заключённые в скобки, при условии наличия предикации;
– Некоторые предложные группы, выражающие причинно-следственные, уступительные или отношения контраста (например, с маркерами «из-за», «для», «с учетом», «несмотря на»);
– Предложения с именами действия, в которых риторическое отношение выражено эксплицитно (например, «X является/стал/был причиной/следствием/свидетельством Y»).

Таким образом, дискурсивная сегментация не сводится к делению предложения на клаузы и требует глубокого анализа дискурсивных ролей высказываний.

Рассмотрим текстовый документ $T = \{t_1, t_2, \ldots, t_n\}$, представленный в виде последовательности токенов, где t_i — i-й токен текста, а n — общее количество токенов в документе. В системах автоматического дискурсивного анализа данная задача решается следующим образом. Для каждой позиции i в последовательности токенов T необходимо присвоить метку класса $y_i \in \{B, I\}$, которая указывает на позицию токена относительно границ сегментации. Метка B (начало, англ. “Begin”) означает, что токен t_i является первым токеном новой ЭДЕ; I (внутри, англ. “Inside”) означает, что токен t_i находится внутри текущей ЭДЕ, но не является её началом. Моделирование структуры дискурса в теории риторических структур как дерева составляющих гарантирует, что каждый токен документа состоит в одной из элементарных единиц. Постановка задачи предусматривает указание в решении как левой, так и правой границы ЭДЕ: в ряде исследований [114;115] задачу решают с точки зрения предсказания первого токена ЭДЕ, в то время как в других работах [76; 116] предсказывают последний токен ЭДЕ. С точки зрения рассматриваемых в настоящей работе методов

сегментации на основе глубокого обучения различия в постановке двух задач малозначительны.

Автоматическая дискурсивная сегментация, таким образом, сводится к задаче разметки последовательностей, целью которой является определение такой разметки $Y = \{y_1, y_2, \ldots, y_n\}$, которая максимально точно соответствует корректному делению текста на ЭДЕ. Иллюстрация из корпуса RRT (см. рисунок 3.1) демонстрирует, что для задачи характера низкая сбалансированность классов, а класс токена относительно принадлежности к границе ЭДЕ зависит от его контекста в предложении.

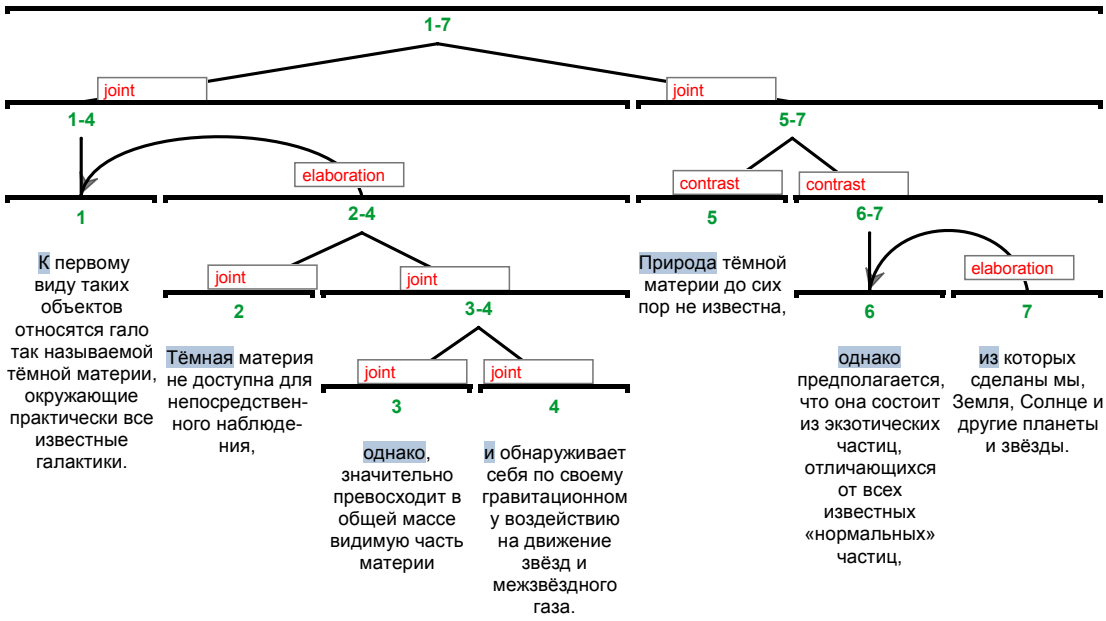


Рисунок 3.1 — Пример дискурсивной сегментации в риторической структуре (`RRT:news1_12`). Цветом выделены токены, для которых $y_i = B$.

Для решения задачи сегментации текста как разметки последовательности токенов предлагается модель глубокого обучения на основе архитектуры BiLSTM-CRF. Сегментаторы на основе этой архитектуры продемонстрировали лучшее качество решения задачи на различных языках в рамках соревнования DISRPT 2019 [117;118].

Входными данными является последовательность векторных представлений токенов документа $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, являющихся конкатенацией предобученных эмбеддингов токенов и результата свёртки

последовательности символов: $x_j = [\mathrm{emb}(t_j), \mathrm{CNN}(t_j)]$. Архитектура модели включает:

1. Один слой двунаправленной LSTM. Входная последовательность векторов токенов обрабатывается в прямом и обратном направлении, что позволяет получить для каждого элемента последовательности скрытое состояние в двух направлениях, с одинаковой эффективностью обрабатывая левый и правый контекст токена:

$$\begin{aligned} \overrightarrow{\mathbf{h}}_t &= \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overrightarrow{\mathbf{h}}_{t-1}), \\ \overleftarrow{\mathbf{h}}_t &= \mathrm{LSTM}_{\mathrm{backward}}(x_t, \overleftarrow{\mathbf{h}}_{t+1}), \\ \mathbf{h}_t &= \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right], \end{aligned} \tag{3.1}$$

   где $\mathbf{h}_t$ — объединённое скрытое состояние для токена t.

   Для предотвращения переобучения к объединённым скрытым состояниям применяется метод случайного разреживания (dropout), а также нормализация выходов внутреннего слоя.

2. CRF-слой. Входными данными является последовательность скрытых состояний $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$. В этом слое вычисляется совместное распределение вероятностей для последовательности меток $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. Функция оценки вероятностей для последовательности меток определяется как:

$$s(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} (\mathbf{W}_{y_t} \mathbf{h}_t + b_{y_t}) + \sum_{t=1}^{T-1} \mathbf{A}_{y_t, y_{t+1}} \tag{3.2}$$

   где $\mathbf{W}_{y_t}$ и b_{y_t} — обучаемые параметры для метки y_t, а $\mathbf{A}_{y_t, y_{t+1}}$ — параметр перехода от метки y_t к метке y_{t+1}.

3.3. Классификация риторических отношений

Поверхностный анализ риторической структуры текста требует решения двух задач классификации, а именно разработки *структурного классификатора* и *классификатора отношения и ядерности.*

На шаге j рекурсивного построения леса риторических структур рассматривается множество дискурсивных единиц:

$$D_j = \{\mathrm{du}_1, \mathrm{du}_2, \ldots, \mathrm{du}_m\}. \tag{3.3}$$

Для построения структур составляющих из набора дискурсивных единиц используется *структурный классификатор*:

$$f_{struct} : D_j \times D_j \to \{0{,}1\},$$

где $f_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1}) = 1$ означает, что соседние дискурсивные единицы du_i и du_{i+1} следует объединить в новую дискурсивную единицу верхнего уровня $\mathrm{du}_{i:i+1} = \mathrm{du}_i \oplus \mathrm{du}_{i+1}$, которая войдёт в D_{j+1}. Если же $f_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1}) = 0$, то непосредственное объединение этих единиц при построении риторического дерева невозможно. При выполнении поверхностного разбора риторической структуры с обучением моделей на частично размеченном корпусе структурный классификатор позволяет воспроизводить риторические отношения на тех уровнях дискурса, для которых имеется достаточный для обучения объём экспертной разметки.

При реализации риторического анализа методами машинного обучения структурный классификатор, обучаемый задаче бинарной классификации на размеченных в корпусе ($f_{struct} = 1$) и синтетических[1] ($f_{struct} = 0$) примерах, генерирует оценочную величину $P_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1})$, интерпретируемую как вероятность наличия риторической связи между указанными

[1]Отрицательные примеры формируются из соседних пар дискурсивных единиц в оригинальном порядке, не объединённых риторическим отношением в размеченном корпусе. Этот подход не гарантирует, что в иных контекстах аналогичная пара не может формировать общее риторическое отношение.

единицами. Если эта оценка вероятности превышает заданное пороговое значение $\tau \in \mathbb{R}$ ($P_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1}) > \tau$), то две дискурсивные единицы объединяются в новую $\mathrm{du}_{i:i+1} \in D_{j+1}$. Этот процесс повторяется до тех пор, пока существует хотя бы одна пара соседних единиц, для которых оценка вероятности превышает τ.

После построения структуры составляющих к каждой объединённой паре единиц применяется *классификатор отношения и ядерности* f_{rel}. С его помощью одновременно определяется тип риторического отношения и позиция ядра между du_i и du_{i+1}:

$$f_{rel} : D \times D \to R, \quad R = \{\text{Конкатенация\_NN}, \text{Детализация\_NS}, \dots\}. \tag{3.4}$$

Совместное предсказание данных характеристик позволяет уменьшить накопление ошибок в риторическом анализе и избежать недопустимых комбинаций типов отношений и расположения ядер.

Таким образом, 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 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.

3.3.1. Методы классификации на основе признакового описания

Исходя из анализа риторических отношений в русскоязычном письменном дискурсе, представленного в главе 2, разработаны усовершенствованные методы классификации риторических отношений для поверхностного анализа дискурсивной структуры.

Набор признаков

Базовый набор признаков описан в разделе 2.3. Для решения задач классификации на основе важности признаков в модели градиентного бустинга, обученной по схеме кросс-валидации, предварительно исключаются дискурсивные маркеры, не релевантные конкретной задаче. Вместо word2vec-векторов используется предобученная модель ELMo, учитывающая контекст при кодировании токенов [119]. Кроме того, введены дополнительные признаки:

1. Признаки нахождения двух дискурсивных единиц на одном уровне детализации: в одном предложении, в одном абзаце.
2. Тональность каждой дискурсивной единицы.

Первый признак наиболее важен для решения задачи структурной классификации, поскольку помогает выявлять целостные структуры на каждом уровне. Хотя теория риторических структур формально не ограничивает построение отдельной структуры для каждого предложения или абзаца, на практике учёт «границ» предложения и абзаца способствует обнаружению закономерных разрывов в размеченных данных. Второй признак направлен на выявление эмоционального контраста между дис-

курсивными единицами, что может быть полезно при классификации отношений уступки, оценки и контраста.

Набор признаков, описанный в разделе 2.3, обозначим как baseline, а расширенный вариант, используемый в анализаторе, — baseline + AF.

Структурный классификатор

В качестве основного метода структурной классификации применяется метод опорных векторов (SVM) с линейным ядром. Предварительные эксперименты показали, что в задачах локальной классификации пар дискурсивных единиц и построения структуры составляющих сложные ансамбли (например, с градиентным бустингом) и дополнительная фильтрация признаков не дают значимого прироста качества. При создании риторического анализатора необходимо также учитывать вычислительную эффективность итогового анализатора, включающего множество моделей.

Классификатор отношения и ядерности

Для задачи классификации типа риторического отношения применяется ансамбль, состоящий из модели градиентного бустинга (CatBoost) на признаках, отобранных методом лассо, и SVM с линейным ядром. Рассмотрим варианты конфигурации классификатора.

Базовый метод. Метод, продемонстрировавший лучшие результаты в исследовании, описанном в главе 2, был адаптирован к задаче совместной классификации типа отношения и ядерности. Рассматриваются все классы «тип отношения + ядерность», размеченные в корпусе.

Базовый метод с дополнительными признаками. В данном случае к признакам базового метода добавляются новые признаки (baseline + AF). Это позволяет уточнить представление о связях между дискурсивными единицами и улучшить точность классификации.

3.3.2. Методы классификации на основе глубокого обучения

Для решения задач классификации пар дискурсивных единиц в подходе на основе глубокого обучения используется модель многостороннего симметричного сопоставления BiMPM (Bilateral Multi-Perspective Matching) [120], ранее применявшаяся для задач семантического сопоставления высказываний [121]. В разработанном риторическом анализаторе модель архитектуры BiMPM используется для выявления и классификации риторических отношений и между дискурсивными единицами на всех уровнях дискурса.

Модель BiMPM последовательно обрабатывает две дискурсивные единицы, кодируя токены при помощи двунаправленного LSTM-кодировщика, затем сопоставляет полученные контекстные векторы несколькими способами (многостороннее сопоставление). Итоговые векторы агрегируются и передаются в полносвязный слой с softmax-активацией для предсказания класса. В качестве базового векторного представления используются контекстные эмбеддинги из предобученной языковой модели и обучаемые заново символьные эмбеддинги.

Ниже приведены основные этапы классификации пары дискурсивных единиц $(\mathrm{du}_i, \mathrm{du}_{i+1})$ в модели BiMPM.

1. На первом этапе каждому токену $t_{i,j}$ в двух дискурсивных единицах $\mathrm{du}_i = (t_{i,1}, ..., t_{i,M})$ и $\mathrm{du}_{i+1} = (t_{i+1,1}, ..., t_{i+1,N})$ ставится в соответствие векторное представление. Вектор токена представляет собой

конкатенацию эмбеддинга из предобученной модели и символьного эмбеддинга:

$$x_j = \big[\mathrm{emb}(t_j); \mathrm{charGRU}(t_j)\big], \tag{3.5}$$

где $\mathrm{emb}(t_j)$ — эмбеддинг токена в предобученной кодирующей модели, а $\mathrm{charGRU}(t_j)$ — символьный эмбеддинг, являющийся результатом преобразования последовательности символов ДЕ в обучаемом GRU-слое [122].

2. На втором этапе получают контекстные представления токенов каждой ДЕ в локальном контексте дискурсивной единицы. Для этого используется LSTM-слой с d скрытыми нейронами, применяемый к векторам токенов ДЕ в прямом и обратном направлении:

$$\begin{aligned} \overrightarrow{\mathbf{h}}_t^{\mathrm{du}_i} &= \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overrightarrow{\mathbf{h}}_{t-1}^{\mathrm{du}_i}), \\ \overleftarrow{\mathbf{h}}_t^{\mathrm{du}_i} &= \mathrm{LSTM}_{\mathrm{backward}}(x_t, \overleftarrow{\mathbf{h}}_{t+1}^{\mathrm{du}_i}), \\ \mathbf{h}_t^{\mathrm{du}_i} &= \big[\overrightarrow{\mathbf{h}}_t^{\mathrm{du}_i}; \overleftarrow{\mathbf{h}}_t^{\mathrm{du}_i}\big]. \end{aligned} \tag{3.6}$$

Аналогично вычисляются представления для du_{i+1}.

3. На третьем этапе выполняется сопоставление контекстных представлений токенов левой и правой дискурсивной единицы несколькими способами. Для этого используется функция многостороннего косинусного сходства f_m, которая принимает два вектора и обучаемую матрицу весов $\mathbf{W}$ для каждого из способов сопоставления:

$$m = f_m(\mathbf{v_1}, \mathbf{v_2}; \mathbf{W}), \tag{3.7}$$

где $\mathbf{v_1}, \mathbf{v_2}$ — векторы токенов размерности $d \times 2$, $\mathbf{W} \in \mathbb{R}^{l \times d \times 2}$ — обучаемая матрица многостороннего сопоставления. Результатом применения функции является вектор $m = [m_1, m_2, \ldots, m_l]$ значений косинусного сходства векторов токенов с учётом взвешивания каждой строкой матрицы многосторонней оценки $\mathbf{W}$:

$$m_k = cosine(\mathrm{W}_k \circ \mathbf{v_1}, \mathrm{W}_k \circ \mathbf{v_2}). \tag{3.8}$$

В архитектуре BiMPM используется четыре стратегии сопоставления. Представления токенов в результате кодирования в LSTM-слое в прямом и обратном порядке сравниваются по отдельности.

а) Полное сопоставление заключается в вычислении косинусного сходства между векторами токенов одной ДЕ и последним выходом кодировщика при обработке второй ДЕ:

$$\begin{aligned} \overrightarrow{m}_t^{\text{full}} &= f_m(\overrightarrow{\mathbf{h}}_t^{\text{du}_i}, \overrightarrow{\mathbf{h}}_N^{\text{du}_{i+1}}; \overrightarrow{\mathrm{W}}_{\text{full}}), \\ \overleftarrow{m}_t^{\text{full}} &= f_m(\overleftarrow{\mathbf{h}}_t^{\text{du}_i}, \overleftarrow{\mathbf{h}}_1^{\text{du}_{i+1}}; \overleftarrow{\mathrm{W}}_{\text{full}}), \end{aligned} \tag{3.9}$$

где $\overrightarrow{\mathrm{W}}_{\text{full}}$ и $\overleftarrow{\mathrm{W}}_{\text{full}}$ — обучаемые матрицы полного сопоставления.

б) При сопоставлении с максимизацией для каждой позиции токена в первой ДЕ выбирается максимальное значение косинусного сходства с токенами второй ДЕ:

$$\overrightarrow{m}_t^{\max} = \max_{j \in (1 \ldots N)} f_m(\overrightarrow{\mathbf{h}}_t^{\text{du}_i}, \overrightarrow{\mathbf{h}}_j^{\text{du}_{i+1}}, \overrightarrow{\mathrm{W}}_{\max}). \tag{3.10}$$

в) Сопоставление с учетом внимания требует вычисления весов контекстных векторов правой дискурсивной единицы как косинусной близости между токенами:

$$\overrightarrow{\alpha}_{t,j} = cosine(\overrightarrow{\mathbf{h}}_t^{\text{du}_i}, \overrightarrow{\mathbf{h}}_j^{\text{du}_{i+1}}). \tag{3.11}$$

Коэффициенты внимания $\alpha_{t,j}$, нормализованные softmax, используются для вычисления вектора внимания токенов правой ДЕ:

$$\overrightarrow{\mathbf{h}}_t^{\text{att}} = \sum_{j=1}^{N} \frac{\exp(\overrightarrow{\alpha}_{t,j} \cdot \overrightarrow{\mathbf{h}}_j^{\text{du}_{i+1}})}{\sum_{k=1}^{N} \exp(\overrightarrow{\alpha}_{t,k})}. \tag{3.12}$$

Таким образом представления правой дискурсивной единицы модифицируются с учётом контекста левой ДЕ. Это позволяет сосредотачиваться при классификации на наиболее важных токенах в правой ДЕ при условии информации,

данной в левой ДЕ, что особенно полезно для анализа сложных риторических отношений. Теперь каждый контекстный вектор $\mathbf{h}_t^{\mathrm{du}_i}$ первой дискурсивной единицы сопоставляется с соответствующим модифицированным вектором $\mathbf{h}_t^{\mathrm{att}}$ второй ДЕ с использованием функции многостороннего косинусного сходства:

$$\overrightarrow{m}_t^{\mathrm{att}} = f_m(\overrightarrow{\mathbf{h}}_t^{\mathrm{du}_i}, \overrightarrow{\mathbf{h}}_t^{\mathrm{att}}; \overrightarrow{\mathbf{W}}_{\mathrm{att}}). \tag{3.13}$$

г) При вычислении максимального значения внимания вместо нормализации весов внимания (3.12) для каждого токена первой ДЕ выбирается наиболее релевантный токен из второй ДЕ. Он определяется максимальным значением коэффициента внимания:

$$j^* = \arg \max_{j \in (1,\ldots,N)} \alpha_{t,j}. \tag{3.14}$$

Выбор наиболее релевантного токена делает модель более чувствительной к сильным семантическим связям между отдельными словами двух дискурсивных единиц.

4. Результаты всех сопоставлений в прямом и обратном направлении конкатенируются в общий вектор сопоставлений $\mathbf{m}_i$. На основе агрегированного вектора осуществляется классификация с использованием полносвязного слоя и функции softmax:

$$\hat{y}_i = \mathrm{softmax}(W_{\mathrm{pred}} \cdot \mathbf{m}_i + b_{\mathrm{pred}}). \tag{3.15}$$

Описанная архитектура используется для классификации отношения и ядерности на основе контекстных представлений дискурсивных единиц. Структурный классификатор также использует бинарные признаки, указывающие на нахождение дискурсивных единиц на одном уровне детализации, то есть в одном предложении ($g_s : D_j \times D_j \to \{0,1\}$) или в одном абзаце

($g_p : D_j \times D_j \to \{0,1\}$). Эти признаки учитываются на этапе 4 классификации с использованием метода BiMPM. Для финальной классификации значения уровня детализации конкатенируются с вектором сопоставлений:

$$\hat{y}_i = \mathrm{softmax}\Big(W_{\mathrm{pred}} \cdot \big[\mathbf{m}_i,\ g_s(\mathrm{du}_i, \mathrm{du}_{i+1}),\ g_p(\mathrm{du}_i, \mathrm{du}_{i+1})\big] + b_{\mathrm{pred}}\Big). \tag{3.16}$$

Признаки нахождения дискурсивных единиц на одном уровне детализации являются важными для построения дерева, в котором большинство [123] предложений или абзацев формируют обособленные дискурсивные единицы.

3.3.3. Ансамблирование классификаторов

При построении анализатора риторической структуры целесообразно сочетать несколько методов и моделей, компенсируя их недостатки. Ансамблирование реализуется путём комбинирования вероятностных оценок, полученных от нескольких классификаторов разных типов.

Классические методы на основе признакового описания позволяют работать с интерпретируемыми признаками разной природы. Это облегчает исследование свойств риторических связей и упрощает отладку риторического анализатора. Признаковые модели могут показывать хорошие результаты даже с относительно небольшими обучающими выборками. Такие модели более устойчивы к несбалансированным классам, поскольку результат определяется разделяющими классы признаками, а также требуют меньше вычислительных ресурсов. Методы глубокого обучения, напротив, могут автоматически выявлять сложные лексические и семантические закономерности между высказываниями без ручного проектирования признаков и зачастую превосходят классические алгоритмы по ключевым метрикам, если доступен достаточный объём обучающих данных.

Однако они требуют значительного объёма размеченных данных и могут быть чувствительны к их недостатку, а также дисбалансу классов.

В предложенном поверхностном риторическом анализаторе обе задачи классификации решаются ансамблированием двух типов моделей. Структурная классификация реализуется ансамблем, сочетающим модель опорных векторов (SVM) и глубокое симметричное сопоставление (BiMPM) с дополнительным учётом признаков уровней детализации. Для классификации типа отношения и ядерности применяется голосование ансамбля, состоящего из ансамбля моделей машинного обучения на признаках baseline+AF и BiMPM-модели. В качестве способа ансамблирования выбран стекинг, где для мета-классификации оценок вероятности используется логистическая регрессия. Этот подход позволяет существенно повысить итоговую точность классификации за счёт оптимального комбинирования сильных сторон различных классификаторов.

Общая схема решения задач классификации в поверхностном риторическом анализаторе изображена на рисунке 3.2.

3.4. Построение риторического леса

Основная сложность в реализации риторического анализа на основе русскоязычного корпуса RRT заключается в неопределенности границ размеченных деревьев внутри документа, что затрудняет построение предсказательных моделей. По этой причине первые эксперименты по риторическому анализу текстов на русском языка посвящены *поверхностному разбору* — построению леса локальных риторических деревьев, имитирующего частичную разметку в корпусе RRT. Построение риторического леса включает в себя разбор локальных структур, описывающих отношения между дискурсивными единицами на нижних уровнях текста. Для реше-

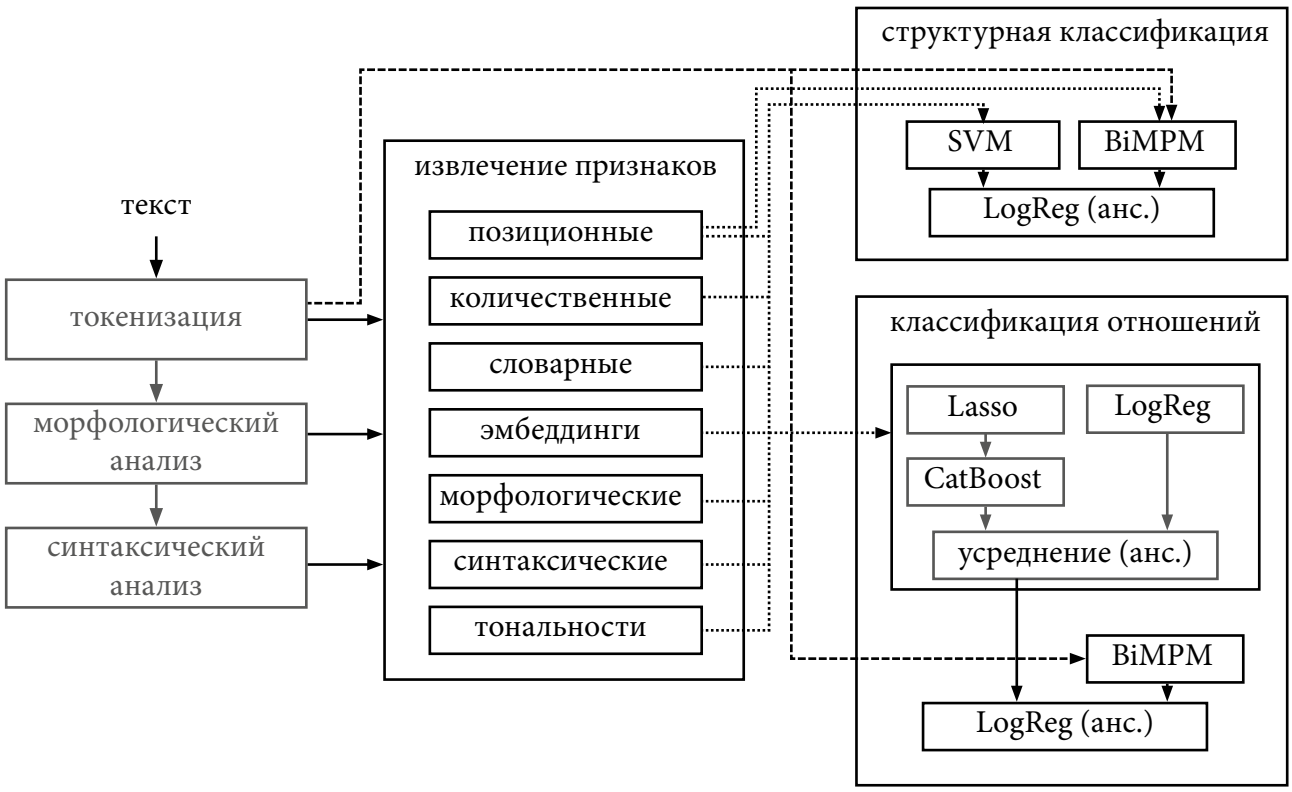


Рисунок 3.2 — Схема решения задач классификации в риторическом анализаторе.

ния этой задачи разработаны методы на основе как классических методов машинного обучения, так и глубокого обучения.

3.4.1. Жадный разбор локальных риторических структур

В качестве базового метода построения леса локальных риторических структур принят непосредственный жадный восходящий разбор с порогом уверенности. Анализатор последовательно объединяет соседние ДЕ на основе предсказанной оценки вероятности $P_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1})$ существования между ними риторических отношений до тех пор, пока структурный классификатор с некоторой уверенностью может предсказать наличие отношений между двумя дискурсивными единицами (т.е., пока уровень дискурсивной структуры, порождаемой восходящим разбором, соответствует размеченным примерам, на которых обучен структурный классификатор).

Формально, на каждом шаге разбора выбирается пара ДЕ $(\mathrm{du}_i^*, \mathrm{du}_{i+1}^*)$ с максимальной предсказанной вероятностью:

$$(\mathrm{du}_i^*, \mathrm{du}_{i+1}^*) = \underset{(\mathrm{du}_i, \mathrm{du}_{i+1}) \in D_j}{\operatorname{argmax}} P_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1}), \tag{3.17}$$

где D_j — множество всех пар ДЕ на текущем шаге j разбора дерева. Если вероятность $P_{struct}(\mathrm{du}_i^*, \mathrm{du}_{i+1}^*)$ превышает заданный порог τ_k, эти две единицы объединяются в новую дискурсивную единицу верхнего уровня $\mathrm{du}_{i:i+1}$. Этот процесс продолжается до тех пор, пока остаются пары, для которых вероятность объединения превышает порог τ_k.

В базовой конфигурации предложенного анализатора алгоритм применяется последовательно к каждому уровню детализации: предложению, абзацу, документу. На уровне предложения исходными дискурсивными единицами являются элементарные дискурсивные единицы, границы которых определяются обученным сегментатором. Для абзацев и документов в качестве исходного набора ДЕ используются дискурсивные единицы, сформированные на предыдущем уровне. Построение дерева завершается, если либо не остаётся пар единиц, которые можно объединить в единицу более высокого уровня, либо вероятность объединения для всех оставшихся пар оказывается меньше порога τ_k, заданного для соответствующего уровня детализации, то есть выполняется условие:

$$\forall (\mathrm{du}_i, \mathrm{du}_{i+1}) \in D_j \quad P_{struct}(\mathrm{du}_i, \mathrm{du}_{i+1}) < \tau_k.$$

Формальное описание алгоритма представлено на рисунке 3.3.

Вероятность $P_{struct}(du_i, du_{i+1})$ вычисляется с использованием бинарного структурного классификатора. Отрицательные примеры для обучения классификатора формируются из соседних фрагментов текста, которые не образуют риторическое отношение в корпусе. Порог τ_k, где $k \in \{\text{предложение}, \text{абзац}, \text{документ}\}$, является гиперпараметром, значение которого устанавливается для каждого уровня детализации эмпирически. Выбор такой статегии обусловлен представлением каждого документа в корпусе не

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 3.3 — Жадный алгоритм построения риторического леса

в виде отдельного риторического дерева, но набора деревьев произвольной длины. Система дискурсивного анализа предусматривает обучение одного структурного классификатора, но для каждого уровня детализации используются разные значения порога уверенности. Для построения деревьев внутри абзаца этот порог значительно ниже, чем для полного документа, но принимает строго положительные значения, поскольку абзац в корпусе не всегда представляет собой отдельное поддерево: разные части абзаца могут не соединяться, но быть частями соседних деревьев абзацного уровня. По той же причине порог уверенности для построения дерева на уровне предложения ниже порога для абзацев, но также строго положителен.

3.4.2. Частичный нисходящий разбор с лучевым поиском

Для улучшения точности построения отдельных риторических деревьев внутри документа предлагается использовать метод нисходящего

разбора, основанный на глубоком обучении с ранжированием вариантов и применением лучевого поиска. В отличие от полностью жадного алгоритма построения леса риторических структур, описанного в подразделе 3.4.1, данный метод учитывает локальный контекст абзаца и позволяет формировать более точные внутренние структуры.

Базовый алгоритм построения леса риторических структур, описанный в 3.4.1, допускает, что целый абзац может не являться самостоятельной дискурсивной единицей ($\tau_{\text{абзац}} < 1$). Анализ корпуса $\text{RRT}_{\text{v2.0}}$, на материале которого проводились исследования методов построения поверхностных структур в текстах на русском языке, показал, что лишь около 15% абзацев не формируют полноценную дискурсивную единицу. Однако в большинстве случаев возможно использовать для разбора локальных подструктур абзацев методы с оптимизацией внутренних представлений на основе глобального контекста известной наибольшей дискурсивной единицы.

Алгоритм частичного нисходящего разбора

Метод, описываемый в этом подразделе, предполагает построение модели глубокого обучения, которая заменяет структурную классификацию с жадной свёрткой на уровнях предложений и абзацев, обеспечивая более точное построение риторических структур внутри отдельных абзацев документа. В отличие от жадного алгоритма восходящего разбора, который рассматривает каждую пару ДЕ в отрыве от контекста, поскольку границы формируемого дерева неизвестны заранее, предложенный подход позволяет приравнять границы формируемых поддеревьев к границам абзацев и эффективно ранжировать возможные риторические подструктуры внутри абзаца. Лес риторических структур на уровне документа при этом все еще формируется жадным восходящим разбором, благодаря чему воз-

можно совмещение деревьев отдельных абзацев в связные дискурсивные единицы верхних уровней.

Механизм лучевого поиска. Лучевой поиск (beam search) — это эвристический алгоритм, который предусматривает рассмотрение множества наиболее перспективных узлов на каждом уровне дерева поиска. Ширина луча b ограничивает количество рассматриваемых вариантов. В контексте нисходящего разбора риторических структур внутри абзаца лучевой поиск позволяет эффективно находить оптимальные разбиения дискурсивных единиц с учётом глобальной вероятности построения корректного поддерева. На каждом шаге нисходящего разбора абзаца рассматриваются все потенциальные способы разбиения текущей дискурсивной единицы (абзац или фрагмент абзаца) на две составляющие. Для каждой возможной структуры вычисляется локальная вероятность, отражающая степень корректности риторического разбиения. Лучевой поиск позволяет одновременно рассматривать b наиболее перспективных вариантов разбиения, последовательно их уточняя и отбрасывая варианты с низкой оценкой вероятности. В результате удаётся найти оптимальное или близкое к оптимальному разбиение дискурсивной единицы, максимизирующее глобальную вероятность построения корректного поддерева абзаца.

Интеграция в анализатор поверхностных риторических структур. Поддеревья абзацев формируются с использованием описанного лучевого поиска, после чего полученные структуры совмещаются в единый лес риторических структур на уровне всего документа. Для этого применяется жадный восходящий алгоритм (рис. 3.3), который позволяет при необходимости связывать дискурсивные единицы-абзацы, исходя из вероятностей, вычисленных моделью для каждого потенциального отношения. Таким образом, глобальная структура документа формируется комбинированием

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, ..., e_n]) ← [h_0, h_1, ..., h_m]                ▷ Состояния кодировщика
L_d ← |E| − 1                                                   ▷ Длина декодирования
beam ← 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 ← 1 to L_d do
    for (logp, s, input_span, tree) ∈ beam[t − 1] do
        (i, j) ← input_span[t − 1]                              ▷ Текущая ДЕ для разбиения
        a, s' ← decoder-step(s, h_{i,j})                        ▷ a — распред. вер. разбиения
        for (k, p_k) ∈ top-B(a) and i < k < j do
            curr_input_span ← input_span
            curr_tree ← tree
            curr_tree[t − 1] ← (i, k, j)
            if k > i + 1 then
                curr_input_span[t] ← (i, k)
            end if
            if j > k + 1 then
                curr_input_span[t + j − k − 1] ← (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*) ← arg max_{logp} beam[L_d]                 ▷ S* – лучшая структура
```

Рисунок 3.4 — Алгоритм нисходящего разбора риторического дерева с использованием лучевого поиска

поддеревьев абзацев, моделирующим связность риторических единиц на более высоких уровнях.

Основные шаги алгоритма построения структуры внутри абзаца представлены на рис. 3.4. На каждом этапе разбиения рассматривается набор возможных риторических поддеревьев с выбором b лучших (по значению вероятности). Поддерево уточняется до тех пор, пока не будет найдена наиболее оптимальная дискурсивная структура. Этот подход повышает качество анализа риторической структуры. Его применение наиболее полезно в тех случаях, когда абзац содержит множество дискурсивно связанных высказываний, и построение верной структуры возможно только с учётом всех дискурсивных единиц в данном контексте.

Модель разбора абзаца на основе глубокого обучения

В контексте нисходящего разбора риторических структур внутри абзаца предлагается использовать модель глубокого обучения, на каждом этапе оценивающую вероятность различных вариантов разбиений дискурсивных единиц с целью построения оптимального поддерева. Алгоритм лучевого поиска позволяет эффективно управлять процессом поиска оптимального дерева абзаца, ограничивая количество рассматриваемых вариантов на каждом шаге и снижая вычислительную сложность разбора документа. Модель принимает на вход последовательность элементарных дискурсивных единиц, полученных на этапе сегментации, и на их основе восстанавливает иерархическую структуру абзаца. Архитектура, предложенная в [75] для решения задачи полнотекстового риторического анализа, включает кодировщик для получения контекстных представлений ЭДЕ и декодировщик для предсказания вероятностей разбиений и отношений между ними. В описываемом анализаторе поверхностных риторических структур для русского языка модель этой архитектуры используется только для построения дерева составляющих из заданного набора ЭДЕ; метод сегментации описан в разделе 3.2, а модуль классификации типа риторического отношения и ядерности — в разделе 3.3. Далее приводится описание основных этапов обработки последовательности ЭДЕ для получения не размеченной типами отношений структуры дерева абзаца.

1. Входными данными является последовательность векторных представлений токенов документа $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, являющихся конкатенацией:

$$x_i = \big[\mathrm{emb}(t_i); \mathrm{charBiLSTM}(t_i)\big], \tag{3.18}$$

где $\mathrm{emb}(t_i)$ — предобученный эмбеддинг токена, а $\mathrm{charBiLSTM}(t_i)$ — символьный эмбеддинг, полученный в обучаемом слое двунаправленной LSTM.

Последовательность векторов токенов обрабатывается в трёхслойном BiLSTM-кодировщике, формирующем контекстные представления токенов:

$$\overrightarrow{\mathbf{h}}_t^{(l)} = \mathrm{LSTM}_{\mathrm{forward}}^{(l)}\big(\overrightarrow{\mathbf{h}}_t^{(l-1)}, \overrightarrow{\mathbf{h}}_{t-1}^{(l)}\big), \quad l \in \{1,2,3\}, \quad \overrightarrow{\mathbf{h}}_t^{(0)} = x_t. \tag{3.19}$$

Выходы последнего слоя $\overrightarrow{\mathbf{h}}_t^{(3)}$ и $\overleftarrow{\mathbf{h}}_t^{(3)}$ используются в дальнейшем как контекстные представления токена t.

2. Для каждой ЭДЕ e_k формируется векторное представление на основе представлений её первого и последнего токенов:

$$\mathbf{v}_k = \big[\overrightarrow{\mathbf{h}}_{a_k}^{(3)}; \overleftarrow{\mathbf{h}}_{b_k}^{(3)}\big], \tag{3.20}$$

где a_k и b_k — индексы первого и последнего токена ЭДЕ. Благодаря двунаправленным LSTM состояния граничных токенов частично учитывают информацию о всей дискурсивной единице.

3. Реализуется нисходящий разбор в соответствии с алгоритмом на рис. 3.4. Последовательность векторов $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ является входными данными, а $\mathrm{du}_{1:m}$ — исходной ДЕ, покрывающей весь абзац от первой до последней ЭДЕ.

    а) На каждом шаге модель определяет оптимальную точку разбиения дискурсивной единицы $\mathrm{du}_{i:j}$ на две части: $\mathrm{du}_{i:k}$ и $\mathrm{du}_{k+1:j}$. Для этого декодировщик, основанный на однонаправленной LSTM, формирует скрытые состояния для рассматриваемой единицы, после чего полносвязный слой выбирает оптимальное значение k из множества возможных позиций $k \in \{i, \ldots, j-1\}$:

$$k^* = \underset{k \in \{i,\ldots,j-1\}}{\mathrm{argmax}} \; P_{\mathrm{split}}(i,k,j).$$

б) Для повышения надёжности разбора анализатор сохраняет b наиболее вероятных вариантов разбиения, что позволяет учитывать альтернативные структуры дерева и снижать вероятность накопления ошибок.

в) Процесс рекурсивно повторяется до тех пор, пока все наиболее вероятные ДЕ не оказываются разбиты на элементарные дискурсивные единицы.

3.5. Экспериментальное исследование методов поверхностного разбора риторической структуры

3.5.1. Описание данных

RRT [48] является первым открытым корпусом дискурсивной разметки для русскоязычных текстов в рамках Теории риторических структур. Он состоит из трёх частей: новостные и научно-популярные статьи (129 текстов); научные тексты, относящиеся к лингвистике и компьютерным наукам (99 текстов); блоги различной тематики: путешествия, спорт и здоровье, психология, IT и технологии, политика (104 текста). Корпус содержит около 408000 токенов, размечен 11 экспертами. Для обучения и оценки моделей из корпуса были исключены документы научного жанра. Научный подкорпус был размечен в ходе первого этапа разработки корпуса с учетом самых ранних инструкций разметчика, которые могут противоречить инструкциям для остальной части корпуса, а также содержат большое количество непроективных структур. Таким образом, дискурсивный анализатор основан на материале 233 текстов, включающем новостные и научно-популярные статьи, а также блоги.

Предобработка корпуса $RRT_{v2.0}$ включает корректировку исходных структур и отношений. Небинарные отношения, такие как Joint (Конкатенация) и Sequence (Последовательность), преобразуются в каскады бинарных отношений[2]. Было обнаружено, что знаки пунктуации, такие как точки и запятые, в некоторых случаях перемещены из начала предыдущей ЭДЕ в начало следующей, что исправляется в ходе предобработки. Некоторые типы отношений, указанные в [48] как объединенные или исключенные из корпуса, например, Antithesis (Антитезис) или Evaluation (Оценка), однако, встречаются в корпусе, и также предобрабатываются в соответствии с инструкциями в указанной работе. Низкочастотные отношения (менее 90 обучающих примеров) объединяются с наиболее частотными некаузальными классами с тем же типом нуклеарности. Таблица 7 описывает предобработку конкретных отношений. С учётом расположения ядра итоговый набор классов составляет 22 отношения.

Таблица 7 — Переименование отношений в предобработке корпуса RRT

В разметке	Предобработка
antithesis	Attribution
cause, effect, cause-effect	Cause-effect
condition, motivation	Condition
evaluation, interpretation, interpretation-evaluation	Interpetation-evaluation
restatement_SN	Condition_SN
restatement_NS	Elaboration_NS
solutionhood_NS	Solutionhood_SN
preparation_NS	Elaboration_NS
elaboration_SN	Preparation_SN
background_NS	Elaboration_SN

[2]Это стандартная предобработка риторической разметки для задач автоматического анализа, подразумеваемая далее для всех корпусов.

3.5.2. Детали экспериментов

Токенизация текста и определение границ предложений выполняются при помощи библиотеки UDPipe [124]. Для морфологического анализа используется библиотека MyStem[3]. Для получения векторных представлений токенов используются модели word2vec [108], fastText [125] и ELMo [119], обученные на материалах Национального корпуса русского языка и Википедии в рамках проекта RusVectores[4], а также мультиязычная языковая модель BERT[5] [16]. Веса предобученных моделей фиксированные. Признаки тональностей левой и правой дискурсивных единиц извлекаются с помощью классификатора на основе fastText, предложенного в работе [126].

Модели глубокого обучения для решения трёх задач: дискурсивной сегментации, классификации пар дискурсивных единиц и построения структуры составляющих на уровне абзаца, обучались на одной видеокарте с 12 Гб видеопамяти. Поскольку во всех задачах, кроме структурной классификации, классы несбалансированы, а структурная классификация требует обработки наибольшего объёма данных, наилучшие результаты получены при использовании небольшого размера обучающего батча.

Сегментация. В качестве контекстных языковых моделей для получения векторных представлений токенов в текстах на русском языке используются предобученная в рамках проекта RusVectores[6] на Национальном корпусе русского языка и русской Википедии модель ELMo [119] и модели BERT [16], предварительно обученные на больших мультиязыковом[7] и русскоязычном[8]

[3] https://yandex.ru/dev/mystem
[4] https://rusvectores.org/ru/models
[5] `bert-base-multilingual-cased`
[6] http://rusvectores.org/
[7] `bert-base-multilingual-cased`
[8] `DeepPavlov/rubert-base-cased`

корпусах. Веса предобученных моделей остаются фиксированными на протяжении всего обучения сегментаторов. При решении задачи дискурсивной сегментации для кодирования последовательности символов применялся сверточный слой с 128 фильтрами размера 3. Размер LSTM-кодировщика: 100. При обучении использовались следующие гиперпараметры: коэффициент `dropout` равен 0,5, скорость обучения (`LR`) равна 0,001, размер батча: 2.

Структурный классификатор. Для обучения бинарного классификатора на основе архитектуры BiMPM количество негативных обучающих примеров было уменьшено, что позволило достичь более сбалансированного распределения классов и сократить объём потребляемой памяти. Уменьшение объёма отрицательных примеров с сохранением наиболее репрезентативных реализовано методом центроидов [127]. Для этого пары дискурсивных единиц, не объединённые риторическим отношением в корпусе, закодированы признаками, описанными в разделе 3.3.1, после чего размерность векторов их представлений снижена до 50 при помощи сингулярного разложения ввиду высокой требовательности методов кластеризации к размерности векторов-примеров. К полученным векторам применена кластеризация методом K-Means с количеством кластеров, равным количеству положительных примеров. Центроиды этих кластеров считаются наиболее репрезентативными негативными примерами и используются при обучении классификатора. В архитектуре классификатора размерность charGRU равна 20, а размерность LSTM-кодировщика: 50. Предобученные векторы ELMo размерности 512 дополнительно сжимаются полносвязным слоем с 100 нейронами в целях экономии памяти. Гиперпараметры обучения: `dropout` = 0,1, `LR` = 0,001, размер батча: 2.

Классификатор отношения и ядерности. Ввиду низкой сбалансированности классов в классификаторах типа и ядерности отношения (CatBoost, BiMPM) функция потерь взвешивается с учётом весов классов,

обратно пропорциональным количеству обучающих примеров. В модели архитектуры BiMPM предобученные векторы токенов ELMo сжимаются до размерности 100. Размерность charGRU: 50, размерность LSTM-кодировщика: 200. При обучении использовались `dropout` = 0,4, скорость обучения (`LR`) = 0,0005 и размер батча: 8.

Модель разбора абзаца на основе глубокого обучения. В качестве модели векторизации токенов используется word2vec. Размерность charBiLSTM равна 50, размерность LSTM-кодировщика: 400, размер луча: 20. При обучении использовались следующие гиперпараметры: `dropout` = 0,3, `LR` = 0,002, размер батча: 40.

3.5.3. Оценка качества

В описываемых экспериментах задачи сегментации и структурной классификации рассматриваются как задачи бинарной классификации. Для оценки качества моделей сегментации вычисляется F1-мера для предсказания границы ЭДЕ, в то время как для оценки качества структурного классификатора — F1-мера для предсказания наличия отношения между дискурсивными единицами. Классификация отношения и ядерности оценивается с использованием макроусреднённой F1-меры.

Подробное описание метрик и критериев оценки методов построения риторических структур приведено в разделе 1.3.4.

3.5.4. Дискурсивная сегментация

В ходе экспериментов качество работы предлагаемого метода дискурсивной сегментации сравнивалось с качеством модели на основе BiLSTM без CRF-слоя [114], в которой в качестве признаков использовались эмбеддинги из мультиязычной языковой модели BERT. Как показано в таблице 8, добавление CRF-слоя, а также использование специфичной для языка контекстной языковой модели, позволяет улучшить качество сегментации на 0,86% F1 относительно базового варианта.

Таблица 8 — Результаты оценки качества сегментации, в %

Метод	Признаки	Валидац. выборка			Тестовая выборка		
		P	R	F1	P	R	F1
baseline	BERT-M	91,79	85,40	88,48	**89,66**	85,56	87,56
BiLSTM+CRF	BERT-M	90,75	**88,72**	89,72	87,80	**88,99**	88,39
	ELMo	**92,10**	87,53	**89,76**	89,09	87,86	**88,42**

Наилучшие результаты продемонстрировала модель BiLSTM-CRF, где для кодирования токенов применялась предобученная модель ELMo. На тестовом наборе данных она достигла значения F1, равного 88,4%. Таким образом, добавление CRF-слоя обеспечивает более точное разбиение текста на элементарные дискурсивные единицы, что важно для точности последующего построения риторического дерева.

3.5.5. Классификация

Результаты оценки качества структурной классификации представлены в таблице 9. Модель глубокого обучения (BiMPM с ELMo-эмбеддингами)

демонстрирует качество, сопоставимое с моделью на основе интерпретируемых признаков (baseline+AF) в терминах F-меры, однако превосходит её по показателю полноты (R). Важно отметить, что SVM-классификатор на основе явных признаков демонстрирует наилучшую точность (P). На тестовой выборке ансамбль обеих моделей с мета-классификатором обеспечивает F1 68,07%, что указывает на преимущество гибридного подхода, сочетающего интерпретируемые признаки и глубокое обучение.

Таблица 9 — Результаты оценки качества структурного классификатора, %

Метод	Признаки	Валидац. выборка			Тестовая выборка		
		P	R	F1	P	R	F1
baseline	baseline+AF	57,42	75,42	65,20	**58,42**	76,38	66,21
BiMPM	ELMo	54,95	**82,95**	66,11	54,54	82,82	65,77
baseline +BiMPM	baseline+AF, ELMo	**57,62**	82,56	**67,87**	57,66	**83,06**	**68,07**

Таблица 10 — Результаты оценки качества классификации отношения и ядерности, в %

Метод	Признаки	Валидац. выборка			Тестовая выборка		
		P	R	F1	P	R	F1
baseline	baseline	45,54	42,07	42,70	42,48	41,32	40,63
	baseline+AF	48,76	44,15	45,45	46,54	44,18	44,19
BiMPM	ELMo	50,85	46,80	46,76	47,35	45,40	44,64
baseline +BiMPM	baseline+AF, ELMo	**54,28**	**48,59**	**49,17**	**49,89**	**47,73**	**47,50**

В таблице 10 приведены сводные результаты оценки точности, полноты и F1 классификации риторического отношения и ядерности. Улучшенная базовая модель (baseline+AF) превосходит исходную базовую модель (baseline) по всем метрикам. Важно отметить, что модель глубокого

обучения (BiMPM с использованием ELMo) демонстрирует сопоставимое качество как на валидационной, так и на тестовой выборке, а их ансамблирование позволяет значительно повысить среднее качество классификации ввиду более точного распознавания нескольких классов относительно отдельных моделей: Condition, Evidence, Interpretation-evaluation, Purpose, Same-unit.

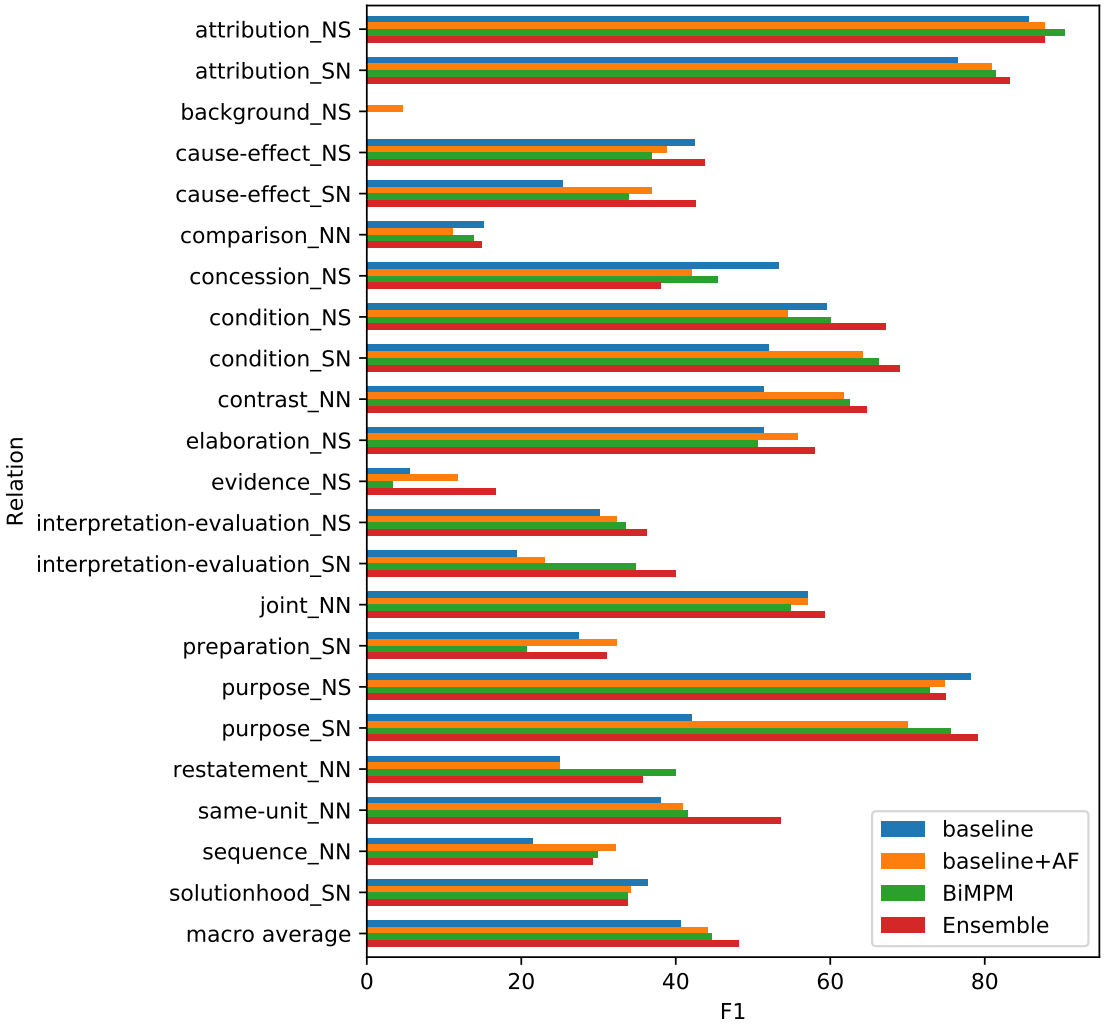


Рисунок 3.5 — Качество классификации отношения и ядерности по классам: F1, в %

Результаты оценки на тестовом корпусе базовой модели классификации риторических отношений (Baseline), улучшенной базовой модели (Baseline+AF), BiMPM-классификатора и ансамбля улучшенной базовой и нейронной моделей (Ensemble) представлены на рисунке 3.5. Нейронный классификатор показал наилучшие результаты в распознавании 12 из

22 риторических отношений. Однако базовая модель с большей точностью распознаёт наиболее частотные классы, такие как Joint (Конкатенация, 21,9% примеров) и Elaboration (Детализация, 20,4% примеров), что объясняется её способностью лучше обобщать типичные признаки массовых классов. Ансамбль базовой модели и BiMPM-классификатора позволил повысить макроусреднённую F1 на 2,86%. Это подтверждает эффективность предложенного метода для точного определения риторических связей между ДЕ.

3.5.6. Построение риторических деревьев

В таблице 11 приведены результаты оценки качества построения леса риторических структур из последовательности токенов текста при помощи двух методов разбора. Можно видеть, что базовый метод построения леса риторических структур позволяет восстановить структуры отношений, размеченных в корпусе, с качеством 27,1% F1 на уровне предложения, 20,3% на уровне абзаца и 17,7% на уровне документа. Эти результаты демонстрируют возможность построения леса риторических деревьев на основе локальных оценок вероятностей. Результаты также показывают значительное улучшение качества построения риторических деревьев при использовании нисходящего разбора с лучевым поиском на уровне абзацев. Прирост F1-метрики составил 10,5% на уровне предложений, 10,4% — на уровне абзацев и 8,9% — на уровне документа по сравнению с базовым методом.

Показатель качества поверхностного дискурсивного анализа с учётом отношений и ядерностей (Full) также улучшен: на уровне предложений прирост составил 10,6% F1, на уровне абзацев — 7,0%, а на уровне документа — 6,2%. Эти результаты показывают, что более тщательный разбор структуры составляющих в границах абзаца способствует более корректному анализу

Таблица 11 — Качество 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**

риторических структур в документе, а также подтверждают эффективность решения задач дискурсивного анализа методами глубокого обучения.

3.6. Выводы

В данной главе представлен комплексный подход к построению риторического леса для текста на русском языке с использованием методов поверхностного разбора дискурсивной структуры. В рамках разработки подходов к поверхностному разбору риторической структуры текста на русском языке предложены методы сегментации текста на элементарные дискурсивные единицы, а также методы классификации риторических отношений между ДЕ. В том числе разработаны признаки дискурсивных единиц для эффективной классификации типа отношения и положения ядра средствами классических методов машинного обучения, а также модели глубокого обучения на основе архитектуры BiMPM для классификации наличия отношения, типа и положения ядра.

Экспериментально показано, что сочетание классических методов машинного обучения с моделями глубокого обучения позволяет достичь вы-

сокого качества решения задачи классификации риторических отношений. Полученные результаты подтверждают эффективность предложенных методов, а также демонстрируют их потенциал для применения в различных задачах обработки текстов на русском языке.

В рамках поверхностного разбора риторической структуры текста предложено два подхода к построению риторических структур. Первый — жадный восходящий разбор с линейной свёрткой дискурсивных единиц в соответствии с локальной оценкой вероятности формирования ими дискурсивного отношения, позволяющий формировать лес риторических деревьев документа в соответствии со статистиками отношений в размеченном корпусе. Второй — нисходящий разбор дискурсивных структур с лучевым поиском, основанный на глубоком обучении, озволяющий наиболее эффективно формировать риторическую структуру в границах абзаца. Оба подхода показали свою эффективность для разбора риторических структур в текстах на русском языке. Метод нисходящего разбора, основанный на глубоком обучении с лучевым поиском, позволил значительно улучшить точность построения риторических деревьев на всех уровнях текста. Эти результаты подчеркивают важность использования современных методов глубокого обучения для дискурсивного анализа текстов на русском языке.

Результаты, полученные в главе 3, представлены в публикациях “RST discourse parser for Russian: an experimental study of deep learning models” [8] и “Discourse-aware text classification for argument mining” [9].

Глава 4. Методы полнотекстового разбора риторической структуры

4.1. Введение

В Теории риторических структур текст представляется в виде дерева составляющих, в котором дискурсивные единицы объединяются единым набором отношений на всех уровнях. Представление структуры текста как одного связного дерева необходимо для качественного решения прикладных задач, в которых важно эффективно оценивать высказывания в разных частях текста в соответствии с положением внутри общей структуры. Настоящая глава посвящена методам полнотекстового разбора риторической структуры текстов на естественном языке.

В этой главе предложен усовершенствованный метод DMRST [76] для полнотекстового нисходящего разбора риторической структуры текста. Этот метод основан на гибридной архитектуре глубокого обучения, объединяющей сегментацию текста на элементарные дискурсивные единицы, построение дерева риторических структур и назначение типов отношений и ядерностей в рамках одной модели. В главе подробно описаны модификации базовой архитектуры DMRST, а также приведена оценка их эффективности для анализа текстов на русском (корпус RRT [48]) и английском (корпус RST-DT [51], корпус GUM [77]) языках. Показано, что качество риторического анализа «с нуля» превосходит предыдущие подходы для английского языка.

С ростом интереса к кросс-языковому анализу естественных языков было обнаружено, что перенос моделей глубокого обучения для решения различных задач анализа между языками является перспективным направлением исследований, в особенности для тех языков, для которых количество размеченных данных ограничено. В области риторического

дискурсивного анализа разметка риторических структур разработана для нескольких языков (английский [51; 77], испанский [83], португальский [79–82], немецкий [86], нидерландский [87], баскский [78], китайский [84; 85; 128], русский [48]). При этом разметка в каждом корпусе отличается нюансами интерпретации Теории риторических структур, формальными ограничениями, способствующими повышению согласованности разметки, а также набором размеченных жанров и доменов. Всё это затрудняет исследование кросс-языковой обобщаемости современных методов риторического анализа. Как показывают исследования, проведённые на материалах нескольких десятков параллельных пар коротких текстов на дальнеродственных языках (англо-испано-баскские [93], испано-китайские [84; 129]), ключевыми факторами, определяющими различия в риторической структуре, являются особенности синтаксиса конкретного языка. Они диктуют различия прежде всего на уровне предложения, в то время как построение глобальной дискурсивной структуры в естественных языках более универсально. В данной главе описаны создание параллельного корпуса RRG для русского языка на основе мультижанрового англоязычного корпуса GUM [77] и разработка двуязычных моделей риторического анализа, выполненные в рамках диссертационного исследования.

Также актуальной становится задача кросс-жанрового анализа текстов, поскольку разнородность размеченных данных, присущая текстам разных жанров и стилей, усложняет задачу построения универсальных риторических анализаторов. В этой главе представлено решение этой проблемы через смешение данных различающихся интерпретаций теории риторических структур. Приведён анализ результатов исследования на материалах риторической разметки в двух крупных корпусах текстов на русском языке.

4.2. Метод гибридного нисходящего разбора

Методы полнотекстового анализа основаны на нейронной архитектуре DMRST [76], которая обеспечивает нисходящий риторический разбор неструктурированных текстов, включая сегментацию, построение деревьев, а также определение ядерности и типов отношений. Согласно оригинальному исследованию, модели на основе этой архитектуры показали высокие результаты для различных языков: английского (51,6% R F1), португальского (46,3% R F1), испанского (50,9% R F1), немецкого (32,3% R F1), нидерландского (39,4% R F1) и баскского (31,2% R F1).

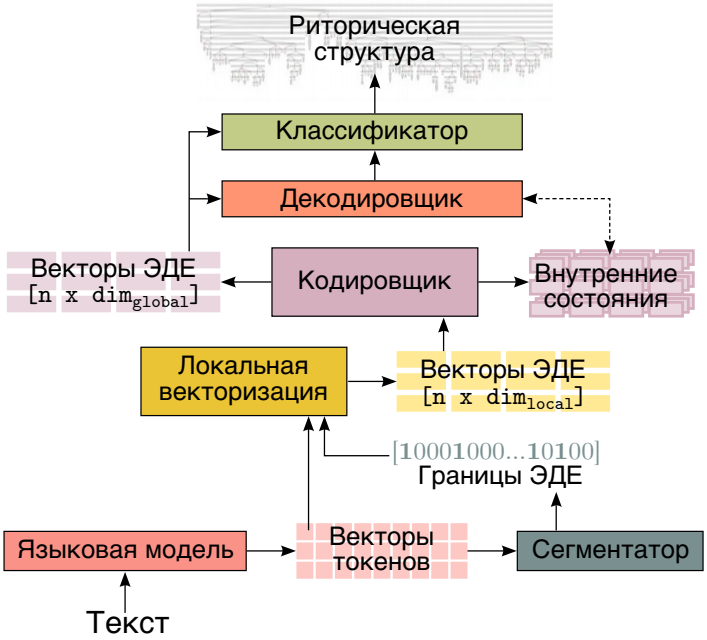


Рисунок 4.1 — Архитектура DMRST

Текст D представляется последовательностью закодированных токенов $D = \{t_1, t_2, \ldots, x_n\}$. В оригинальной работе, как и в этом исследовании, для кодирования текста используется предобученная контекстная языковая модель, а токенами являются частотные для корпуса предобучения языковой модели фрагменты слов [13]. Кодирующие языковые модели имеют ограниченный размер контекста (2^9, реже 2^{10} или 2^{11} токенов). Использование их в анализе развернутого дискурса, как и выбор мультиязычных

предобученных моделей, размер словаря которых предполагает разделение даже общеупотребимых слов в отдельных языках преимущественно на короткие токены, требует применения метода «скользящего окна». Последовательность токенов кодируется по k токенов за шаг, а отступы слева и справа позволяют учитывать близкий контекст при вычислении векторных представлений токенов фрагмента в фокусе. Последовательность закодированных токенов поступает на вход единой гибридной модели анализа риторических структур, схематично изображенной на рисунке 4.1 и состоящей из следующих модулей:

1. Сегментатор. Модуль сегментации предназначен для разбиения текста $T = \{t_1, t_2, \ldots, t_n\}$ на элементарные дискурсивные единицы. Для этого сегментатор определяет множество индексов $S = \{s_1, s_2, \ldots, s_k\}$, где каждый индекс s_i указывает на позицию последнего токена соответствующей ЭДЕ. Таким образом, текст представляется в виде сегментов

$$e_i = \{t_{s_{i-1}+1}, \ldots, t_{s_i}\}, \quad i \in \{1, 2, \ldots, k\}, \quad s_0 = 0. \tag{4.1}$$

   Предсказание оптимального разбиения текста осуществляется нахождением

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P(S \mid T), \tag{4.2}$$

   где $P(S \mid D)$ — вероятность варианта сегментации текста T на последовательность ЭДЕ, которая оценивается обученным на размеченных данных нейронным модулем.

   В оригинальной модели DMRST данный модуль реализован с помощью двух полносвязных слоёв, решающих задачу классификации отдельного токена документа. Первый слой оценивает вероятность нахождения токена на левой границе ЭДЕ ($y_i^{(1)} \in \{B, I\}$) и используется исключительно для расчёта общей функции потерь при обучении модели как регуляризатор весов основного выходного слоя. Такая стратегия обучения классификатора токена направлена

на учёт контекста одного соседнего токена[1]. Второй слой определяет, является ли токен правой границей ЭДЕ ($y_i^{(2)} \in \{B-1, I\}$), а его предсказания, нормализованные softmax, применяются для оценки вероятности сегментации $P(S \mid T)$.

2. Модуль локальной векторизации ЭДЕ. Для каждой ЭДЕ, заданной последовательностью векторов токенов $e_i = \{t_{s_{i-1}+1}, \ldots, t_{s_i}\}$, строится локальное векторное представление, которое зависит исключительно от токенов данной ЭДЕ и не учитывает глобальный контекст документа[2]. Обозначим через $x_j = \mathrm{emb}(t_j)$ векторное представление токена t_j, полученное с помощью модели-кодировщика. Тогда локальное векторное представление $\mathbf{v}_{\mathrm{local}}(e_i)$ ЭДЕ e_i строится как агрегированное представление векторов всех её токенов:

$$\mathbf{v}_{\mathrm{local}}(e_i) = f_{\mathrm{local}}\left(x_{s_{i-1}+1}, x_{s_{i-1}+2}, \ldots, x_{s_i}\right), \tag{4.3}$$

где f_{local} — функция агрегации (например, рекуррентный слой, взвешивание с механизмом внимания или другой метод сжатия последовательностей векторов). В оригинальной имплементации DMRST применяется простое усреднение векторов токенов.

3. Модуль глобальной векторизации ЭДЕ (кодировщик). Данный модуль формирует представления ЭДЕ с учётом контекста всего документа. На вход модуля подаётся последовательность локальных векторных представлений $\{\mathbf{v}_{\mathrm{local}}(e_i)\}_{i=1}^{m}$, кодирующих собственную информацию в каждом из m терминальных узлов риторической структуры. Она преобразуется в последовательность глобальных

[1] В предложенном далее модуле сегментации рассматривается задача разметки последовательности с учётом контекста всего документа. Для решения используется один обучаемый рекуррентный слой с CRF-оценкой переходов, ввиду чего отсутствует необходимость в дополнительных слоях.

[2] В то время как контекст документа может изначально учитываться при кодировании токенов в языковой модели, локальность представления здесь подразумевает, что анализатор использует для вычисления представления только закодированные любым методом векторы токенов внутри ЭДЕ.

векторов ЭДЕ, отражающих взаимосвязи между элементарными высказываниями в контексте всего дискурса:

$$\{\mathbf{v}_{\text{global}}(e_i)\}_{i=1}^{m} = f_{\text{global}}\big(\{\mathbf{v}_{\text{local}}(e_i)\}_{i=1}^{m}\big), \tag{4.4}$$

где f_{global} — функция преобразования, формирующая глобальные векторные представления на основе входной последовательности локальных векторов. В архитектуре DMRST для моделирования преобразования используется двунаправленный GRU-слой.

4. Рекуррентный модуль предсказания структуры (декодировщик).

Исходная дискурсивная единица $\text{du}_{1:m}$, которая включает все элементарные дискурсивные единицы документа, последовательно разбивается на две подструктуры на каждом шаге нисходящего разбора при помощи указательной сети [130]. Для организации рекурсивной декомпозиции реализуется стек, в который изначально помещается единица $\text{du}_{1:m}$.

а) Рассмотрим текущую дискурсивную единицу $e_{i:j}$, содержащую ЭДЕ с индексами от i до j. На данном шаге для каждого кандидата u в диапазоне $i \leqslant u \leqslant j$ вычисляется оценка внимания:

$$s_{t,u} = \sigma\big(h_t, \mathbf{v}_{\text{global}}(e_u)\big), \tag{4.5}$$

где h_t — текущее скрытое состояние декодировщика, отражающее информацию о уже выполненных шагах разбора, а $\sigma(x, y)$ — функция, реализующая операцию скалярного произведения (возможно, с дополнительным нелинейным преобразованием), которая служит для оценки совместимости текущего состояния с каждым кандидатом на разбиение. В качестве декодировщика используется однослойный GRU, обновление которого осуществляется согласно следующей

рекуррентной формуле:

$$h_t = \mathrm{GRU}\Big(f_{\mathrm{agg}}\big(\mathbf{v}_{\mathrm{global}}(e_i), \ldots, \mathbf{v}_{\mathrm{global}}(e_j)\big), h_{t-1}\Big), \tag{4.6}$$

где f_{agg} представляет собой функцию агрегации глобальных векторов ЭДЕ, входящих в текущую дискурсивную единицу $\mathrm{du}_{i:j}$ на шаге t (например, усреднение), а h_{t-1} — скрытое состояние, полученное на предыдущем шаге разбора. Этот подход обеспечивает сохранение информации о глобальном дискурсе и позволяет учитывать контекст уже выполненных разбиений при оценке разбиений локальных высказываний.

б) Оценки нормализуются с помощью softmax, определяется оптимальная точка разбиения:

$$k^* = \underset{u \in \{i,\ldots,j-1\}}{\mathrm{argmax}}\ \mathrm{softmax}(s_{t,u}). \tag{4.7}$$

Текущая дискурсивная единица $e_{i:j}$ делится на две подъединицы:

$$e_{i:j} \longrightarrow \big(e_{i:k^*},\ e_{k^*+1:j}\big). \tag{4.8}$$

Если какая-либо из полученных единиц не является элементарной (то есть содержит более одной ЭДЕ), она добавляется в стек для последующего разбора. Разбор продолжается до тех пор, пока все единицы не будут разбиты на элементарные.

5. Модуль классификации отвечает за предсказание риторического отношения и положения ядра между двумя дискурсивными единицами, представленными последовательностями глобальных векторов ЭДЕ. Пусть $\mathbf{v}(e_{i:k})$ и $\mathbf{v}(e_{k+i:j})$ — глобальные векторные представления левой e_L и правой e_R дискурсивных единиц, полученных на предыдущих этапах. Целью модуля является присвоение паре соединенных в риторической структуре ДЕ класса rel_nuc, где

$rel \in \mathcal{R}$ — риторическое отношение, а $nuc \in \{NS, SN, NN\}$ — положение ядра. Для моделирования взаимодействия между векторами составляющих дискурсивные единицы ЭДЕ используется биафинное преобразование. Пусть $\varphi(\mathbf{v}(e_L), \mathbf{v}(e_R))$ — биафинная функция, определяемая как:

$$\varphi(\mathbf{v}(e_L), \mathbf{v}(e_R)) = \mathbf{v}(e_L)^\top \mathbf{W}\mathbf{v}(e_R) + \mathbf{U}(\mathbf{v}(e_L) \oplus \mathbf{v}(e_R)) + \mathbf{b}, \qquad (4.9)$$

где $\mathbf{W}$ — матрица преобразования, $\mathbf{U}$ — обучаемая матрица для линейной комбинации векторов, $\mathbf{b}$ — вектор смещения. Результат биафинного преобразования $\varphi(\mathbf{v}(e_L), \mathbf{v}(e_R))$ передается в полносвязный слой, который предсказывает метку rel_nuc.

При создании полнотекстового риторического анализатора для русского языка и анализе кросс-языковой и кросс-жанровой обобщаемости риторического анализа исходная архитектура DMRST и детали обучения моделей были усовершенствованы для повышения точности сегментации и построения риторической структуры текста.

Одним из основных изменений стала реализация сегментатора на основе архитектуры BiLSTM-CRF [114] вместо полносвязной нейронной сети, использовавшейся в оригинальной версии DMRST. Это позволило существенно улучшить процесс сегментации текстов на ЭДЕ, что, в свою очередь, положительно повлияло на общую точность анализа. Следует отметить, что модуль сегментации аналогичной архитектуры применялся и в поверхностном риторическом анализаторе, описанном в главе 3. Однако при поверхностном анализе он функционировал отдельно от модуля анализа структуры и его веса обучались отдельно. Синхронная оптимизация весов сегментатора и модуля рекуррентного разбиения дискурсивных единиц, в свою очередь, позволяет более эффективно оценивать границы высказываний в соответствии с определениями риторических отношений в корпусе.

Другим ключевым изменением стало замещение простого усреднения подсловных эмбеддингов, применявшегося в оригинальной модели,

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 4.2 — Алгоритм гибридного нисходящего риторического разбора

дополнительным слоем BiLSTM для локального кодирования ЭДЕ. Это изменение позволило более точно учитывать контекст фраз и повысить качество представлений дискурсивных единиц.

Еще одна модификация касается механизма динамического взвешивания функций потерь (DWA [131]). Динамическое взвешивание необходимо для того, чтобы каждый модуль анализатора получал необходимое внимание во время обучения:

$$\mathcal{L}_{total} = \sum_{k=1}^{3} \lambda_k \mathcal{L}_k, \;\; w_k(i-1) = \frac{\mathcal{L}_k(i-1)}{\mathcal{L}_k(i-2)} \tag{4.10}$$

$$\lambda_k(i) = \text{softmax}(\frac{w_k(i-1)}{Temp}) \times 3, \tag{4.11}$$

где значение потерь $\mathcal{L}_{total}$ является взвешенной суммой потерь для отдельных модулей с весами λ_i; w_k — относительные скорости убывания функции для задач 1 (сегментация), 2 (построение дерева) и 3 (классификация), i — индекс итерации, а $Temp$ контролирует жесткость взвешивания. Однако в такой постановке динамическое взвешивание опирается только на значение потерь за последние два батча (формула 4.10). В риторическом анализе это означает риск усиления локальных трендов, особенно при небольшом размере батча, включающего риторические деревья разных глубин и сложности. Чтобы решить эту проблему, мы вводим параметр размера окна DWA b:

$$w_k(i-1) = \frac{\sum_{j=1}^{b} \mathcal{L}_k(i-j)}{\sum_{j=b+1}^{2b} \mathcal{L}_k(i-j)} \tag{4.12}$$

Анализируя более широкий диапазон значений потерь, модель может эффективно выявлять долгосрочные тренды и корректировать веса модулей соответствующим образом. Эта модификация улучшила стабильность обучения с меньшими батчами, особенно на наборе данных RRT, включающем как риторические структуры для развернутых абзацев текста, так и большое количество размеченных по отдельности предложений.

4.3. Кросс-языковой риторический анализ

4.3.1. Актуальность кросс-языкового анализа

Кросс-языковой дискурсивный анализ позволяет переносить знания и модели, обученные на одном языке, на другой язык, что значительно расширяет возможности анализа текста в условиях ограниченности ресурсов. Такой подход предоставляет ряд преимуществ, включая повышение качества анализа за счёт использования дискурсивной информации на различных языках и возможность обучения на гораздо более обширных корпусах.

Следует отметить, что ранее кросс-языковая обобщаемость методов дискурсивного анализа изучалась на параллельном корпусе локальных дискурсивных отношений TED-MDB (TED Multilingual Discourse Bank [95]). Однако этот корпус включает разметку лишь шести текстов на шести языках (английский, русский, польский, португальский, немецкий, турецкий), содержащих от 560 до 661 примера дискурсивных отношений для каждого языка. Он может быть использован только для оценки возможностей кросс-языкового переноса в подзадачах выделения дискурсивных коннекторов и классификации отношений, но не для обучения методов построения целостной дискурсивной структуры. Как показано в обзоре корпусов риторической разметки (раздел 1.3.3), для обучения моделей построения дискурсивной структуры текстов методами глубокого обучения используют крупные корпусы, насчитывающие не менее 100 риторических структур или 2600 примеров дискурсивных отношений. Существующие корпусы риторической разметки для русского (RRT) и английского (RST-DT, GUM) языков, используемые в разработке автоматических анализаторов, содержат от 21494 до 26106 примеров дискурсивных отношений.

Несмотря на значительный интерес к данной проблеме, кросс-языковой анализ риторической структуры остаётся сложной задачей вследствие разнообразия интерпретаций теории риторической структуры при разметке корпусов на разных языках и недостатка параллельных дискурсивных данных. Существующие крупные корпусы значительно различаются методами разметки, набором жанров, источниками текстов и определениями типов отношений, что может приводить к недооценке потенциала кросс-языкового переноса в риторическом анализе. Таким образом, для исследования возможностей кросс-языковой адаптации методов дискурсивного анализа необходимо создание обширного параллельного корпуса риторической разметки.

В рамках диссертационного исследования разработана русскоязычная версия риторического раздела многослойного мультижанрового корпуса Джорджтаунского университета (GUM) версии 9.1 [77]. Параллельная версия корпуса охватывает все 213 оригинальных документов 12 жанров. Этот первый большой параллельный корпус полнотекстовой риторической разметки является ценным ресурсом для кросс-языкового анализа дискурса, позволяющим разрабатывать двуязычные риторические модели и оценивать потенциал кросс-языкового переноса методов дискурсивного разбора между дальнеродственными языками.

Как показывают предыдущие исследования [84; 93; 129; 132], различия в риторических структурах между языками в основном выражаются на низших структурных уровнях, тогда как глобальная организация документа имеет более универсальный характер. В рамках настоящего диссертационного исследования разработан риторический анализатор (см. раздел 4.2), основанный на нисходящем разборе. Анализаторы данного типа начинают построение риторического дерева с определения универсальной высокоуровневой структуры документа, которая является относительно универсальной, тогда как языковые особенности определяются прежде всего синтаксической структурой на низших уровнях дискурса.

Исследование, описанное в этом разделе, посвящено изучению возможностей адаптации гибридного нисходящего анализатора к различным жанрам на втором языке с использованием разработанных методов риторического анализа и русскоязычной версии корпуса полнотекстовой риторической разметки.

4.3.2. Методология создания параллельного корпуса

Корпус русскоязычной разметки на основе многослойного корпуса Джорджтаунского университета (RRG) создан вручную адаптацией разметки риторических структур из корпуса $\text{GUM}_{9.1}$. Далее описаны ключевые этапы создания параллельного корпуса RRG.

Перевод

В рамках исследования особое внимание уделялось ручному переводу с английского языка 213 текстов 12 жанров: 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`). Такой подход позволил обеспечить как литературность перевода, так и адаптацию к жанровым особенностям письменного дискурса на русском языке. Он отличается от стандартной практики исследований в области межъязыкового анализа риторических структур, в которых используется машинный перевод отдельных элементарных дискурсивных единиц с сохранением структуры дерева из исходных данных. При подготовке версии для русского языка особенное

внимание уделялось точности перевода терминологии и имён сущностей, что потребовало углублённого исследования источников. Перевод на русский язык также требовал установления рода глаголов, что было выполнено для частей корпуса, описывающих устный дискурс (`vlog` и `conversation`) на основе анализа аудиозаписей.

Сопоставление риторических структур

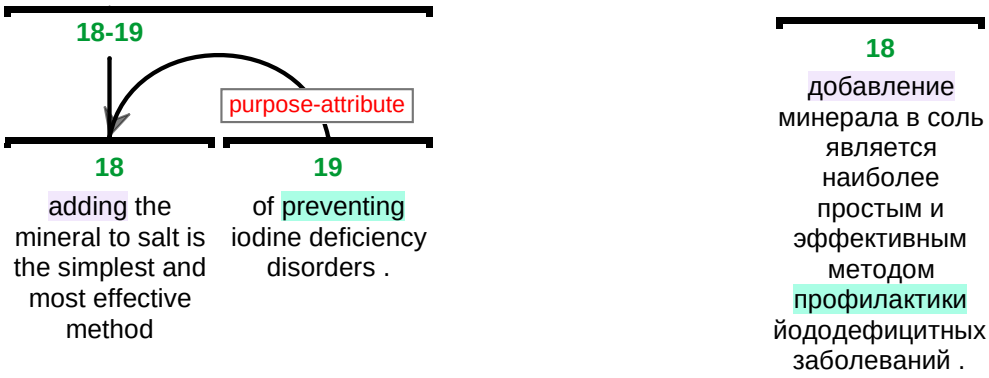
Переведённые тексты были вручную сопоставлены с оригинальными риторическими структурами на уровне элементарных дискурсивных единиц. Сопоставление осуществлялось в соответствии с инструкциями по дискурсивной сегментации на русском языке, разработанными для разметки корпуса риторических структур RRT [113]. При согласовании риторических деревьев допускалось добавление или удаление дискурсивных единиц в связи с различиями в сегментации текстов на двух языках. Ядерность и риторические отношения назначались согласно инструкции по разметке корпуса GUM RST[3]. Поскольку методология исследования предполагала уточнение риторических отношений на уровне около предложений, а не построение структуры «с нуля», ожидались минимальные расхождения в разметке риторической структуры. Как показано в таблице 12, корпус RRG содержит 95,9% от числа ЭДЕ в корпусе для английского языка. Анализ выявил преобладание соответствий «N-к-одному» среди ЭДЕ, что главным образом объясняется различиями между языками:

– Перевод глагольных форм в существительные в русском языке. Например, `adding` переводится как *добавление*, а `preventing` — как *профилактика* (см. пример на рисунке 4.3).

[3] https://wiki.gucorpling.org/gum/guidelines

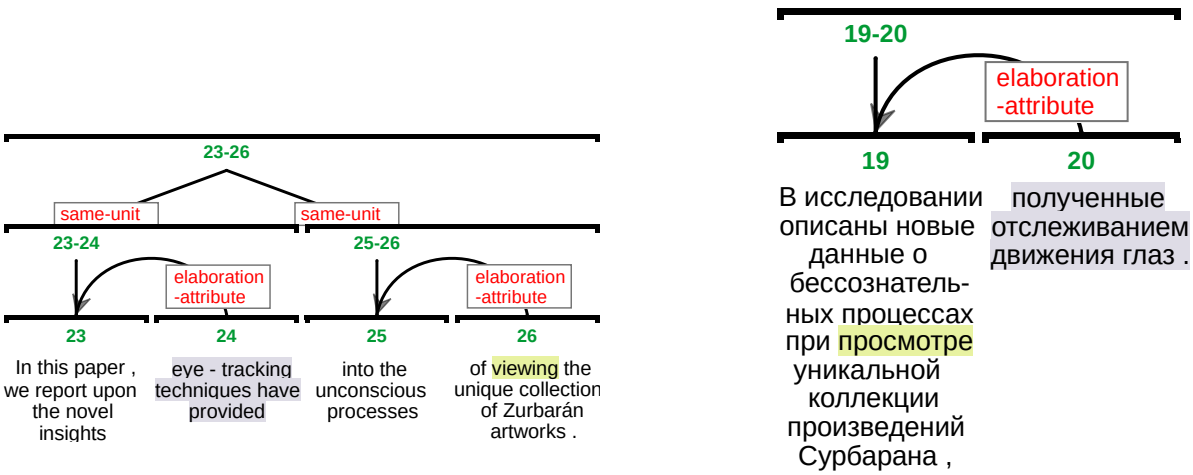
– В русском языке предпочтение чаще отдаётся активным конструкциям, что сокращает количество выделяемых в ЭДЕ страдательных оборотов (например, [*there's sufficient iodine*] [*added into the food supply*] можно перевести как [*в пищевые продукты поступает достаточное количество йода*].

– Некоторые ЭДЕ в RRG соответствуют сокращённым подструктурам с отношением служебного типа SAME-UNIT из GUM (см. рисунок 4.4), подразумевающим продиктованный ограничением структуры составляющих разрыв единой дискурсивной единицы.



(а) Разметка в GUM. (б) Соответствующая разметка в RRG.

Рисунок 4.3 — Соответствие ЭДЕ N-к-одному; `news_iodine`.

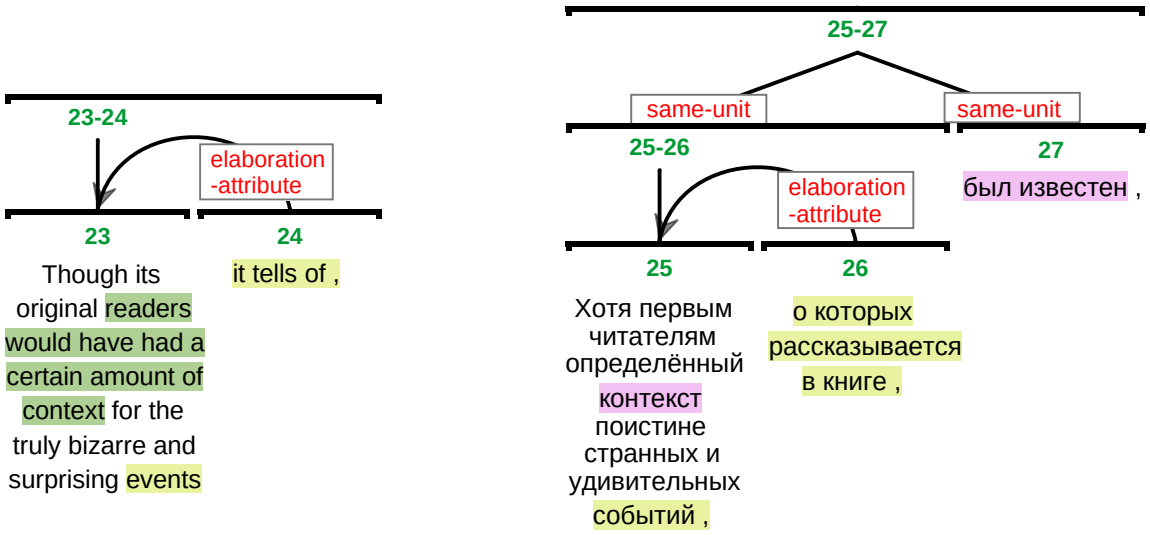


(а) Разметка в GUM. (б) Соответствующая разметка в RRG.

Рисунок 4.4 — Соответствие ЭДЕ N-к-одному; `fiction_wedding`.

Соответствия «один-к-N» возникают, когда одна ЭДЕ в GUM делится на несколько единиц в RRG. Чаще всего это наблюдается при переводе

предложных фраз и при наличии большей вариативности синтаксической конструкции в некотором жанре письменного дискурса (см. рисунок 4.5).



(а) Разметка в GUM. (б) Соответствующая разметка в RRG.

Рисунок 4.5 — Соответствие ЭДЕ один-к-N; `fiction_wedding`.

Шлифовка разметки

Для повышения согласованности разметки и устранения ошибок в черновой версии корпуса RRG был проведён анализ распределения классов риторических отношений. Этот анализ позволил выявить явные ошибки разметки, в том числе унаследованные из оригинального английского корпуса. К таким относятся редкие или маловероятные классы (например, присутствовавший в одном из 213 документов класс Restatement_SN: перефразировка не может встречаться в тексте ранее перефразируемого фрагмента текста).

Для дальнейшего уточнения разметки использовался классификатор типов риторических отношений для русского языка (см. раздел 3.3), обученный на черновой версии корпуса. Этот классификатор служил инструментом для обнаружения выбросов, помогая идентифицировать по-

тенциальные ошибки разметки. Особое внимание уделялось случаям, когда классификатор с высокой уверенностью (низкой энтропией) предсказывал некорректный класс, исключая истинную (размеченную) метку из трёх наиболее вероятных вариантов. Обнаруженные ошибки в структуре и назначении классов были исправлены в соответствии с инструкциями по разметке риторических отношений корпуса GUM.

4.4. Кросс-жанровый риторический анализ на основе смешанной разметки

4.4.1. Актуальность смешения данных в риторическом анализе

Одной из главных проблем автоматического риторического анализа с помощью машинного обучения является ограниченность размеченных данных, обусловленная высокой трудоёмкостью разметки. Для решения задач анализа локального дискурса, таких как сегментация предложений и классификация дискурсивных отношений, в корпусах предлагается достаточно данных для глубокого обучения. Однако задача оптимального разбора структуры полного текста, являющаяся краеугольной в дискурсивном анализе, решается на основе в лучшем случае нескольких сотен размеченных примеров. При этом обучению анализаторов на смешении множества корпусов препятствует разнообразие наборов отношений в разных корпусах и стратегий объединения отношений в классы верхнего уровня. Ранее предпринимались попытки такого смешения [76;91], однако они требуют гармонизации отношений в корпусах разных языков и жанров. Такое согласование просто в реализации, но приводит к потере информации о точных риторических отношениях и расположении ядер из экспертной разметки; кроме того, оно не учитывает особенности ограничений на разметку, ко-

торые могут противоречить в разных корпусах в зависимости от наборов отношений или уточнений определений схожих отношений. В этом разделе предлагается метод обучения риторического анализатора на смешении любых данных риторической разметки, независимо от набора отношений, принятого в конкретном корпусе. Метод оценивается на смешении двух больших корпусов на русском языке: RRT и RRG, а также приводятся оценки качества обучения со взвешенным добавлением автоматически переведённых англоязычных корпусов. Показано, что метод позволяет (1) эффективно увеличить объём данных для обучения анализатора без искажения разметки риторических структур и (2) получить универсальный дискурсивный анализатор, свободный от недостатков моделей, обучаемых на отдельных корпусах с характерными для них жанрами, источниками и глубиной структур.

4.4.2. Метод смешения данных в риторическом анализе

Несмотря на то, что в каждом корпусе TPC-разметки используются разные наборы и определения отношений, в связи с чем они также различаются деталями сегментации, общие принципы построения структур из заданного набора элементарных дискурсивных единиц более универсальны. В связи с этим предлагается архитектура анализатора, где модуль построения структуры универсален для любых обучающих данных риторической разметки и может быть обучен на смешении данных из N корпусов (рисунок 4.6).

В предлагаемой архитектуре добавлены отдельные модули сегментации и классификации для каждого корпуса из обучающего набора (N корпусов). На этапе обучения модели входное значение $i \in 0..N$ соответствует порядковому номеру корпуса в обучающем наборе. Параметры

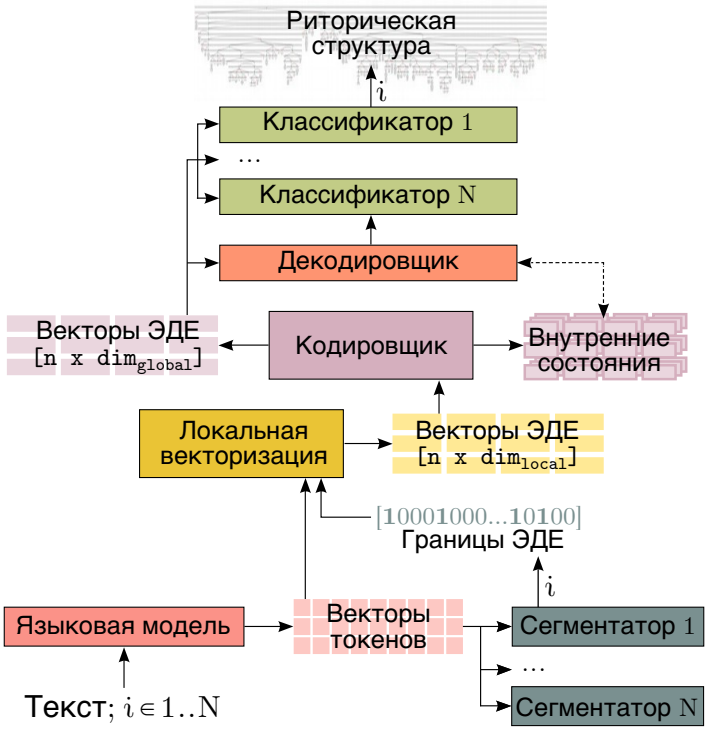


Рисунок 4.6 — Схема риторического анализатора, универсального для N наборов обучающих данных.

кодирующей языковой модели и модуля построения дерева составляющих обновляются при обработке каждого обучающего примера, что способствует обучению на большем количестве данных и, соответственно, более тщательному выявлению закономерностей. Параметры языковой модели оптимизируются с использованием алгоритма обратного распространения ошибки для улучшения кодирования текстов на основе общих для языка принципов дискурсивной сегментации и классификации. Использование отдельных модулей для сегментации и классификации каждого корпуса минимизирует риск искажения экспертной разметки при обучении модели. Кроме того, важным преимуществом предлагаемого подхода является возможность динамического выбора набора риторических отношений на этапе разбора текста обученной моделью: пользователь самостоятельно задаёт значение i, что позволяет адаптировать детализацию отдельных аспектов дискурса в зависимости от прикладной задачи.

4.5. Экспериментальное исследование методов полнотекстового нисходящего разбора риторической структуры

4.5.1. Описание данных

В этом разделе качество полнотекстового разбора и возможности обобщения риторического анализа оцениваются на материале риторической разметки для двух языков: русского и английского. Используется четыре корпуса риторической разметки:

RRT [48] В исследованиях по анализу полнотекстового риторического анализа в русском языке используется корпус $\text{RRT}_{\text{v2.1}}$, являющийся улучшенной версией разметки версии 2.0. Исправлены все обнаруженные непроективные структуры, несвязанные ЭДЕ внутри поддеревьев, проблемы с xml-файлами разметки; выверены все метки начала абзаца, иллюстраций, иллюстраций с подписью; исправлены формальные ошибки сегментации с разделением внутри слов и до пунктуации; улучшена разметка предложений как самостоятельных поддеревьев в блоговой части корпуса; согласованы формальные дискурсивные связи во всем корпусе (заголовок–текст, заголовок–подзаголовок–текст, списки, иллюстрация–подпись, текст–выводы, цитата–атрибуция–цитата); исправлены очевидные ошибки назначения классов основных отношений по аналогии со «шлифовкой разметки», описанной в разделе 4.3.2. Полученный корпус включает новостные статьи и блоги из различных источников: 5 новостных источников и 17 источников блогов, посвященных путешествиям, историям из жизни, информационным технологиям, косметике, ЗОЖ, политике, экологии и психологии. Положительной стороной корпуса, отличающей его от других корпусов риторической разметки, является разнообразие ис-

точников дискурса. Важной отрицательной стороной корпуса является поверхностная разметка документов: число риторических структур, размеченных в одном документе, достигает 42, в среднем корпус содержит 11,7 деревьев на документ. В связи с этим были разработаны методы поверхностного риторического анализа, описанные в главе 3. Однако целью риторического анализа является в первую очередь построение единой полнотекстовой структуры документа, и прочие известные корпуса риторической разметки разработаны с этим условием. Поскольку деревья в корпусе RRT включают случайное количество абзацев, а риторические классы высокоуровневой организации текста, такие как Заголовок или Смена-темы в нем отсутствуют, эксперименты с полнотекстовым разбором на основе этой разметки мы проводим, разделяя каждый размеченный документ на неэлементарные несвязанные поддеревья и полагая каждое дерево отдельным размеченным документом. При этом сохраняется оригинальное разделение корпуса на обучающую/проверочную/тестовую выборки; документы просто разделяются на файлы `docname_part_*_.rs3`, рассматриваемые как независимые тексты. Документы, содержащие только одну ЭДЕ (обычно это заголовки текстов), исключаются. Важно заметить, что в полученном корпусе, используемом в экспериментах, 12.8% деревьев состоят из 2-4 элементарных дискурсных единиц. Это высокое значение, в первую очередь диктующее введение описанной в разделе 4.2 модификации вычисления весов функции потерь: поведение значений функции потерь, когда в соседних батчах находятся риторические структуры из двух и двадцати ЭДЕ, необходимо стабилизировать перед адаптацией скорости обучения отдельных модулей.

RST-DT [51]. Эталонный размеченный корпус для разработки и оценки систем риторического анализа. Включает разметку 385 текстов новостей WSJ различной длины.

GUM [77] Открытый корпус многоуровневой лингвистической разметки, формируемый студентами курса компьютерной лингвистики. Содержит разметку множества жанров письменного и устного, дискурсивную разметку в рамках теории риторических структур. Версия 9.1 содержит 213 документов 12 жанров.

RRG Версия корпуса $\text{GUM}_{9.1}$ на русском языке, разметка которой описана в разделе 4.3.2.

Общий анализ четырех корпусов, представленный в Таблице 12, демонстрирует различия между корпусами, выходящие за рамки различия в наборах жанров, размеров деревьев и набора риторических отношений. Например, в корпусе RST-DT обнаружено 79,4% формирующих самостоятельное поддерево неэлементарных предложений[4] (тех, в которых размечено хотя бы одно дискурсивное отношение). Этот показатель, так же как и явно присутствующая разметка границ предложений и абзацев, способствовал развитию исследований риторического анализа внутри предложения [123; 133–136]. В корпусе GUM соответствие между формальными границами предложений и риторическими поддеревьями менее жесткое. Кроме того, в разметке GUM RST, используемой для обучения и оценки анализаторов, полностью отсутствуют метки предложений и абзацев, в отличие от RST-DT. Эти различия подчеркивают, что вариации в риторической структуре, даже в пределах одного и того же жанра (см. таблицу 13) возникают не только из-за разнообразия наборов отношений и текстовых источников в разных корпусах, как предполагают [137], но также из-за различий в формальных ограничениях, устанавливаемых для улучшения согласованности между разметчиками.

[4]Предсказание границ предложений осуществлялось библиотеками `spaCy` (английский) и `razdel` (русский). Этот подход минимизирует влияние потенциальных ошибок от автоматических сегментаторов предложений на сравнение корпусов.

Таблица 12 — Статистики корпусов

Корпус	Жанров	Источ.	Док.	Классов	Токенов на дерево			Неэлем. предлож., %	ЭДЕ	ЭДЕ на дерево	Пар ДЕ
					мин.	макс.	мед.				
RST-DT	1	1	385	41	30	2624	396	79,4	21789	56,6	21404
GUM	12	12+	213	27	167	1879	989	72,5	26319	123,6	26106
RRT	2	17+	233	24	2	1148	89	76,7	28372	11,7	25957
RRG	12	12+	213	27	137	1629	833	76,9	25223	118,4	25010

Хотя авторы корпуса RST-DT [123] кратко упоминают о том, что 95% предложений размечены самостоятельными поддеревьями, наш анализ, основанный на автоматической сегментации предложений и подсчете в бинаризованных деревьях (стандартный формат для автоматического риторического анализа), обнаружил более умеренную оценку в 86% для всех типов предложений.

4.5.2. Метод гибридного разбора

Детали экспериментов

В этом исследовании используется мультиязыковая модель `xlm-roberta-large` [138], известная исключительным качеством кодирования текста для решения задач классификации без дообучения на размеченных данных [139]. Гиперпараметры, заданные для отдельных модулей DMRST, указаны в таблице 14. В каждом эксперименте результаты усредняются по пяти запускам с различными начальными значениями модели (фиксированные разбиения корпусов: GUM и RRG, RRT) или разными разбиениями на обучающую/валидационную выборки (RST-DT).

Таблица 13 — Неэлементарные предложения, покрываемые самостоятельными подструктурами, в %

Корпус	Жанр	Англ.	Рус.
RST-DT	*news*	79,4	–
GUM, RRG	*academic*	72,0	76,9
	bio	61,1	72,2
	conversation	65,8	68,7
	fiction	70,4	78,5
	interview	71,4	78,1
	news	69,0	79,2
	reddit	73,0	77,4
	speech	85,8	87,5
	textbook	78,5	76,4
	vlog	75,3	77,3
	voyage	71,3	71,5
	whow	77,5	78,4
RRT	*blogs*	–	71,6
	news	–	82,9

Сегментация

Оценка качества сегментации при анализе текстов гибридным методом представлена в Таблице 16, как и другие метрики риторического анализа текстов с нуля.

Английский язык Из предложенных ранее методов сегментации лучшее качество на корпусе RST-DT принадлежит методу DisCut [140]. Этот классификатор последовательности токенов в предложении на основе языковой модели `XLM-RoBERTa-large` достиг качества 97,6% F1[5]. Предложенный на-

[5]Согласованность между аннотаторами для сегментации на подмножестве из 53 [51] дважды размеченных текстов корпуса RST-DT имеет оценку 98,3% F1 [123]. Однако эта оценка относится к ограниченной небольшой части корпуса, не совпадающей с его официальной тестовой частью.

Таблица 14 — Используемые параметры

	RST-DT	**GUM**	**RRG**	**RRT**
batch size (кол-во деревьев)	2	1	1	6
b_{DWA} (кол-во деревьев)	12	12	12	24
`Языковая модель`				
размерность внутренних представлений		1024		
размер фокуса скользящего окна		400		
скорость обучения		2e-05		
`Анализатор`				
размерность внутренних представлений	1024	1024	1024	768
dropout (вход сегментатора)		0.4		
dropout (вход кодировщика)		0.5		
скорость обучения		1e-04		
`ToNy`				
размерность внутренних представлений		200		
`E-BiLSTM`				
размерность внутренних представлений		512		

ми анализатор DMRST+`ToNy` позволяет улучшить качество сегментации, демонстрируя среднее значение по пяти запускам в 97.9% F1. Итоговая модель также превосходит модели базовой архитектуры DMRST на корпусе $\text{GUM}_{9.1}$, достигая среднего значения F1 95,5% (+ 0,8% F1 в среднем).

Русский язык В предыдущих исследованиях для сегментации использовался отдельный модуль на основе архитектуры ToNy [114] (см. раздел 3.2). При обучении и оценке на корпусе $\text{RRT}_{\text{v2.1}}$ этот метод демонстрирует качество сегментации 89,1% F1. Гибридный анализ методом DMRST позволил достичь среднего качества сегментации 92,4% F1. Модификации архитектуры DMRST, описанные в разделе 4.2, не оказали значительного влияния на качество сегментации на RRT ввиду необходимости в большем размере батча (таблица 14) для стабильного обучения на корпусе. В то же время модификации последовательно улучшали качество сегментации на корпусе RRG, в среднем с 96,3% F1 до 96,9% F1.

Построение структуры

Результаты экспериментов выявили важное противоречие между двумя задачами, решаемыми одной гибридной моделью. Модели, демонстрирующие наилучшее качество построения риторического дерева из последовательности «стандартных» (размеченных экспертами) ЭДЕ, показывают худшие результаты при разборе текста с нуля (на предсказанной сегментации), и наоборот (см. таблицы 15 и 16). Эта закономерность свидетельствует о том, что внутренние представления кодировщика во время обучения модели оптимизируются для решения двух потенциально конфликтующих задач. Сегментация требует анализа локальных дискурсивных признаков внутри предложения. Модели и нейронные модули, решающие задачу сегментации, оптимизируют в первую очередь представления, соответствующие синтаксическим закономерностям. В то же время построение структуры на уровне документа требует учёта более далёкого контекста, что включает в себя оптимизацию представлений, отражающих закономерности в глобальном дискурсе. Кроме того, глубокие модели разбора структуры, обученные и тестируемые на экспертной сегментации, не способны корректировать характерные ошибки автоматического сегментатора. Таким образом, прямое сравнение моделей, требующих отдельного модуля сегментации, и гибридных моделей на стандартном наборе ЭДЕ может быть недостоверно. Далее в этой главе описываются результаты оценки разбора текстов с нуля (end-to-end parsing) как целевой задачи риторического анализа.

Результаты оценки качества риторического анализа с нуля представлены в таблице 16.

Модифицированный метод DMRST продемонстрировал лучшее качество по сравнению с оригинальной версией на материалах корпуса RST-DT, достигнув F1-метрики 55,3% для полной риторической структуры. На те-

Таблица 15 — Оценка построения риторического дерева из набора стандартных ЭДЕ, F1, в %. Отсутствующие значения не приведены в цитируемых работах.

	Корпус	Метод	Segm	N	R	Full
Англ.	RST-DT	Human	78.7	66.8	57.1	55.0
		HILDA [60]	68.6	55.9	45.8	44.6
		DPLP [63]	64.1	54.2	46.8	46.3
		CODRA [62]	65.1	55.5	45.1	44.3
		[141]	65.3	54.2	45.1	44.2
		[66]	64.5	54.0	38.1	36.6
		HILDA [142]	65.1	54.6	44.7	44.1
		[61]	59.5	47.2	34.7	34.3
		[94]	62.7	54.5	45.5	45.1
		[68]	71.4	60.3	49.2	48.1
		[143]	67.1	57.4	45.5	45.0
		[71]	67.2	55.5	45.3	44.3
		[75]	74.3	64.3	51.6	50.2
		[73]	73.1	62.3	51.5	50.3
		[74]	76.3	65.5	55.6	53.8
		DMRST + Маш. перевод [76]	76.7	66.2	56.5	–
		[144]	76.4	66.1	54.5	53.5
		[56]	77.8 ± 0.3	**68.0 ± 0.5**	**57.3 ± 0.2**	55.4 ± 0.4
		DMRST (this work)	**78.7 ± 0.4**	**68.0 ± 0.6**	**57.3 ± 0.2**	**55.7 ± 0.3**
		+ `ToNy`	78.4 ± 0.7	67.4 ± 0.8	56.8 ± 0.9	55.2 ± 0.9
		+ `ToNy` + `E-BiLSTM`	78.5 ± 0.5	67.5 ± 0.7	57.0 ± 0.5	55.3 ± 0.5
	GUM v9.1	DMRST (this work)	72.7 ± 0.7	60.8 ± 0.6	52.8 ± 0.5	51.7 ± 0.4
		+ `ToNy`	72.8 ± 0.3	**61.4 ± 0.6**	**53.1 ± 0.5**	**52.0 ± 0.5**
		+ `ToNy` + `E-BiLSTM`	**73.1 ± 0.3**	61.3 ± 0.2	53.0 ± 0.3	**52.0 ± 0.3**
Рус.	RRT	DMRST (this work)	81.0 ± 0.5	63.3 ± 0.9	54.2 ± 0.9	54.0 ± 0.9
		+ `ToNy`	80.9 ± 1.0	**63.4 ± 0.9**	**54.7 ± 0.9**	**54.6 ± 0.9**
		+ `ToNy` + `E-BiLSTM`	**81.2 ± 0.4**	62.9 ± 0.9	53.8 ± 1.2	53.6 ± 1.2
	RRG	DMRST (this work)	**71.5 ± 0.4**	**57.6 ± 0.2**	**49.1 ± 0.3**	**47.9 ± 0.2**
		+ `ToNy`	71.1 ± 0.5	56.6 ± 1.4	48.2 ± 1.5	47.2 ± 1.4
		+ `ToNy` + `E-BiLSTM`	70.7 ± 0.4	56.4 ± 0.5	48.3 ± 0.5	47.1 ± 0.5

стовой части корпуса GUM улучшенная модель также продемонстрировала лучшее качество, достигнув 52,0% F1 для полной риторической структуры. Полученные результаты демонстрируют эффективность предложенных модификаций в условиях многожанрового корпуса. Несмотря на высокое качество на корпусах для английского языка, на корпусе RRT модифицированная модель не позволила улучшить качество риторического анализа,

Таблица 16 — Качество риторического анализа, F1, в %

	Корпус	Метод	Segm.	S	N	R	Full
Англ.	RST-DT	SegBot [116] & [71]	92,2	62,3	50,1	40,7	39,6
		RPFS [75]	96,3	68,4	59,1	47,8	46,6
		DMRST [76]	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**
Рус.	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**

что может объясняться необходимостью в большом размере батчей для стабильного обучения на корпусе с сильно различающимися по сложности примерами (см. параметры обучения моделей в таблице 14). Среднее значение метрики Full F1 составило 43,8%, что указывает на сложности в адаптации модели к специфике русского языка и важность качественной сегментации для всего риторического анализа.

4.5.3. Двуязычные модели

Для оценки качества созданного параллельного корпуса и возможности переноса моделей риторического анализа между языками на основе модифицированной архитектуры DMRST, описанной в разделе 4.2, были разработаны двуязычные модели. Исследуются возможности архитектуры `+ToNy+E-BiLSTM`, продемонстрировавшей лучшее качество анализа на тесто-

вых примерах параллельного корпуса для каждого языка. Рассматривается два сценария: (1) адаптация в условиях ограниченных параллельных данных и (2) межъязыковая адаптация анализатора на полностью размеченных параллельных данных. Анализируется качество риторического разбора на втором языке. Оценивается, насколько увеличение параллельной разметки для обучения влияет на качество риторического разбора различных жанров на русском языке с фиксированным набором данных для английского языка.

Непосредственный перенос

В таблице 17 указано качество анализа для каждого языка на тестовой выборке соответствующего варианта корпуса по жанрам.

Используя в экспериментах тексты, отличающиеся только языком, мы изолируем влияние на анализ риторической структуры дискурсивных языковых особенностей, предполагая более объективную оценку межъязыковых возможностей анализа по сравнению с типичными подходами, использующими смешанные источники. Как показано в таблице 18, анализатор риторических структур, обученный только на данных разметки для английского языка, достигает впечатляющих результатов на материалах русскоязычного письменного дискурса при отсутствии обучающих данных на русском языке только за счет использования предобученной мультиязычной языковой модели. Его качество сравнимо с моноязыковым русскоязычным анализатором структуры. Хотя анализатор для русского языка демонстрирует лучшее качество по всем метрикам (Seg: +1,4%, Span: +2,6%, Nuc: +1,9%, Full: +2,4%), разрыв остается относительно небольшим, что демонстрирует эффективность оригинального парсера на основе мультиязыковой модели при межъязыковом переносе. Обратное направле-

Таблица 17 — Оценка моноязыковых анализаторов по жанрам, в %

Английский язык					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	94,6 ± 0,6	72,7 ± 1,3	64,0 ± 1,9	56,9 ± 1,7	56,3 ± 1,7
bio	97,7 ± 0,6	68,1 ± 1,8	57,0 ± 2,9	53,2 ± 2,1	51,5 ± 2,1
conversation	95,5 ± 0,3	49,5 ± 1,3	39,0 ± 1,5	29,8 ± 1,4	29,3 ± 1,6
fiction	93,9 ± 0,7	59,3 ± 2,4	47,8 ± 2,9	39,7 ± 2,2	38,5 ± 2,3
interview	95,1 ± 0,4	73,8 ± 0,6	65,7 ± 1,3	55,3 ± 1,2	55,1 ± 1,1
news	94,6 ± 0,8	69,0 ± 1,9	60,4 ± 2,4	56,7 ± 2,0	55,0 ± 2,1
reddit	93,3 ± 0,6	60,5 ± 1,1	51,5 ± 1,4	44,5 ± 1,6	44,0 ± 1,4
speech	97,5 ± 0,4	79,1 ± 1,7	67,4 ± 2,4	57,8 ± 1,8	57,6 ± 2,0
textbook	97,5 ± 0,3	78,7 ± 1,3	66,1 ± 1,8	57,4 ± 2,0	57,0 ± 1,9
vlog	95,6 ± 0,5	61,9 ± 1,0	48,8 ± 2,0	43,5 ± 1,5	41,7 ± 1,7
voyage	94,6 ± 0,5	67,2 ± 1,9	51,6 ± 2,2	44,6 ± 2,0	44,1 ± 1,9
whow	97,3 ± 0,3	75,7 ± 0,9	64,3 ± 1,9	58,6 ± 1,7	57,0 ± 1,7
все	95,5 ± 0,1	66,9 ± 0,5	56,1 ± 0,3	48,8 ± 0,4	47,9 ± 0,4
Русский язык					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	94,6 ± 0,5	72,6 ± 1,8	62,9 ± 1,1	55,8 ± 0,8	55,7 ± 0,8
bio	98,5 ± 0,3	69,0 ± 1,8	58,4 ± 1,1	52,8 ± 1,2	52,2 ± 1,2
conversation	95,5 ± 0,5	48,5 ± 1,2	33,8 ± 1,1	27,4 ± 1,2	25,9 ± 1,4
fiction	96,2 ± 0,6	61,0 ± 1,1	47,3 ± 1,2	38,2 ± 0,6	36,7 ± 1,0
interview	96,6 ± 0,3	71,6 ± 1,9	60,3 ± 0,7	47,3 ± 0,8	47,3 ± 0,8
news	96,3 ± 0,5	65,7 ± 2,1	54,2 ± 3,2	47,6 ± 1,2	45,9 ± 1,7
reddit	97,7 ± 0,3	61,1 ± 1,3	48,6 ± 1,6	42,7 ± 1,4	41,5 ± 1,7
speech	96,0 ± 0,6	70,5 ± 2,5	58,7 ± 1,9	50,9 ± 0,5	50,2 ± 0,5
textbook	97,4 ± 0,3	76,0 ± 2,0	62,7 ± 2,3	54,6 ± 2,0	53,6 ± 2,0
vlog	97,9 ± 0,3	65,8 ± 2,1	43,2 ± 2,1	38,8 ± 1,5	35,5 ± 1,3
voyage	99,0 ± 0,1	73,7 ± 0,9	58,1 ± 0,8	50,4 ± 1,2	49,3 ± 1,0
whow	97,8 ± 0,5	75,5 ± 1,7	64,4 ± 2,3	55,5 ± 2,1	54,1 ± 2,2
все	96,9 ± 0,2	66,5 ± 0,4	53,3 ± 0,6	45,8 ± 0,5	44,6 ± 0,4

ние переноса (с русского на английский язык) привело к значительному падению качества. Оценка F1 для сегментации на английском русскоязычной моделью составляет всего 86,9%. Это несоответствие связано с сильной зависимостью от знаков препинания для разделения элементарных дискурсивных единиц в русском языке. В корпусе GUM всего 18,5% ЭДЕ

заканчиваются запятыми, в то время как в RRG доля таких ЭДЕ составляет 37,5%. Это привело к переобучению сегментатора для русского языка на графическом признаке конца ЭДЕ, менее распространенном в английском языке.

Таблица 18 — Оценка непосредственного межъязыкового переноса моноязыковых анализаторов по жанрам, в %

Русский → **английский**					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	83,1 ± 1,3	52,0 ± 4,3	43,2 ± 3,2	39,0 ± 3,0	38,7 ± 2,9
bio	94,4 ± 0,5	63,0 ± 1,8	50,1 ± 2,8	45,9 ± 3,0	44,8 ± 3,2
conversation	91,6 ± 0,6	42,4 ± 1,7	30,8 ± 2,0	23,5 ± 1,2	22,8 ± 1,4
fiction	85,3 ± 0,8	47,8 ± 2,6	35,9 ± 2,6	28,8 ± 1,9	27,7 ± 1,7
interview	83,2 ± 1,4	43,9 ± 3,6	37,1 ± 2,3	29,6 ± 2,9	29,5 ± 2,7
news	84,5 ± 1,8	45,9 ± 3,3	38,7 ± 3,4	36,9 ± 2,9	34,8 ± 2,6
reddit	83,1 ± 1,4	37,1 ± 2,7	30,7 ± 1,8	24,9 ± 2,5	24,6 ± 2,3
speech	83,7 ± 1,6	44,6 ± 2,1	34,8 ± 1,3	29,8 ± 2,4	29,5 ± 2,5
textbook	87,8 ± 1,4	56,2 ± 2,3	45,7 ± 2,3	39,9 ± 2,3	39,2 ± 2,1
vlog	88,1 ± 1,9	52,7 ± 3,4	35,7 ± 3,1	32,8 ± 3,5	30,2 ± 3,9
voyage	85,1 ± 1,2	46,6 ± 2,6	34,9 ± 2,3	28,8 ± 1,7	28,7 ± 1,5
whow	90,6 ± 1,8	58,7 ± 3,8	49,8 ± 3,9	42,9 ± 3,2	42,1 ± 3,0
все	86,9 ± 1,0	49,0 ± 2,2	38,6 ± 2,1	33,1 ± 1,9	32,2 ± 1,9
Английский → **русский**					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	93,1 ± 0,9	69,2 ± 0,8	61,5 ± 0,2	52,1 ± 0,9	52,1 ± 0,9
bio	97,3 ± 0,4	66,3 ± 1,0	54,6 ± 0,5	47,5 ± 0,8	46,3 ± 0,9
conversation	94,4 ± 0,7	45,5 ± 2,5	32,9 ± 3,3	23,2 ± 2,5	22,1 ± 2,4
fiction	94,9 ± 0,7	60,0 ± 2,4	48,1 ± 2,8	38,1 ± 1,8	37,2 ± 1,7
interview	95,6 ± 0,8	69,7 ± 1,3	58,2 ± 1,0	46,9 ± 0,8	46,1 ± 1,0
news	93,5 ± 1,2	61,9 ± 1,2	51,8 ± 1,7	45,8 ± 0,9	44,4 ± 1,0
reddit	97,1 ± 0,4	59,8 ± 2,1	48,5 ± 1,7	41,1 ± 0,8	40,6 ± 0,8
speech	94,5 ± 0,5	69,8 ± 1,1	56,6 ± 0,9	48,6 ± 1,6	47,8 ± 1,3
textbook	95,1 ± 0,3	71,2 ± 0,9	58,0 ± 1,8	51,9 ± 0,6	51,4 ± 0,6
vlog	97,2 ± 0,1	61,6 ± 1,8	41,5 ± 0,6	36,1 ± 1,0	33,3 ± 0,7
voyage	96,7 ± 0,3	71,6 ± 1,4	55,2 ± 1,6	48,8 ± 2,4	46,8 ± 1,9
whow	96,5 ± 0,5	74,0 ± 1,6	61,9 ± 1,8	54,1 ± 1,7	52,0 ± 1,9
все	95,5 ± 0,3	63,9 ± 0,7	51,4 ± 1,0	43,4 ± 0,6	42,2 ± 0,6

Добавление параллельных обучающих данных

Цель этого эксперимента состоит в оценке объёма данных, необходимых для успешного межъязыкового переноса в полнотекстовом риторическом анализе. Необходимо заметить, что задача построения риторического анализатора требует трудоемкой экспертной разметки, поэтому важно оценить объём данных, при добавлении которых в обучающую выборку можно достигнуть удовлетворительного качества риторического анализа на втором языке. Качество межъязыкового переноса оценивается при разных объёмах параллельной разметки: от 25% до 100% корпуса на целевом языке. Оценка предполагает идеальный сценарий, включающий полную параллельную разметку на втором (русском) языке.

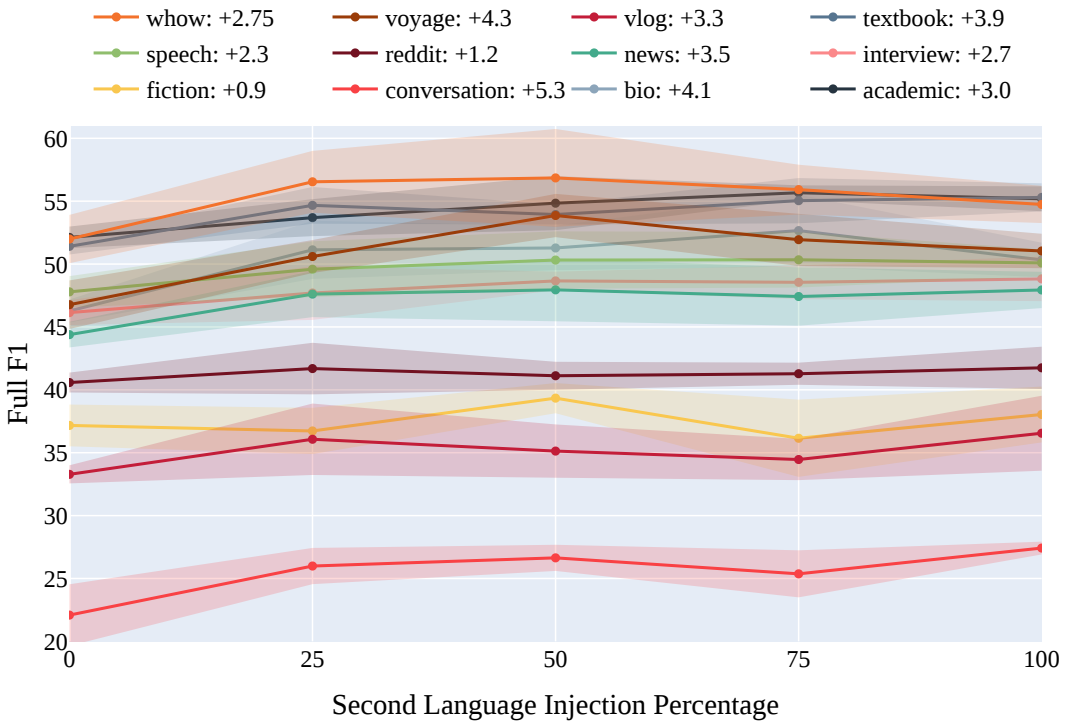


Рисунок 4.7 — Влияние добавления русского языка в обучающие данные на качество риторического анализа по жанрам

В таблице 19 приведены оценки качества модели по мере увеличения числа размеченных примеров на втором языке. Наблюдается постепенное

Таблица 19 — Оценка качества моделей, обученных с добавлением данных на втором языке

Англ.	Рус.	**Английский**				
		Segm.	S	N	R	Full
100%	0%	95,5 ± 0,1	66,9 ± 0,5	56,1 ± 0,3	48,8 ± 0,4	47,9 ± 0,4
	25%	95,5 ± 0,1	66,4 ± 0,7	55,1 ± 1,0	48,2 ± 1,0	47,4 ± 1,0
	50%	95,5 ± 0,1	66,6 ± 0,5	55,4 ± 0,6	48,7 ± 0,6	47,7 ± 0,7
	75%	95,6 ± 0,2	67,2 ± 0,2	55,7 ± 0,5	48,9 ± 0,6	47,9 ± 0,5
	100%	95,3 ± 0,1	66,4 ± 0,7	55,2 ± 0,6	48,6 ± 0,6	47,6 ± 0,7
Англ.	Рус.	**Русский**				
		Segm.	S	N	R	Full
100%	0%	95,5 ± 0,3	63,9 ± 0,7	51,4 ± 1,0	43,4 ± 0,6	42,2 ± 0,6
	25%	96,4 ± 0,3	66,3 ± 0,6	53,8 ± 0,6	45,9 ± 0,7	44,9 ± 0,6
	50%	96,6 ± 0,2	67,0 ± 0,5	54,2 ± 0,6	46,6 ± 0,8	45,5 ± 0,8
	75%	96,8 ± 0,2	67,0 ± 0,4	54,0 ± 0,5	46,2 ± 0,5	45,0 ± 0,5
	100%	96,8 ± 0,1	66,9 ± 0,4	54,3 ± 0,3	46,5 ± 0,4	45,4 ± 0,4

улучшение полного качества модели при адаптации с помощью увеличивающегося объёма данных. Интересно, что точность разметки риторических отношений достигает максимального значения уже примерно на уровне 50% разметки на втором языке. Качество анализа для различных жанров проиллюстрировано на рисунке 4.7. Более подробная оценка по жанрам представлена в таблице 20. Наилучшая адаптация к русскому языку была обнаружена для таких жанров, как *wikihow* (инструкция), *textbook* (учебник), *academic* (академический текст), *voyage* (путеводитель), *bio* (Википедия), *speech* (официальная речь), *interview* (интервью), and *news* (новости). В жанрах спонтанного устного дискурса было достигнуто наихудшее качество анализа, однако добавление параллельных данных позволило сильнее всего адаптировать анализ именно в именно этих жанрах (*vlog* (влог): с 33,3% до 36,6% F1; *conversation* (беседа): с 22,1% до 27,4% F1). Важно заметить, что хотя письменные разделы корпуса могут быть достаточно точно адаптированы на русский язык, передача нюансов спонтанной

Таблица 20 — Оценка двуязычного анализатора по жанрам, в %

Английский + русский → английский					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	94,2 ± 0,4	71,6 ± 1,1	63,1 ± 2,0	55,9 ± 2,1	55,5 ± 2,3
bio	97,6 ± 0,3	70,0 ± 0,9	58,4 ± 1,0	54,0 ± 1,4	52,5 ± 1,5
conversation	95,1 ± 0,1	51,5 ± 1,5	39,2 ± 0,7	31,1 ± 1,4	30,2 ± 1,3
fiction	93,3 ± 0,6	59,2 ± 2,8	48,8 ± 2,3	41,2 ± 1,8	40,2 ± 1,8
interview	94,6 ± 0,5	71,7 ± 1,2	63,5 ± 1,8	55,2 ± 1,3	54,7 ± 1,2
news	94,8 ± 0,7	67,5 ± 2,4	59,2 ± 1,8	54,5 ± 1,6	52,9 ± 1,7
reddit	92,6 ± 0,8	58,5 ± 1,5	48,9 ± 2,3	43,0 ± 2,2	42,3 ± 2,2
speech	97,3 ± 0,3	75,7 ± 1,6	64,8 ± 1,9	57,2 ± 1,1	57,2 ± 1,1
textbook	97,5 ± 0,4	77,3 ± 1,7	65,3 ± 2,0	57,3 ± 0,8	56,4 ± 0,9
vlog	95,9 ± 0,4	62,8 ± 2,0	46,1 ± 2,6	42,8 ± 2,8	40,6 ± 2,7
voyage	94,2 ± 0,5	65,7 ± 2,5	49,5 ± 3,0	43,7 ± 2,6	43,4 ± 2,6
whow	97,2 ± 0,3	75,5 ± 1,3	65,0 ± 1,8	58,3 ± 1,9	56,8 ± 1,6
все	95,3 ± 0,1	66,4 ± 0,7	55,2 ± 0,6	48,6 ± 0,6	47,6 ± 0,7
Английский + русский → русский					
Жанр	**Segm**	**S**	**N**	**R**	**Full**
academic	94,9 ± 0,6	72,9 ± 1,7	63,2 ± 1,6	55,3 ± 1,0	55,2 ± 1,0
bio	98,4 ± 0,4	68,1 ± 1,9	57,5 ± 1,7	51,4 ± 1,4	50,3 ± 1,4
conversation	95,3 ± 0,4	47,8 ± 1,0	34,8 ± 1,3	28,9 ± 0,5	27,4 ± 0,5
fiction	96,6 ± 0,3	62,8 ± 1,9	49,6 ± 0,7	39,2 ± 2,0	38,0 ± 2,2
interview	96,9 ± 0,1	70,0 ± 1,7	60,2 ± 1,9	49,2 ± 1,8	48,8 ± 1,8
news	96,8 ± 0,7	68,5 ± 0,6	56,8 ± 1,7	49,6 ± 1,0	47,9 ± 1,4
reddit	97,2 ± 0,3	60,9 ± 1,6	49,4 ± 2,0	42,5 ± 1,6	41,7 ± 1,7
speech	96,3 ± 0,5	69,9 ± 2,4	57,5 ± 1,0	50,7 ± 1,1	50,1 ± 1,1
textbook	97,1 ± 0,3	77,1 ± 0,6	64,6 ± 1,0	56,1 ± 1,3	55,3 ± 1,1
vlog	97,8 ± 0,5	66,0 ± 1,7	46,0 ± 3,1	39,8 ± 3,4	36,5 ± 3,0
voyage	98,5 ± 0,3	76,4 ± 1,5	60,0 ± 1,9	51,7 ± 1,5	51,0 ± 1,4
whow	97,8 ± 0,3	75,9 ± 1,5	64,5 ± 2,5	56,3 ± 1,1	54,7 ± 1,5
все	96,8 ± 0,1	66,9 ± 0,4	54,3 ± 0,3	46,5 ± 0,4	45,4 ± 0,4

русской речи при переводе разметки, описывающей устный дискурс на английском языке (*vlog*, *conversation*) может быть затруднительна.

Двуязычная модель превосходит моноязыковую на корпусе RRG (44,6% F1), достигая 45,4% Full F1. Это улучшение может быть связано с потенциальными ограничениями предобученной модели `XLM-RoBERTa`

Таблица 21 — Оценка моноязыковой и двуязычной моделей. Full F1, в %

Язык теста	Английский		Русский		
Обуч. данные	GUM	GUM+RRG	GUM	RRG	GUM+RRG
academic	56,3	55,5 (–0,8)	52,1	55,7	55,2 (–0,5)
bio	51,5	52,5 (+1,0)	46,3	52,2	50,3 (–1,9)
conversation	29,3	30,2 (+0,9)	22,1	25,9	27,4 (+1,5)
fiction	38,5	40,2 (+1,7)	37,2	36,7	38,0 (+1,3)
interview	55,1	54,7 (–0,4)	46,1	47,3	48,8 (+1,5)
news	55,0	52,9 (–2,1)	44,4	45,9	47,9 (+2,0)
reddit	44,0	42,3 (–1,7)	40,6	41,5	41,8 (+0,3)
speech	57,6	57,2 (–0,4)	47,8	50,2	50,1 (–0,1)
textbook	57,0	56,4 (–0,6)	51,4	53,6	55,3 (+1,7)
vlog	41,7	40,6 (–1,1)	33,3	35,5	36,6 (+1,1)
voyage	44,1	43,4 (–0,7)	46,8	49,3	51,0 (+1,7)
whow	57,0	56,8 (–0,2)	52,0	54,1	54,7 (+0,6)
all	47,9	47,6 (–0,3)	42,2	44,6	45,4 (+0,8)

в обработке русского языка из-за отсутствия баланса в количестве данных для разных языков в корпусе предобучения CC-100 (токенов для русского языка 23,4 млрд, для английского — 55,6 млрд [138]). Двуязычное обучение, при котором модели предоставляются дискурсивные параллельные данные, помогает смягчить несбалансированность при предобучении и улучшая кодирование моделью текста на русском языке. Несмотря на небольшое снижение качества полного анализа текстов на английском языке, двуязычный парсер превзошел моноязыковой в анализе текстов 9 из 12 жанров на русском языке (см. таблицу 21), что подчеркивает эффективность билингвального обучения для межъязыкового переноса.

4.5.4. Обучение на основе смешанных данных

Анализ совместимости корпусов RRT и RRG

В этом разделе исследуется совместимость между наборами данных в русскоязычном анализе риторических структур путем сравнения наборов отношений, принятых в RRT и RRG.

Для классификации пар дискурсивных единиц, связанных в размеченных структурах, был обучен классификатор отношений для русского языка, описанный в разделе 3.3. Он представляет собой ансамбль классификатора на признаках и модели глубокого обучения на основе ELMo. На тестовых данных корпуса RRT классификатор продемонстрировал качество 48,9% F1, тогда как на тестовых данных RRG — 46,3%. Подробные результаты оценок приведены в таблице 22. Результаты перекрестной классификации отношений между корпусами проиллюстрированы на рисунке 4.9. Эти результаты указывают на соответствие между большинством классов из двух наборов данных, одновременно демонстрируя сложность унификации риторических корпусов.

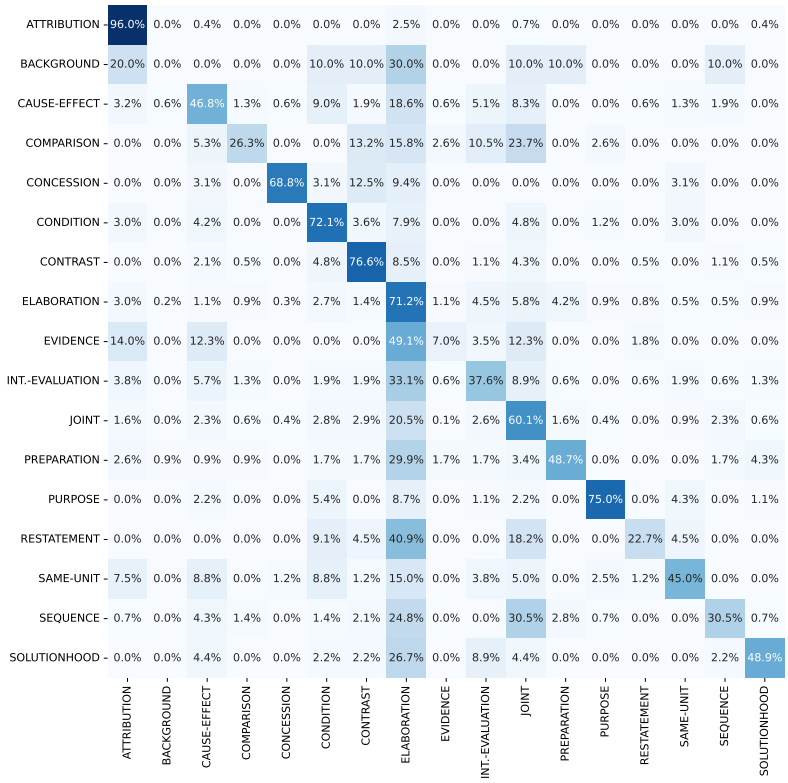
В таблице 22 показаны оценки классификации риторических отношений в каждом корпусе. Задача рассматривается в контексте гибридного риторического анализатора, с объединением отношения и ядерности в один тип отношения. На рисунке 4.8 приведены матрицы ошибок для тех же моделей классификации, с учетом только риторического отношения. Хотя классификатор, обученный на корпусе RRG, демонстрирует лучшее качество классификации для некоторых отношений, зеркальных для RRT (Contingency/Condition, Purpose, Topic/Solutionhood), большую трудность для него представляют причинно-следственные отношения (16,7% для Causal в RRG по сравнению с 46,8% для Cause-Effect в RRT).

Таблица 22 — Качество классификации риторических отношений в русскоязычных корпусах.

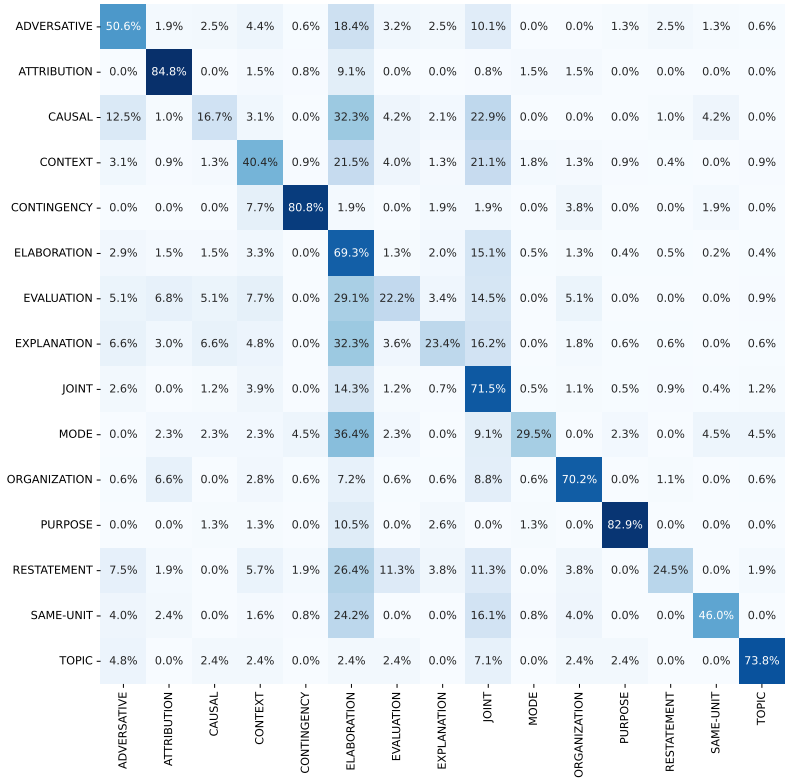
	P	R	F1	Кол-во		P	R	F1	Кол-во
RRT					RRG				
Attribution_NS	87,21	97,40	92,02	77	adversative_NN	24,32	17,31	20,22	52
Attribution_SN	77,05	94,95	85,07	198	adversative_NS	35,85	33,33	34,55	57
Background_SN	00,00	00,00	00,00	10	adversative_SN	36,23	51,02	42,37	49
Cause-effect_NS	50,88	37,18	42,96	78	attribution_NS	84,00	72,41	77,78	29
Cause-effect_SN	43,18	48,72	45,78	78	attribution_SN	69,47	88,35	77,78	103
Comparison_NN	35,71	26,32	30,30	38	causal_NS	29,55	16,46	21,14	79
Concession_NS	83,33	90,91	86,96	22	causal_SN	07,14	05,88	06,45	17
Concession_SN	40,00	20,00	26,67	10	context_NS	60,56	42,16	49,71	102
Condition_NS	53,47	75,00	62,43	72	context_SN	35,24	30,58	32,74	121
Condition_SN	62,38	67,74	64,95	93	contingency_NS	71,43	71,43	71,43	14
Contrast_NN	70,94	76,60	73,66	188	contingency_SN	86,49	84,21	85,33	38
Elaboration_NS	52,72	71,21	60,59	639	elaboration_NS	50,66	69,33	58,54	551
Evidence_NS	26,67	08,89	13,33	45	evaluation_NS	33,80	23,30	27,59	103
Evidence_SN	00,00	00,00	00,00	12	evaluation_SN	50,00	07,14	12,50	14
Int.-evaluation_NS	45,24	39,58	42,22	144	explanation_NS	54,41	26,62	35,75	139
Int.-evaluation_SN	33,33	15,38	21,05	13	explanation_SN	20,00	03,57	06,06	28
Joint_NN	72,18	60,12	65,60	682	joint_NN	60,69	71,48	65,64	568
Preparation_SN	56,44	48,72	52,29	117	mode_NS	46,43	31,71	37,68	41
Purpose_NS	89,06	78,08	83,21	73	mode_SN	00,00	00,00	00,00	3
Purpose_SN	55,00	57,89	56,41	19	organization_NS	73,68	96,55	83,58	29
Restatement_NN	33,33	22,73	27,03	22	organization_SN	78,57	65,13	71,22	152
Sequence_NN	59,72	30,50	40,38	141	purpose_NS	85,07	82,61	83,82	69
Solutionhood_SN	51,16	48,89	50,00	45	purpose_SN	75,00	85,71	80,00	7
same-unit_NN	59,02	45,00	51,06	80	restatement_NN	37,50	32,14	34,62	28
					same-unit_NN	82,61	45,97	59,07	124
					topic_SN	63,27	73,81	68,13	42
Macro avg.	51,58	48,41	48,92	2896	*Macro avg.*	50,69	45,64	46,30	2584

Это может быть связано с зависимостью классификатора от дискурсивных коннекторов. Всего 23,6% пар дискурсивных единиц в RRG с коннектором, характерным для причинно-следственной связи, действительно формируют причинно-следственное отношение, в то время как в RRT доля таких пар составляет 47,7%. Наиболее распространенными отношениями с неоднозначным маркером причины в RRG являются Explanation (13,9%), Elaboration (11,6%), Joint (10,0%) и Context (8,4%).

Соотношения риторических классов отношение_ядерность в двух корпусах иллюстрирует рисунок 4.9. Уверенно предсказанные отношения

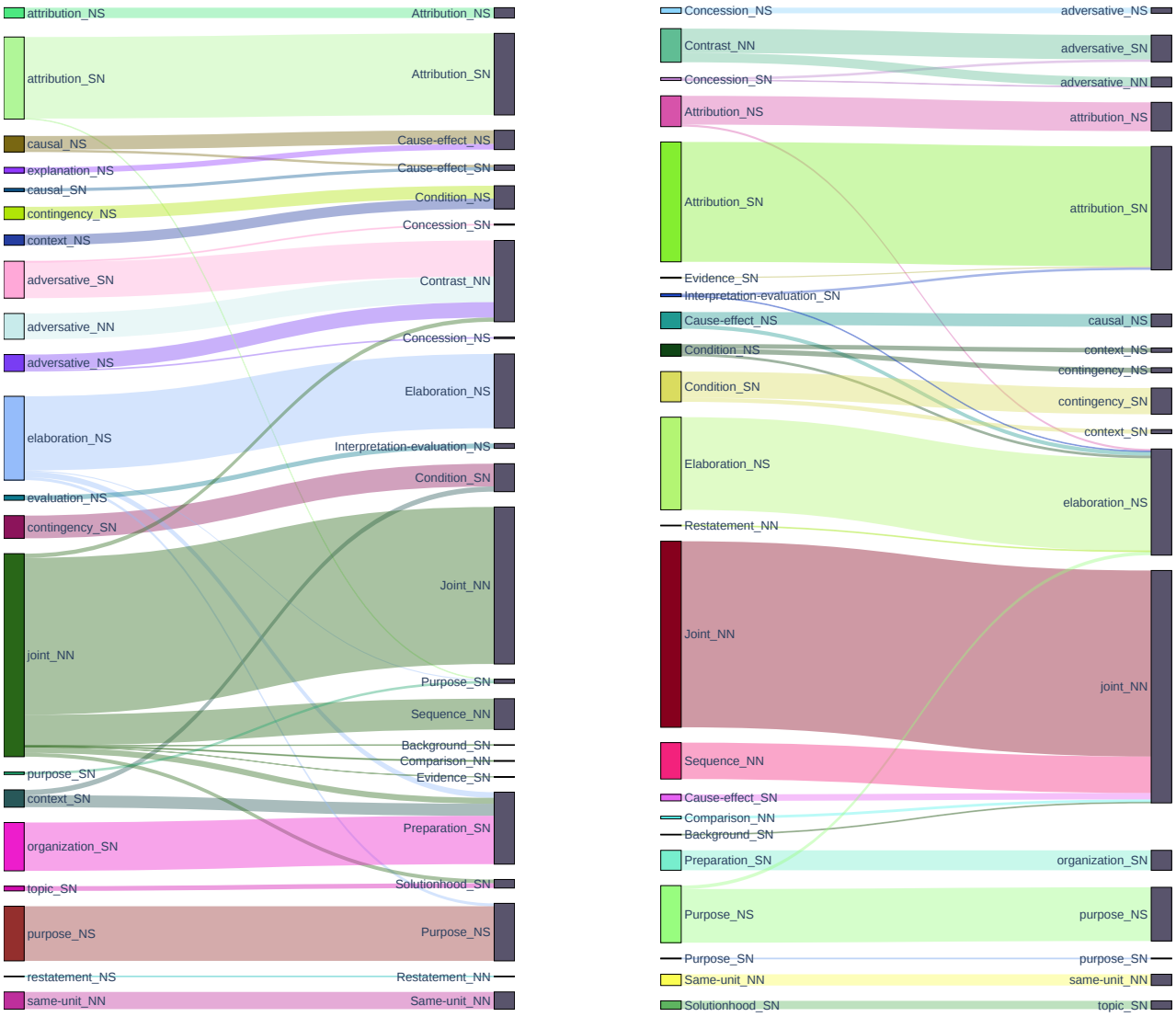


(а) RRT



(б) RRG

Рисунок 4.8 — Матрицы ошибок для классификации риторических отношений в русскоязычных корпусах; без учета положения ядра.



(а) Классификтор RRT → RRG (б) Классификатор RRG → RRT

Рисунок 4.9 — Иллюстрация согласованности в перекрёстной классификации на двух корпусах.

(энтропия выше 75-го процентиля) показаны справа, истинные отношения из целевого корпуса — слева. Включены только частые переходы ($> 2{,}5\%$ истинного класса). Таким образом, изображены тенденции совпадений между отношениями в двух типах разметки. Классы Organization_NS, Mode, Context_SN, and Organization_NS в корпусе RRG не соответствуют конкретным классам в RRT при анализе используемых классификатором признаков дискурсивных единиц. Классификатор, обученный на корпусе RRT, последовательно присваивает класс Condition классу contingency из RRG (детальное отношение contingency-condition), так и context

(context-circumstance) из RRG. В корпусе RRG детализированные классы Антитеза, Уступка, Контраст объединяют в общий тип отношения верхнего уровня Adversative, используемый при разработке методов автоматического анализа. В RRT отсутствует иерархия классов, и объединенная в RRG категория соответствует в нем двум типам отношения: Contrast и Concession, что приводит к несоответствию определений положения ядра. На материалах обоих корпусов классификаторы склонны к схожим ошибкам. К примеру, несмотря на наличие собственного специального подкласса Evidence в более общей категории explanation, классификатор, обученный на корпусе RRG, склонен к ошибочной классификации примеров отношения Evidence из RRT как attribution; такую же ошибку в 14% случаев совершает и RRT-классификатор. Это указывает на склонность обеих моделей интерпретировать ссылки на источники информации как атрибуции вне зависимости от предполагаемого значения. Классу Cause-effect из RRT соответствуют частные случаи подклассов Justify и Motivation из общей категории RRG explanation, охватывающего как причинные, так обоснователные отношения (кроме подкласса Evidence).

Риторический анализ на основе смешанных данных

Эксперименты проводились на объединенном корпусе, состоящем из данных двух крупных русскоязычных корпусов: RRT и RRG. Исследовались возможности обобщения данных разметки с различающимися принципами в условиях одного языка (русского).

Прежде всего, в задаче риторического анализа текстов на русском языке были протестированы современные моноязыковые модели. Результаты оценки качества гибридного метода, описанного в разделе 4.2, с использованием предобученных кодирующих языковых моделей для русского

языка (таблицы 23, 24) показали, что наилучшее качество полного разбора для русского языка в среднем демонстрирует анализатор, использующий для кодирования русского языка относительно компактную модель `ai-forever/ruBert-base`. Эта модель использовалась в экспериментах со смешением корпусов RRT и RRG.

Таблица 23 — Показатели качества (3 запуска, F1, в %) полного риторического разбора; корпус RRT.

Языковая модель	Seg	S	N	R	Full
`ai-forever/ruRoberta-large` [145]	90,1 ± 0,0	61,5 ± 0,4	46,7 ± 0,5	39,8 ± 0,7	39,8 ± 0,7
`DeepPavlov/rubert-base-cased` [146]	91,3 ± 0,5	63,2 ± 0,7	48,7 ± 0,4	41,5 ± 0,6	41,3 ± 0,6
`hivaze/ru-e5-large`	91,4 ± 0,4	64,4 ± 0,8	49,4 ± 0,8	41,5 ± 0,9	41,4 ± 0,9
`Tochka-AI/ruRoPEBert-e5-base-2k`	91,6 ± 0,2	64,2 ± 0,6	50,0 ± 0,7	43,0 ± 0,6	42,8 ± 0,6
`deepvk/deberta-v1-base`	91,9 ± 0,4	64,6 ± 0,3	50,2 ± 0,2	42,9 ± 0,3	42,8 ± 0,3
`ai-forever/ruBert-large` [145]	92,0 ± 0,1	65,3 ± 0,5	50,8 ± 0,5	43,0 ± 0,5	42,9 ± 0,5
`ai-forever/ruBert-base` [145]	91,6 ± 0,3	64,5 ± 0,4	50,8 ± 0,5	43,5 ± 0,6	43,3 ± 0,6

Таблица 24 — Показатели качества (3 запуска, F1, в %) полного риторического разбора; корпус RRG.

Языковая модель	Seg	S	N	R	Full
`deepvk/deberta-v1-base`	92,1 ± 0,6	55,6 ± 0,7	43,3 ± 0,4	36,2 ± 0,6	35,3 ± 0,8
`DeepPavlov/rubert-base-cased`	94,7 ± 0,2	56,8 ± 0,5	44,3 ± 0,7	37,4 ± 0,7	36,4 ± 0,6
`Tochka-AI/ruRoPEBert-e5-base-2k`	94,4 ± 0,5	58,1 ± 0,8	45,7 ± 0,5	39,8 ± 0,2	38,5 ± 0,2
`ai-forever/ruBert-large`	95,0 ± 0,5	60,2 ± 0,6	47,9 ± 0,4	40,5 ± 0,5	39,4 ± 0,5
`hivaze/ru-e5-large`	95,1 ± 0,2	61,4 ± 0,3	48,2 ± 0,3	41,5 ± 0,4	40,3 ± 0,4
`ai-forever/ruBert-base`	94,9 ± 0,1	61,6 ± 0,3	49,4 ± 0,3	41,8 ± 0,3	41,0 ± 0,4
`ai-forever/ruRoberta-large`	94,6 ± 0,5	63,2 ± 0,5	50,5 ± 0,5	42,5 ± 0,5	41,5 ± 0,6

Проведены следующие эксперименты: (1) Обучение на смешении корпусов с учетом в архитектуре только различных наборов классов отношений (общая сегментация); (2) Обучение на смешении корпусов с учетом наборов отношений и диктуемой ими сегментации; (3) Добавление к смешанным русскоязычным данным автоматически переведенных на уровне ЭДЕ корпусов RST-DT и GUM_{10}[6]. Для автоматического перевода англоязычных данных

[6] Добавленные в последней версии корпуса 22 документа дополнительных жанров: эссе, письмо, подкаст, записи судебных заседаний.

использована модель NLLB [147]. При обучении риторического анализатора значение функции потерь для примеров из автоматически переведенных данных взвешивается с коэффициентом 1/3.

Таблица 25 — Показатели качества (3 запуска, F1, в %) полного риторического разбора; RRT+RRG.

Обучающие данные		RRT			RRG		
		Seg	S	Full	Seg	S	Full
RRT		91,6 ± 0,3	64,5 ± 0,4	43,3 ± 0,6	87,3 ± 0,2	44,6 ± 1,0	-
RRG		82,6 ± 0,4	48,4 ± 0,5	-	94,9 ± 0,1	61,6 ± 0,3	41,0 ± 0,4
RRT + RRG	(1)	91,4 ± 0,2	64,7 ± 0,4	43,9 ± 0,8	94,0 ± 0,2	58,4 ± 0,3	37,8 ± 0,7
	(2)	91,4 ± 0,3	65,0 ± 0,2	43,1 ± 1,0	94,6 ± 0,5	59,9 ± 0,7	38,4 ± 1,8
RRT + RRG + RST-DT + GUM_{10}	(3)	90,9 ± 0,5	64,7 ± 1,2	43,8 ± 1,0	94,8 ± 0,3	60,8 ± 0,7	39,6 ± 1,1

Результаты экспериментов представлены в таблице 25. Обучение на двойном наборе русскоязычных данных приводит к улучшению качества автоматического анализа при оценке на корпусе экспертной разметки RRT. Между корпусами наблюдаются важные отличия в сегментации, обусловленные как различиями в наборах отношений, так и формальными ограничениями (например, в RRT, в отличие от других корпусов, не обязательно выделять в отдельную дискурсивную единицу содержание скобок); разделение сегментаторов для отдельных корпусов также позволило улучшить качество разбора.

4.6. Выводы

В этой главе предложены модификации метода DMRST для полнотекстового нисходящего разбора риторической структуры текста, позволяющие более точно моделировать элементарные дискурсивные единицы в рамках гибридного разбора, а также обучать риторический анализатор этой

архитектуры на смешении непосредственно несовместимых данных. Экспериментально показано, что качество предложенного метода риторического анализа текстов «с нуля» превосходит предыдущие подходы для английского языка. Для русского языка достигнуто качество анализа текстовых фрагментов различных жанров и источников (45,3% Full F1), сопоставимое с качеством анализа новостных заметок на английском языке (53,0% Full F1).

Приведены оценки кросс-языковой обобщаемости полнотекстового риторического анализа. Разработан параллельный корпус риторической разметки RRG, обучены двуязычные модели, демонстрирующие высокое качество анализа текстов на русском (45,4% Full F1) и английском (47,6% Full F1) языке. Показано, что перенос знаний из разметки на английском языке позволяет достигать результатов, сопоставимых с моноязыковыми моделями, даже при отсутствии значительного объёма разметки текстов на русском языке.

Одной из ключевых проблем обобщаемости риторического анализа является разнообразие наборов и определений риторических отношений в размеченных корпусах. Предложенный подход, предусматривающий обучение универсального анализатора с отдельными модулями сегментации и классификации для каждого набора данных, позволяет обучать модели риторического анализа на непосредственно несовместимых данных риторической разметки. Это позволяет увеличить объём обучающих данных и получать более жанрово-универсальные анализаторы. Обучены модели жанрово-универсального полнотекстового разбора текстов на русском языке, использующие информацию о риторических структурах из корпусов RRT и RRG.

Полученные результаты подтверждают эффективность предложенных методов, а также демонстрируют потенциал кросс-языкового и кросс-жанрового риторического анализа. Результаты исследований опубликованы в работах «Методы анализа риторических структур в текстах на рус-

ском языке» [3] и “Bilingual Rhetorical Structure Parsing with Large Parallel Annotations” [2].

Глава 5. Приложения методов анализа риторических структур в задачах обработки естественного языка

5.1. Введение

Автоматический анализ риторической структуры текста позволяет определять его иерархическую дискурсивную структуру. Эти структуры используются для решения различных задач обработки языка, которые рассмотрены в работе:

- **Классификация текстов**. Анализ риторической структуры применим при классификации текстов по типу, жанру, тематике или авторской позиции. Комбинации признаков дискурсивных единиц в риторическом дереве в соответствии с его структурой [9; 32; 33; 148–150] позволяют более точно определить точку зрения автора или основную мысль текста, чем анализ последовательностью слов. Формальные дискурсивные признаки могут также отражать характерные шаблоны изложения, присущие различным типам дискурса или авторам [151–156].
- Кореференция — это дискурсивный феномен, при котором разные языковые выражения (например, местоимения, имена собственные) обозначают одну и ту же сущность. Важным фактором выбора говорящим языкового выражения для каждого упоминания сущности (*референциального выбора*) является расстояние между ними в тексте. Иерархическая структура текста с выделением ядерных высказываний позволяет получить более точные с точки зрения активации сущностей в памяти говорящего метрики, чем подсчёт последовательности слов между упоминаниями, тем самым позволяя достичь более точного автоматического **разрешения кореференции**.

– Схема аргументации в тексте-рассуждении во многом совпадает с его риторической структурой. Таким образом, использование РРС в **анализе аргументации** позволяет значительно улучшить качество результатов по сравнению с методами, обрабатывающими неструктурированный текст [25; 42; 44; 157].

Глава посвящена методам решения прикладных задач с использованием риторических структур. В разделе 5.2 предлагается метод классификации текстов, позволяющий улучшить качество классификации, принимая во внимание иерархическую риторическую структуру. В разделе 5.3 описаны исследования признаков риторической структуры применительно к разрешению кореференции в текстах на русском языке. В разделе 5.4 предложен способ анализа структур аргументации с нуля, опирающийся на результаты риторического разбора текста-рассуждения.

5.2. Применение анализа риторических структур к задачам классификации текстов

5.2.1. Введение

Риторическая структура текста отражает сложные процессы убеждения и аргументации, используемые автором для выражения своих идей. Ядром риторического отношения является дискурсивная единица, наиболее близкая к основной мысли текста. Понятие ядерности легло в основу работ по определению класса (чаще всего, тональности) всего текста на основе его иерархической риторической структуры. В работе [148] прилагательные в тексте взвешивают в зависимости от их положения относительно основного ядра предложения. Идея развита в последующем исследовании [149], посвященном детальному анализу стратегий взвешивания всей оценочной

лексики с учетом ядерности и типа риторического отношения. Определение тональности предложения как комбинации полярностей элементарных дискурсивных единиц успешно реализуется с помощью простых правил для контрастных отношений (уступка, контраст/антитезис) [150]. Учёт риторической структуры текста продемонстрировал преимущество и в классификации текстов с применением глубокого обучения. Учет риторической структуры также показал свою эффективность в классификации текстов с применением глубокого обучения. Для этого предлагаются различные методы: настройка параметров для каждого отношения с построением графа отношений, зеркального для каждой риторической структуры [158]; обучение представлений дискурсивных единиц на каждом уровне риторической структуры [32]; использование архитектуры Tree LSTM для кодирования дерева предложения с назначенными ядрами (без типов отношений) [159] или для получения представления дерева всего документа на основе обучаемых представлений предложений и риторических отношений между ними [33].

Другим направлением применения дискурсивного разбора в классификации текстов является анализ формальных дискурсивных признаков, выявляющий различия между классами, определяемые не основной идеей, а способами её выражения. Чаще всего обнаруживают различия в распределении типов риторических отношений. В отличие от анализа композиции высказываний, направленного на выявление основной мысли автора и чаще всего применяющийся в анализе тональности или аргументации, данный вид анализа ориентирован на определение специфических шаблонов, характерных для разных типов дискурса. Например, дискурса людей в различных психологических состояниях [151; 152], дезинформации [153], автоматически сгенерированного дискурса [154; 155]. Так, исследование с использованием риторических анализаторов для русского и английского языков [156] выявило высокую эффективность признаков наличия отдельных риторических отношений (Атрибуция, Фон, Сравнение и проч.) для

идентификации использования в текстах новостей различных техник убеждения.

В этом исследовании рассматривается возможность улучшения представления текста на естественном языке для решения задачи классификации за счет использования его иерархической риторической структуры. Основным недостатком предыдущих подходов является переобучение, связанное с синхронным обучением представлений элементарных или всех дискурсивных единиц в дереве. Предлагаемый в этой главе двухэтапный метод позволяет сохранять преимущества современных методов классификации текста как последовательности символов и повышать их качество за счет постобработки предсказаний классификатора в небольшой модели глубокого обучения, вычисляющей общий класс текста на основе его риторической структуры.

5.2.2. Метод интеграции дискурсивной структуры текста в классификацию с использованием языковых моделей

Метод классификации с учетом иерархии дискурсивной структуры основан на классификации отдельных высказываний с последующим рассчетом общего класса текста по дискурсивному дереву. Классификация отдельных высказываний (составляющих риторическое дерево ДЕ) осуществляется посредством любого современного метода классификации текста как последовательности слов. Расчёт общего класса текста по его риторической структуре с классифицированными дискурсивными единицами реализуется через глубокое обучение структурной реккурентной сети Tree LSTM. Таким образом преодолеваются ограничения современных методов классификации текста, внимание в которых реализовано для обработки последовательностей символов, а не сложных дискурсивных структур.

На первом шаге получают предсказания распределений классов для каждой дискурсивной единицы из структуры. Затем обходят бинарное риторическое дерево снизу вверх, преобразуя информацию из левой и правой частей каждого узла дерева для получения представления ДЕ на верхнем уровне, пока не будет достигнут корень дерева (ДЕ, покрывающая весь текст). Предлагается использовать рекуррентную архитектуру Binary Tree LSTM [160]. В ней скрытое состояние и состояние ячейки неэлементарных ДЕ определяется скрытыми состояниями их левых и правых составляющих, а не последовательностью слов внутри этих единиц. Это позволяет обрабатывать самодостаточные фразы в рамках сложного дискурса.

По сравнению с предыдущими исследованиями, использовавшими архитектуру Tree LSTM для обработки риторической структуры в задачах классификации документов, предлагается не классифицировать каждый узел дерева на основе текстовых признаков [159; 161] или лексических оценок [158; 162], а использовать результат работы предварительно обученного классификатора фраз и тип риторических отношений в качестве признаков для каждого узла. Целью является предсказание единственной метки для корня документа. Для всех текстов в рамках рассматриваемых задач определяется один общий класс.

Таким образом, предлагается модель глубокого обучения для агрегирования меток классов, предсказанных для всех дискурсивных единиц документа с помощью современного классификатора последовательностей токенов. Во-первых, это обеспечивает лучшее качество классификации текста как последовательности токенов при помощи языковых моделей, которые сами по себе учитывают некоторые аспекты локального дискурса [163]. Во-вторых, было установлено, что методы, основанные на одновременном обучении представлений ДЕ высокой размерности совместно с Tree LSTM, склонны к сильному переобучению [162]. Следовательно, важно, чтобы исходные представления узлов риторического дерева были компактными

и информативными, что и достигается за счет кодирования с помощью предварительно обученного классификатора.

При формировании исходного векторного представления каждого узла используются шесть общих типов отношений из корпуса RRT:

- Связность (Background, Elaboration, Restatement, Interpretation-evaluation, Preparation),
- Каузально-аргументативные: контраст (Concession, Contrast, Comparison),
- Каузально-аргументативные: причина (Purpose, Evidence, Cause-effect),
- Каузально-аргументативные: условие (Condition),
- Структурные (Sequence, Joint, Same-unit),
- Атрибуция (Attribution).

Алгоритм классификации текста предложенным методом можно формально описать следующим образом:

1. Пусть задан текст D. При помощи методов автоматического риторического анализа определим риторическое дерево текста $T = (V,E)$, где V — множество узлов (ДЕ), а E — множество рёбер, задающих бинарные иерархические отношения между ДЕ. Корневой узел $r_{\text{root}} \in V$ соответствует всему тексту.
2. Для каждой дискурсивной единицы $v_i \in V$ вычислим векторное представление:

$$\mathbf{u}_i = [\text{FC}(\text{Enc}(\text{du}_i)); r_i], \tag{5.1}$$

где:

- $\text{Enc}(\text{du}_i) \in \mathbb{R}^e$ — векторное представление последовательности токенов ДЕ, полученное предварительно обученным кодировщиком (например, языковой моделью);
- $\text{FC} : \mathbb{R}^e \to \mathbb{R}^K$ — распределение вероятностей классов для текущей ДЕ как независимого документа, полученное

в классификаторе последовательности токенов, обучаемом отдельно;

– $r_i \in \mathbb{R}^{12}$ — бинарный вектор, кодирующий тип дискурсивного отношения и ядерность узла. Так, отношение Attribution_NS соответствует меткам Attribution_Nucleus и Attribution_Satellite для его левой и правой составляющей. Корневой узел, не имеющий родительского узла, имеет тип отношения Root. Так как все структурные отношения являются многоядерными, одноразмерный вектор r_i для меток дискурсивных единиц (таких как Coherence_Nucleus, Coherence_Satellite, Root) в нашей модели имеет длину 12.

3. Для учёта иерархической структуры и агрегации информации от дочерних узлов дерева к корневой ДЕ применим глубокую архитектуру Tree LSTM. Пусть для каждого узла $j \in V$ $\mathbf{u}_j$ — входной вектор (формула 5.1); $\mathbf{c}_j \in \mathbb{R}^d$ — вектор состояния; $\mathbf{h}_j \in \mathbb{R}^d$ — выходной вектор. Для узла j с левым потомком $L(j)$ и правым потомком $R(j)$ состояния Tree LSTM вычисляются следующим образом:

$$\mathbf{i}_j = \sigma(W^{(i)}\mathbf{u}_j + U_L^{(i)}\mathbf{h}_{L(j)} + U_R^{(i)}\mathbf{h}_{R(j)} + b^{(i)}), \tag{5.2}$$

$$\mathbf{f}_{j,L} = \sigma(W^{(f)}\mathbf{u}_j + U_{L,L}^{(f)}\mathbf{h}_{L(j)} + U_{L,R}^{(f)}\mathbf{h}_{R(j)} + b_L^{(f)}), \tag{5.3}$$

$$\mathbf{f}_{j,R} = \sigma(W^{(f)}\mathbf{u}_j + U_{R,L}^{(f)}\mathbf{h}_{L(j)} + U_{R,R}^{(f)}\mathbf{h}_{R(j)} + b^{(f)}), \tag{5.4}$$

$$\mathbf{o}_j = \sigma(W^{(o)}\mathbf{u}_j + U_L^{(o)}\mathbf{h}_{L(j)} + U_R^{(o)}\mathbf{h}_{R(j)} + b^{(o)}), \tag{5.5}$$

$$\tilde{\mathbf{c}}_j = \tanh(W^{(u)}\mathbf{u}_j + U_L^{(u)}\mathbf{h}_{L(j)} + U_R^{(u)}\mathbf{h}_{R(j)} + b^{(u)}), \tag{5.6}$$

$$\mathbf{c}_j = \mathbf{i}_j \odot \tilde{\mathbf{c}}_j + \mathbf{f}_{j,L} \odot \mathbf{c}_{L(j)} + \mathbf{f}_{j,R} \odot \mathbf{c}_{R(j)}, \tag{5.7}$$

$$\mathbf{h}_j = \mathbf{o}_j \odot \tanh(\mathbf{c}_j), \tag{5.8}$$

где W и U — матрицы параметров, b — обучаемый вектор, $\mathbf{i}_j$, $\mathbf{f}_{j,L}$, $\mathbf{f}_{j,R}$, $\mathbf{o}_j$ — векторы LSTM-вентилей, $\sigma(\cdot)$ — сигмоидная функция, $\odot$ — поэлементное умножение.

Таким образом, каждый узел агрегирует информацию о дочерних ДЕ в скрытое состояние и память, передавая эту информацию выше по дереву.

4. В результате рекурсивного вычисления состояний получим векторное представление корневой ДЕ $\mathbf{h}_{\text{root}} \in \mathbb{R}^d$.

   Финальное предсказание класса документа осуществляется следующим образом:

$$y = \text{softmax}(W_{\text{final}}^{\top}\mathbf{h}_{\text{root}} + b_{\text{final}}), \tag{5.9}$$

   где $W_{\text{final}} \in \mathbb{R}^{d \times K}$ и $b_{\text{final}} \in \mathbb{R}^K$ — обучаемые параметры классификатора на K классов.

5.2.3. Экспериментальное исследование метода классификации текстов при помощи риторических структур

Данные

Набор данных для совместной классификации позиций и аргументов предоставлен организаторами конкурса RuArg-2022. Метка позиции отражает точку зрения автора по отношению к заданному утверждению. Наличие аргументов «за», «против» или смешанных обозначается меткой аргументации. В наборе данных представлено три утверждения, касающихся COVID-19: «Ношение масок полезно для общества», «Вакцинация полезна для общества» и «Введение и соблюдение карантина полезно для общества».

На рисунке 5.1 показано распределение длин текстов в ЭДЕ, выделенных автоматическим анализатором. Если текст состоит из l элементарных дискурсивных единиц, его риторическая структура включает $l - 1$ отношений. Каждый поднабор данных содержит около 25% простых предложений, не содержащих автоматически распознаваемой дискурсивной структуры. Из этого можно сделать вывод, что для 75% данных можно улучшить качество классификации за счет анализа связности текста. Большинство примеров содержат только одно дискурсивное отношение между двумя элементарными единицами; в официальном тестовом наборе такое наблюдается в 35,9% примеров.

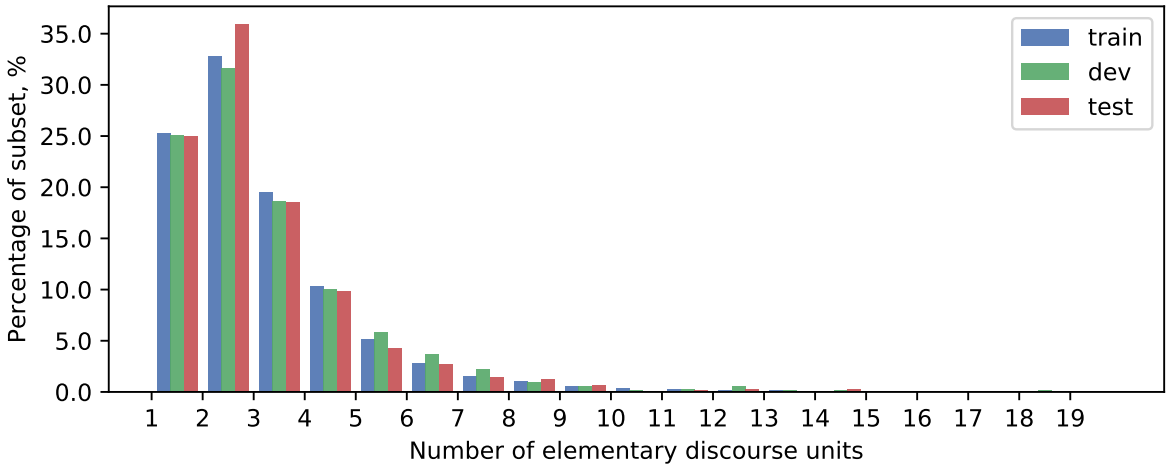
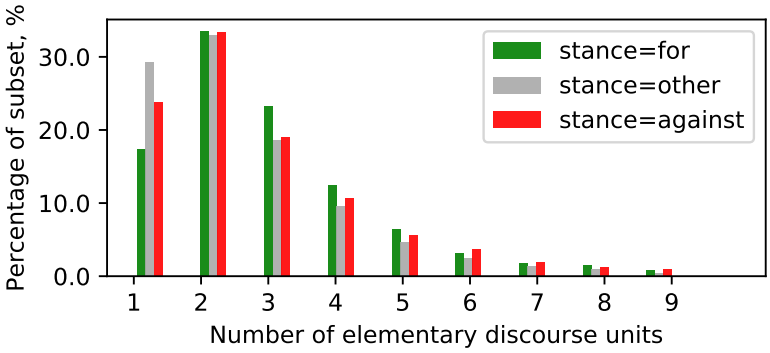


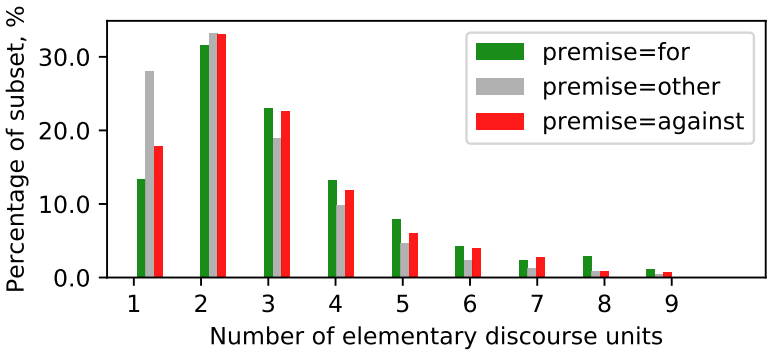
Рисунок 5.1 — Длины текстов в ЭДЕ, набор данных RuArg-2022

На рисунке 5.2 показано распределение длины текстов в различных классах (вне зависимости от тематики) в размеченных обучающих данных. Интересно, что полярные мнения в корпусе чаще выражаются в текстах со сложной риторической структурой. Простые предложения чаще встречаются среди примеров смешанного класса *Other*, что особенно заметно в подзадаче классификации аргументов (рисунок 5.2б). Это показывает, что большинство примеров в этом классе не содержат типичных маркеров аргументации — причинно-следственных, условных или других значимых дискурсивных отношений [164–166], — что соответствует определению класса *Other* в подзадаче классификации аргументов. Другим

интересным наблюдением является то, что примеры, в которых автор выражает положительную позицию или приводит аргумент в пользу, имеют наиболее сложные структуры в обучающем наборе.



(а) Stance detection



(б) Premise classification

Рисунок 5.2 — Длины текстов в ЭДЕ в размеченной обучающей выборке

Детали экспериментов

В качестве базовой модели классификации используется модель BERT, предобученная на русскоязычных текстах из социальных сетей (`DeepPavlov/rubert-base-cased-conversational`); в качестве вектора текста используется закодированное значение фиктивного токена `[CLS]`, добавляемого к последовательности входных токенов. В рамках решения задачи каждый текстовый классификатор имеет два выхода — для классификации позиции автора и класса аргументации. Классификация

осуществляется в двух полносвязных слоях, обучаемых синхронно с дообучением внутренних представлений. Классификаторы для разных тематик обучаются отдельно.

Данные разных классов в наборе данных несбалансированы, с преобладанием класса *Irrelevant*. Поэтому в функцию потерь добавлены весовые коэффициенты для каждого класса при обучении как BERT-классификатора, так и RST-LSTM. Веса задаются в соответствии с распределением классов в обучающем наборе данных.

Для автоматической настройки гиперпараметров, как при дообучении модели BERT, так и при обучении RST-LSTM, использован фреймворк для оптимизации Optuna [167]. Оптимальный размер скрытого слоя Tree LSTM для моделей, связанных с различными темами, варьируется от 50 до 125.

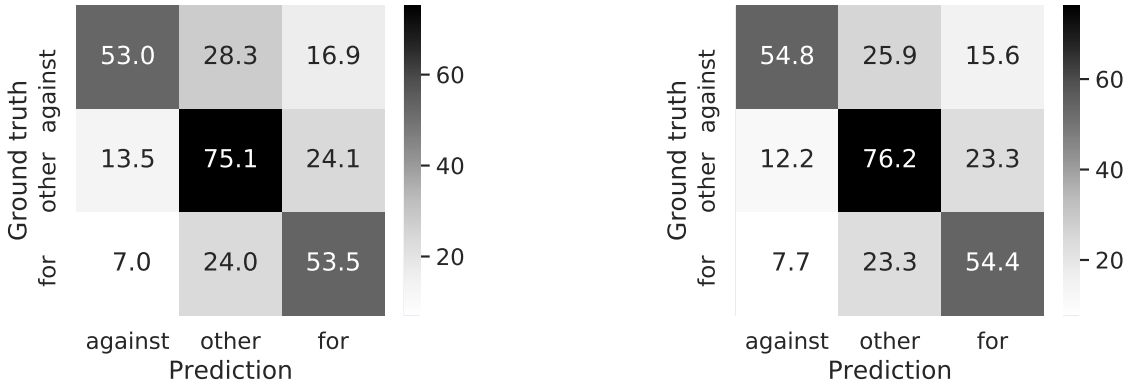
Для оценки используется метрика, предложенная в конкурсе RuArg-2022: макро-усреднённый F1-скор, исключающий метку *Irrelevant*. Для точного сравнения подходов, использующих риторическую структуру и без её учета, мы применяем 5-кратную кросс-валидацию на размеченном обучающем наборе данных. Для получения финальных предсказаний на официальных тестовом и валидационном наборах используется усреднение предсказаний пяти моделей, обученных на разных разбиениях данных в рамках кросс-валидации. Это аналогично ансамблю, где каждая модель обучена на 80% обучающих данных.

Результаты

В Таблице 26 представлены результаты оценки качества базовой модели и модели, использующей риторический анализ.

Таблица 26 — 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**



(а) BERT (б) BERT + RST-LSTM

Рисунок 5.3 — Усредненные матрицы ошибок на кросс-валидации.

Классификация позиций: Учёт дискурсивной структуры позволил улучшить классификацию во всех представленных темах. Относительно базового BERT-классификатора среднее значение F1 улучшено на 1,2% для классификации текстов, в которых обнаружена дискурсивная структура, и на 0,9% во всем тестовом наборе данных. На рисунке 5.3 показаны усредненные матрицы ошибок при определении позиции автора текста в текстах, обладающих дискурсивной структурой (все темы, без метки класса *Irrelevant*). Второй шаг классификации RST-LSTM позволяет лучше различать полярные позиции *For* и *Against*. Для базового классификатора метка *Other* оказалась наиболее сложной, так как она может обозначать

как отсутствие позиции, так и наличие примеров высказываний позиции обоих полярных классов; эта метка также является наиболее частой.

Классификация аргументов: В среднем качество классификации улучшилось на 0,5% как для текстов с распознаваемой риторической структурой, так и для всего тестового набора. Улучшение наиболее заметно в темах «маски» (+1,6% F1) и «карантин» (+2,0% F1), в то время как в теме «вакцинация» второй шаг ухудшил качество классификации (-1,3% F1). Это может указывать на недостатки подхода, при котором одна структура используется для предсказания двух типов меток классов одновременно. Также стоит отметить, что в этой теме наблюдаются самая большая вариация оценок F1 как в базовой, так и в структурной классификации.

Таблица 27 — Исследование влияния типов риторических отношений: F1, в %.

Дискурсивные Отношения	Позиция автора	Аргументация
Все	60,1	62,0
- Связность	- 0,1	- 0,1
- Контраст	- 0,1	- 0,1
- Причина	- 0,0	- 0,1
- Условие	- 0,1	- 0,1
- Атрибуция	- 0,1	- 0,0
Только структурные	- 0,2	- 0,6

Исследование влияния признаков риторической структуры: В Таблице 27 представлены результаты исследования влияния меток риторических отношений и ядерности на результаты классификации. Было обнаружено, что исключение отдельных типов отношений незначительно влияет на результаты классификации на дискурсивных деревьях. Однако замена всех дискурсивных отношений на многоядерные структурные

отношения приводит к снижению F1 на 0,2% в задаче классификации позиции и на 0,6% в классификации аргументов, что подчеркивает важность признаков, связанных с классами риторических отношений, в извлечении аргументов.

5.3. Разрешение кореференции с использованием риторических признаков

Задача разрешения кореференции заключается в нахождении и группировке упоминаний, ссылающихся на один объект реального мира. Формально, пусть D — документ, содержащий множество упоминаний сущностей $\mathcal{M} = m_1, m_2, \ldots, m_N$. Необходимо разбить множество $\mathcal{M}$ на кластеры $\mathcal{C} = C_1, C_2, \ldots, C_K$ таким образом, что все упоминания $m_i, m_j \in C_k$ относятся к одной и той же сущности.

Разрешение кореференции является одной из наиболее сложных задач обработки естественного языка, так как требует учета как лингвистических особенностей, так и общих знаний о мире. При разработке методов глубокого обучения для решения поставленной задачи исследователи преимущественно стараются усовершенствовать векторные представления лексики сущностей и оценку вероятности референциальной связи между ними [168–173]. Однако существующие методы не позволяют в явном виде учитывать референциальный выбор в иерархическом дискурсе, полагаясь только на поверхностные текстовые признаки, такие как лексика упоминаний и количество слов между ними. Для извлечения неявных признаков используются глубокие языковые модели, которые, при достоинствах моделирования семантико-синтаксических связей внутри предложения, способны ограниченно оценивать иерархию дискурса на уровне локальных маркеров. В данном исследовании предлагается метод, который позволяет явно учи-

тывать феномен референциального выбора в иерархическом дискурсе на основе дискурсивного анализа в рамках теории риторических структур. Метод использует преимущества классической нейронной архитектуры модели разрешения кореференции и улучшает её за счёт включения признаков, отражающих иерархию риторической структуры.

В этом разделе описан предложенный метод разрешения кореференции, в котором используются не только кодирование последовательности токенов в предобученной языковой модели, но и риторические признаки: линейное расстояние, риторическое расстояние и расстояние от упоминания сущности до её наименьшего общего предка с возможным антецедентом в риторическом дереве. Метод разрешения кореференции оценивается на материалах длинных текстов, исследуется применимость разработанных поверхностного и полнотекстовых риторических анализаторов для русского языка.

5.3.1. Метод разрешения кореференции на основе анализа риторической структуры

Задача разрешения кореференции формально заключается в следующем. Для документа D, содержащего множество упоминаний $\mathcal{M}$, необходимо определить функцию $\varphi : \mathcal{M} \to \mathcal{C}$, которая отображает каждое упоминание m_i в соответствующий кластер C_k.

В основе метода разрешения кореференции лежит классический подход, включающий выделение возможных сущностей в неразмеченном тексте и двухэтапное ранжирование возможных анафорических отношений между ними. Алгоритм разрешения кореференции в предложенном методе можно описать следующим образом:

1. Закодировать текст D пословно. Векторное представление $\mathbf{v}_t$ слова t вычислить как среднее векторов подтокенов[1] $\{\mathbf{v}_{t_1}, \mathbf{v}_{t_2}, \ldots, \mathbf{v}_{t_n}\}$, полученных из предобученной языковой модели: $\mathbf{v}_t = \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_{t_i}$.
2. Выделить все возможные подстроки длины не более L слов в документе D. Обозначим множество таких подстрок как $\mathcal{S}(D) = \{s \mid s \subset D, |s| \leqslant L\}$.
3. Для каждой подстроки $s \in \mathcal{S}(D)$ оценить вероятность того, что она является самостоятельным упоминанием сущности $p_{\text{span}}(s)$. Векторное представление фразы v_s вычисляется как взвешенная сумма векторных представлений составляющих ее слов $\mathbf{v}_s = \sum_{t_i \in s} \alpha_i \cdot \mathbf{v}_{t_i}$, где веса α_i определяются через механизм внимания:

$$\alpha_i = \frac{\exp(\mathbf{w}_{\text{att}}^{\top}\mathbf{v}_{t_i} + b_{\text{att}})}{\sum_{t_j \in s}\exp(\mathbf{w}_{\text{att}}^{\top}\mathbf{v}_{t_j} + b_{\text{att}})}, \tag{5.10}$$

   где $\mathbf{w}_{\text{att}}$ и b_{att} — обучаемые параметры механизма внимания. Оценка вероятности $p_{\text{span}}(s)$ вычисляется в полносвязном слое:

$$p_{\text{span}}(s) = \sigma(\mathbf{w}_{\text{span}}^{\top}\mathbf{v}_s + b_{\text{span}}), \tag{5.11}$$

   где σ — функция активации, $\mathbf{w}_{\text{span}}$ и b_{span} — обучаемые параметры. Фразы ранжируются по оценкам вероятности $p_{\text{span}}(s)$, выбираются k_{span} наиболее вероятных упоминаний: $\mathcal{S}_{\text{selected}} \subset \mathcal{S}(D)$.
4. Для всех пар выбранных упоминаний (s_i, s_j), где $s_i, s_j \in \mathcal{S}_{\text{selected}}$ и $i < j$, грубо оценить вероятность наличия референциальной связи $p_{\text{ref}}(s_i, s_j)$. Для этого суммировать вероятности того, что s_i и s_j являются упоминаниями, и билинейное преобразование их векторных

[1] Глубокие языковые модели обладают словарём ограниченной длины, в котором одно слово может быть разбито на несколько фрагментов — подтокенов (англ. subtoken). Словарь заданного объёма определяется из корпуса предобучения модели с помощью алгоритма сжатия кодированием пар байтов (Byte Pair Encoding) и его модификаций. Для мультиязычных моделей, таких, как используемая в настоящем исследовании, наиболее характерно фрагментарное кодирование слов, поскольку в их словарях содержится ограниченное количество токенов для кодирования каждого из множества языков, в том числе символы алфавита и наиболее частотные для текстов данного языка последовательности символов.

представлений:

$$p_{\text{ref}}(s_i, s_j) = p_{\text{span}}(s_i) + p_{\text{span}}(s_j) + \mathbf{v}_{s_i}^{\top} W_{\text{bilin}} \mathbf{v}_{s_j}, \tag{5.12}$$

где W_{bilin} — обучаемая матрица весов билинейного преобразования. Пары фраз ранжировать по полученным оценкам наличия связи, выбрать k_{ref} наиболее вероятных связей $\mathcal{R}_{\text{selected}}$.

5. На этом этапе вычисляются окончательные оценки вероятностей референциальных связей для каждой пары упоминаний $(s_i, s_j) \in \mathcal{R}_{\text{selected}}$. Векторное представление пары $\mathbf{v}_{s_i,s_j}$ формируется путём конкатенации векторных представлений и дополнительных признаковых векторов:

$$\mathbf{v}_{s_i,s_j} = [\mathbf{v}_{s_i}; \mathbf{v}_{s_j}; \varphi(s_i, s_j)], \tag{5.13}$$

где $\varphi(s_i, s_j)$ — вектор признаков, характеризующих связь между упоминаниями s_i и s_j, включая расстояние между ними в токенах, риторическое расстояние и другие дискурсивные признаки (см. ниже). Окончательная оценка вероятности вычисляется в полносвязном слое:

$$P_{\text{final}}(s_i, s_j) = \sigma(\mathbf{W}_{\text{final}}^{\top} \mathbf{v}_{s_i,s_j} + b_{\text{final}}). \tag{5.14}$$

6. Для каждого упоминания s_j выбирается антецедент s_i с наибольшей оценкой $P_{\text{final}}(s_i, s_j)$. Цепочка упоминаний, соединённых референциальными связями, образует кластер C_k.

Дискурсивные признаки

Для повышения способности модели анализировать референциальные отношения предлагаются дискурсивные признаки, основанные на риторической структуре текста.

Для каждой пары упоминаний (s_i, s_j) определяются элементарные дискурсивные единицы (ЭДЕ) u_i и u_j, в которых они находятся в риторическом дереве документа. Затем вычисляются метрики, описывающие референциальное расстояние в структуре дискурса [174]:

– Линейное расстояние (D_{Lin}) соответствует количеству ЭДЕ, находящихся между упоминаниями s_i и s_j в тексте.
– Риторическое расстояние (D_{Rh}) определяется количеством ядерных ЭДЕ, встречающихся между двумя упоминаниями в иерархическом дереве риторической структуры. Учитывает активированность дискурсивных единиц во внимании автора текста — важный аспект референциального выбора — через положения ядер в риторических связях.

Также используется признак, оценивающий степень общности дискурса, включающего оба упоминания [175]:

– Расстояние референта до наименьшего общего предка (LCA) (D_{LCA}) вычисляется из утверждения, что упоминание s_j всегда правее возможного антецедента s_i в тексте. $\text{LCA}(u_j, u_i)$ — наименьшая дискурсивная единица риторического дерева, содержащая оба упоминания. Расстояние D_{LCA} вычисляется как разница в глубине между u_j и $\text{LCA}(u_j, u_i)$ в риторическом дереве.

Особенности предобученной языковой модели

Для векторизации упоминаний сущностей предлагается использовать языковую модель архитектуры LUKE [176]. На этапе предобучения эта модель обучается не только задаче маскированного языкового моделирования (MLM), характерной для кодирующих языковых моделей, но и задаче восстановления маскированных сущностей (MEP). Разметка сущностей

в обучающем корпусе используемой реализации `studio-ousia/mluke-large-lite` собрана из гиперссылок в многоязычных, включая русский язык, выгрузках Википедии.

Факторы, снижающие использование памяти

Решение задачи разрешения кореференции с использованием глубоких нейронных сетей предъявляет высокие требования к объёму памяти. Классические подходы [177;178], на которых основан предложенный метод, требуют кодирования всех возможных фраз до заданной длины (характерной для упоминаний сущностей в данном языке) в документе. В последние годы были предложены альтернативные методы, предполагающие обработку отдельных токенов вместо фраз [172; 179; 180]. Однако даже такие методы, полагаясь на дообучение больших языковых моделей, могут требовать от 40 до 80 ГБ видеопамяти [172;181].

В ходе разработки предлагаемого метода были исследованы возможности уменьшения требований к объёму памяти для классических подходов к разрешению кореференции, основанных на кодировании всевозможных фраз. Каждая из предложенных моделей обучалась с использованием до 32 ГБ видеопамяти. Мы оптимизировали стандартную архитектуру модели и её реализацию следующим образом:

– Ключевым фактором, позволившим обучать модель кореференции на большом наборе данных с ограниченной памятью, стало исключение полного дообучения языковой модели. В наших экспериментах веса языковой модели были заморожены за исключением последних k слоёв, где значение k подбиралось эмпирически в зависимости от доступного объёма видеопамяти. В описываемых экспериментах $k = 8$ из 23 слоёв языковой модели дообучались.

– После начального кодирования токенов для уменьшения размерности их представлений использовалась двунаправленная LSTM. В наших экспериментах размерности векторных представлений были уменьшены с $\mathbf{e}_{\text{LM}} \in \mathbb{R}^{1024}$ до $\mathbf{e}_{\text{LSTM}} \in \mathbb{R}^{100}$.
– Каждый абзац текста кодировался языковой моделью отдельно. Это позволило обрабатывать длинные документы без обрезки текста и одновременно сжимать высокоразмерные, частично обучаемые представления текста.

Кроме того, были применены стандартные техники уменьшения требований к объёму памяти:

– Мы устанавливали размер батча равным 1. Аккумуляция градиентов по нескольким батчам не улучшила результаты обучения.
– Все вычисления проводились с использованием смешанной точности.

5.3.2. Экспериментальное исследование метода в задачах разрешения кореференции

В данном разделе описано экспериментальное исследование метода разрешения кореференции на основе анализа риторической структуры текста. Цель исследований — оценить влияние различных риторических признаков признаков, полученных с помощью различных типов риторических анализаторов, на качество разрешения кореференции в русскоязычных текстах.

В предварительном эксперименте использовался поверхностный риторический анализатор (см. раздел 3), который предоставляет упрощённую модель текста в виде леса риторических деревьев. Оценено влияние при-

знаков, извлеченных при помощи поверхностного анализатора, на качество разрешения кореференции.

Во втором эксперименте применялись полнотекстовые риторические анализаторы, обученные на различных корпусах риторической разметки для русского языка (см. раздел 4). Эти анализаторы позволяют более точно и детально восстановить риторическую структуру всего документа, что обеспечивает возможность извлечения более информативных дискурсивных признаков для модели кореференции.

В обоих экспериментах основное внимание уделяется изучению того, как различные дискурсивные признаки, такие как линейное расстояние между элементарными дискурсивными единицами, риторическое расстояние и расстояние до наименьшего общего предка в риторическом дереве, влияют на эффективность разрешения кореференции.

Далее в разделе представлены подробные описания используемых данных, методологии оценки, а также результаты экспериментов и их анализ.

Детали экспериментов

В качестве предобученной кодирующей языковой модели используется `studio-ousia/mluke-large-lite`. Токенизация и разделение текста на предложения осуществляются при помощи библиотеки razdel[2]. Границы именованных сущностей определяются при помощи модели разметки последовательности BIO-тегов сущностей на основе языковой модели `ru_core_news_lg` из библиотеки SpaCy[3]. Модели реализованы при помощи библиотеки pytorch[4]. Вычисления метрик в риторическом дереве реализованы на языке Cython, что позволило ускорить нахождение линейного

[2]https://github.com/natasha/razdel
[3]https://spacy.io
[4]https://pytorch.org

расстояния D_{Lin} в среднем на 15%, D_{Rh} на 22% и D_{LCA} на 62% по сравнению с имплементацией на языке Python.

Экспериментальное исследование с использованием поверхностного риторического анализатора

Данные Предварительное исследование метода проводилось в рамках соревнования RuCoCo-2023 [182]. Организаторы соревнования предоставили закрытые валидационную и тестовую выборки. С точки зрения настройки модели под задачу, требовалось определить максимальную длину сущности в корпусе. Средняя длина сущности составляет 2, а максимальная — 42 токена. Максимальной длиной упоминания в нашей системе установлено 13 токенов. Эта длина характерна для 99,7% сущностей в корпусе. Также, поскольку используется поверхностный риторический разбор, важно оценить длину документов в абзацах. Если общее дерево документа создаётся объединением деревьев абзацев, этот показатель важен для оценки долгих дискурсивных зависимостей. Медианное количество параграфов — 9, а максимальное количество разделённых строк составляет 162. Некоторые новостные статьи необычайно длинные, а некоторые из них содержат списки. Объединение леса риторических структур в одно дерево может повлиять на оценку референтных дистанций в длинных текстах.

Метод оценки Метрикой соревнования является Link-based Entity Aware (LEA) [183]. В этой метрике вес каждой сущности определяется её размером, причём более крупные сущности считаются более важными. LEA также оценивает разрешённые отношения кореференции, а не только отдельные упоминания. Валидация моделей проводится во время обучения на 5% официального тренировочного набора данных. Мы выполняем слу-

чайное разбиение данных 4 раза и представляем усреднённый результат. Итоговые результаты на официальных валидационной и тестовой выборках соревнования получены теми же моделями.

В таблице 28 представлены результаты экспериментов на валидационной выборке соревнования RuCoCo-2023. В таблице 29 приведены результаты оценки на тестовой выборке соревнования.

Таблица 28 — Оценка моделей на валидационной выборке соревнования.

	Precision	Recall	F1	Top-1 F1 (leaderboard)
Baseline	78,7 ± 0,7	69,1 ± 0,7	73,5 ± 0,5	74,3
$+D_{Lin}$	78,6 ± 1,8	68,3 ± 2,2	73,0 ± 0,5	74,0
$+D_{Rh}$	78,5 ± 1,5	69,3 ± 1,0	**73,6 ± 0,9**	**74,6**
$+D_{LCA}$	75,0 ± 0,8	70,9 ± 1,0	72,9 ± 0,4	73,5

Из-за строгого ограничения на количество решений в финальной стадии соревнования мы смогли оценить только две лучшие модели, Baseline и Baseline+D_{Rh}, на закрытой тестовой выборке.

Таблица 29 — Оценка моделей на тестовой выборке соревнования.

	Precision	Recall	F1	Top-1 F1 (leaderboard)
Baseline	79,1 ± 0,8	66,9 ± 0,6	**72,5 ± 0,3**	72,8
$+D_{Rh}$	79,3 ± 1,6	66,6 ± 1,9	72,4 ± 0,5	**73,3**

Признаки D_{Lin} и D_{LCA} оказались неэффективными для решения задачи разрешения кореференции на валидационной выборке (таблица 28). D_{Lin} — линейное расстояние в ЭДЕ — не добавляет существенно новую информацию по сравнению с линейным расстоянием в токенах, которое уже используется в модели. В свою очередь, D_{LCA} — расстояние от упоминания до наименьшего общего предка — может оцениваться с недостаточной точностью в условиях поверхностного риторического разбора, когда из

предсказанного риторического леса искусственно формируется единое скошенное вправо дерево. В этом случае глубина правой ветви больше зависит от порядка изложения слабо связанных дискурсивных единиц, чем от фактической структуры текста.

Средние результаты оценки качества модели, использующей признак риторического расстояния D_{Rh}, незначительно отличаются от результатов базовой модели на обоих наборах данных. Однако её оценки имеют больший разброс, что позволяет модели с наибольшим значением оценки F1 попасть на вершины рейтингов соревнования. Из этого можно предположить, что риторическое расстояние является наиболее устойчивым признаком из рассмотренных, несмотря на недостатки частичного риторического разбора.

Экспериментальное исследование с использованием полнотекстовых риторических анализаторов

Данные В экспериментах используется два больших корпуса референциальной разметки: AnCor [184] и RuCoCo [182]. **Корпус AnCor** предназначен для исследования автоматического разрешения анафоры и кореференции в текстах на русском языке. В корпусе содержится 523 текста таких жанров, как новости (∼50%), блоги (∼24%), юридические тексты (∼11%) и другие. Среднее число токенов в документе: 225. Размечено 25159 упоминаний, организованных в 5678 референциальных цепочек. В рамках описанных в этой работе экспериментов для валидации моделей случайно выбираются 5% обучающего корпуса; разметка тестовой части корпуса из соревнования RU-EVAL-2019 общедоступна и используется для тестирования всех моделей. **RuCoCo** — самый большой из существующих корпус референциальной разметки для русского языка содержит 307 новостных текстов, в которых размечено 38877 референциальных цепочек из 150405

упоминаний. Среднее число токенов в документе: 440. В рамках описываемых экспериментов из размеченного корпуса случайно выбираются по 5% документов для валидации и тестирования моделей; значения качества усредняются по 3 разбиениям.

Методы оценки Для оценки методов разрешения кореференции в экспериментах используются классические метрики качества: MUC [185], B^3 [186], CEAF [187] и $CONLL = \frac{MUC+B^3+CEAF}{3}$ [188], а также активно используемая в актуальных исследованиях, включая работы на основе корпуса RuCoCo, метрика LEA [183]. **Метрика MUC** основана на сравнении связей между упоминаниями, если представить референциальные связи между упоминаниями в виде графа. Идеальная система связывает все упоминания одной сущности. Ошибки системы проявляются в виде лишних или пропущенных связей. Точность в MUC рассчитывается как доля корректных связей среди всех предсказанных, а полнота — как доля корректных связей из эталонной разметки, найденных системой. Несмотря на интуитивную понятность, MUC имеет недостаток: она поощряет избыточную объединённость упоминаний, например, при отнесении всех упоминаний к одному кластеру. В отличие от MUC, **метрика B^3** основывается на оценке самих упоминаний. Для каждого упоминания вычисляются точность (доля истинных кореферентов в предсказанном кластере) и полнота (доля истинных кореферентов упоминания, попавших в тот же предсказанный кластер). Средние значения точности и полноты по всем упоминаниям дают итоговую оценку. **Метрика CEAF (CEAF-e)** оценивает соответствие между предсказанными и эталонными кластерами, используя матрицу совпадений, где каждая ячейка представляет количество совпадений между парами кластеров. Благодаря методу сопоставления гипотезы и эталона, CEAF устойчива к различиям в размерах кластеров и снижает влияние крупных кластеров на общий результат, хотя и является вычислительно затратной. В **метрике LEA** оцениваются пары упоминаний в каждом предсказанном кластере

с учетом их соответствия эталонным кластерам. LEA учитывает вес каждого упоминания, пропорциональный количеству связей, в которых оно участвует, что позволяет более точно анализировать внутреннюю структуру кластеров, а не просто количество правильно определенных кластеров.

Риторические анализаторы В ходе исследования изучается применимость различных вариантов полного анализатора риторических структур для русского языка к прикладным задачам. Используются варианты аналитора на основе архитектуры, описанной в разделе 4.2, демонстрирующие наилучшее качество разбора структуры на русскоязычных корпусах. В том числе лучший анализатор, обученный на корпусе **RRT**; двуязычный анализатор, обученный на смеси оригинального англоязычного корпуса GUM v9.1 и его русскоязычной адаптации **RRG**; кросс-жанровый анализатор для русского языка **RRT + RRG**, обученный на смешении корпусов RRT, RRG и переведенных RST-DT и GUM_{v10}. Поскольку эксперименты, описанные в разделе 4, показали, что мультиязыковая модель `xlm-roberta-large` позволяет достичь лучшего качества риторического разбора на русскоязычных корпусах (RRT, RRG) по сравнению с актуальными моноязыковыми моделями, все модели в настоящем исследовании используют эту мультиязыковую модель для получения векторных представлений текстов. В таблице 30 приведены оценки качества используемых в экспериментах полнотекстовых анализаторов на тестовых выборках русскоязычных корпусов риторической разметки.

Таблица 30 — Оценки качества риторических анализаторов, F1, в %.

Модель	Обучающие данные	Тестовые данные							
		RRT				RRG			
		Seg	Span	Nuc	Full	Seg	Span	Nuc	Full
RRT	RRT	92,4	66,5	52,4	45,3	86,7	49,1	35,2	-
RRG	RRG, GUM	85,5	53,6	38,7	-	96,9	66,5	53,3	44,6
RRT + RRG	RRT, RRG, RST-DT, GUM10	91,6	66,5	53,0	45,3	95,3	64,2	52,0	42,9

Основные результаты экспериментов по разрешению кореференции с использованием различных полнотекстовых анализаторов дискурса на русском языке приведены в таблице 31, где представлены средние показатели качества моделей на двух датасетах для трех запусков обучения. На датасете AnCor с фиксированной тестовой выборкой базовый метод продемонстрировал среднее качество, сопоставимое с лучшими результатами, ранее полученными для этого корпуса (58.1% CoNLL F1 [189], использовались аугментация данных и информация о говорящем). Это свидетельствует о высокой эффективности базового метода без необходимости привлечения дополнительных данных.

Таблица 31 — Оценка качества разрешения кореференции; F1, в %.

		AnCor					RuCoCo				
		MUC	B3	CEAFe	CoNLL	LEA	MUC	B3	CEAFe	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+RRG	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**

Признаки расстояний между сущностями в дереве риторического разбора значительно улучшили качество разрешения кореференции на материалах корпусов AnCor и RuCoCo. Особенно заметное повышение качества наблюдается на крупнейшем датасете RuCoCo, где использование каждого парсера и каждого типа признаков привело к улучшению качества разрешения кореференции. Этот результат показывает, что модель, обученная на больших объемах данных, способна эффективно выявлять закономерности в последовательностях токенов, а признаки, извлеченные

из иерархической структуры дискурса, способствуют дополнительному повышению точности модели.

На менее объемных данных корпуса AnCor лучшие результаты достигнуты при учете признаков, полученных при помощи анализатора, основанного на корпусе RRT. Этот анализатор обучен разметке риторических структур, характерных для корпуса AnCor: новостных статей и блогов небольшого размера. В отличие от AnCor, корпус RuCoCo содержит более развернутые тексты, риторическая структура которых была успешно обработана анализаторами, знакомыми с высокоуровневыми структурами из корпуса RRG.

Примечательно, что учет признака риторического расстояния D_{Rh} способствовал улучшению качества разрешения кореференции на большинстве датасетов и для большинства парсеров. Это подчеркивает важность учета структурных особенностей дискурса при решении задач кореференции, а также устойчивость данного признака к различиям в подходах к риторическому разбору текстов.

Наилучшее качество разрешения кореференции на корпусе RuCoCo достигнуто при учете признака расстояния от правого упоминания до наименьшего общего предка D_{LCA}: +1,1% LEA F1 по сравнению с базовым методом при разборе текстов жанрово-универсальной дискурсивной моделью RRT+RRG. Этот же признак позволил достичь наилучшего качества и на корпусе AnCor: +0,5% LEA F1 по сравнению с базовым методом при использовании анализатора компактных структур RRT. В то же время добавление признака линейного расстояния между ЭДЕ D_{Lin} не привело к значительному улучшению качества ни в одном из экспериментов, поскольку этот признак, как и в экспериментах с поверхностным анализатором, не добавляет полезной информации к уже имеющейся в модели.

5.4. Использование анализа риторических структур в построении структуры аргументации в тексте-рассуждении

Автоматический анализ аргументации является одной из ключевых задач в области анализа естественного языка. Анализ аргументации включает в себя выявление посылок, утверждений и выводов в структуре аргументации текста-рассуждения. Несмотря на то, что анализ аргументации и риторический анализ имеют разные цели, многие исследователи отмечают связь между этими двумя областями. Ряд работ [42; 43; 157; 190–192] посвящен поиску корреляций между аргументативными и риторическими структурами в тексте. В этих работах рассматривается единственный результат автоматического риторического анализа или единственная ручная разметка структуры каждого текста. Однако одна и та же аргументативная структура может быть найдена в текстах похожих рассуждений с различными риторическими структурами, в особенности если риторическое дерево извлекают автоматически. Это необходимо учитывать при сопоставлении структур дискурса и аргументации.

Анализ риторических структур не является задачей с однозначным решением и граничит с прагматическим анализом естественного языка. Согласно [92], оценка качества ручной разметки экспертом-лингвистом новостей из эталонного корпуса RST-DT составляет около 55,0% Full F1 для фиксированной сегментации ЭДЕ. Теория риторических структур не предполагает единственно верной интерпретации структуры дискурса. Несовершенства автоматических анализаторов усиливают различия в риторических структурах схожих документов. Объективная интерпретация дискурса может требовать сложных навыков, таких как рассуждение на основе общего знания или оценка относительной значимости отдельных утверждений. Необходимо отметить, что на качество риторического анализа также значительно влияет жанр и домен. Для изолированной подзадачи

классификации риторических отношений в парах смежных ЭДЕ в новостях, академических текстах, выступлениях TED, постах и комментариях Reddit и художественной литературе, размеченных в корпусе GUM RST [77]), сообщается о средней ошибке анализа при кросс-жанровом переносе в 60% [193]. В другом исследовании демонстрируется значительное снижение качества построения неразмеченных риторических структур при переносе анализатора с обучающей выборки корпуса RST-DT на тестовую из мультижанрового корпуса GUM [137]. Оценка качества снижается в среднем на $\sim$ 11 Span F1 и на $\sim$ 16 Nuc F1. Проведенные нами исследования мультижанрового переноса в русском языке, описанные в разделе 4.5.4, демонстрируют схожую деградацию качества анализа при переносе риторического анализатора на жанры, отсутствовавшие в обучающей выборке.

Существующие исследования по анализу совместимости структур дискурса и аргументации опираются на единственную риторическую аннотацию для каждого текста. В данном исследовании показано, что такой подход может снижать объективность анализа соотношений между структурами. Одна и та же аргументативная структура может быть выражена через различные риторические структуры. Различие между структурами диктуется как естественной вариацией выражения мыслей автором, так и неоднозначностью и несовершенствами риторического анализа.

В данном исследовании предлагается новый подход к исследованию взаимосвязи между структурами дискурса и аргументации, предполагающий использование множества вариантов риторической структуры для одной структуры аргументации. Для решения этой задачи разработан метод анализа структуры аргументации на основе глубокого обучения — биафинный анализатор структуры на основе дискурса (DBAP).

В исследовании используется корпус Argumentative Microtexts [194]. В этом корпусе текст-рассуждение рассматривается как гипотетический диалектический обмен между автором, который представляет и защищает свое утверждение, и его оппонентом. Аргументация может быть пред-

ставлена графом с узлами, соответствующими отдельным утверждениям (аргументационные дискурсивные единицы, АДЕ), и ребрами, показывающими отношения поддержки либо атаки между АДЕ.

В разделе 5.4.1 предлагаются методы анализа аргументации как анализа зависимостей между АДЕ, в том числе на основе нескольких вариантов риторической структуры. В разделе 5.4.2 описаны экспериментальные исследования предложенных методов; приведены результаты анализа взаимосвязи между структурами дискурса и аргументации; приведены первые результаты построения структур аргументации для версии корпуса Argumentative Microtexts для русского языка.

5.4.1. Метод анализа структуры аргументации с использованием вариантов риторической структуры

В целях анализа аргументации в коротких текстах-рассуждениях предлагается подход, основанный на классическом методе *Evidence Graphs* (EG) [195]. Данный метод предполагает преобразование графа аргументации текста в дерево зависимостей. Однако, в отличие от метода EG, где разметка аргументативной структуры достигается посредством сложного взаимодействия между отдельными классификаторами структуры, функции, роли и центрального утверждения, наш подход основан на прямом предсказании зависимостей между текстовыми фрагментами в единой нейронной модели. Роли аргументативных дискурсных единиц и центральное утверждение автоматически выводятся из дерева зависимостей с помощью простых правил.

Биаффинный анализатор аргументации

Анализ аргументации формулируется как задача построения дерева зависимостей. Терминальными узлами в данном дереве могут выступать либо заранее выделенные АДЕ, либо элементарные дискурсные единицы, предсказанные дискурсивным анализатором. В последнем случае вводится дополнительная структурная функция для объединения нескольких ЭДЕ в одну АДЕ.

Алгоритм анализа структуры аргументации в тексте, заданном последовательностью предварительно сегментированных дискурсивных единиц, можно описать следующим образом:

1. Из последовательности из n дискурсивных единиц $u_1, u_2, \ldots, u_n$ (элементарных или аргументативных) каждая ДЕ кодируется при помощью CLS-пулинга в предобученной ЯМ в вектор $\mathbf{v}_i \in \mathbb{R}^{d_{LM}}$:

$$\mathbf{v}_i = \text{Encoder}(w_1 w_2 \ldots w_k), \tag{5.15}$$

   где $w_1 w_2 \ldots w_k$ — токены соответствующей ДЕ.

2. Применяется биаффинная модель анализа зависимостей [196]. В рассматриваемой модели структуры аргументации метки дуг представляют собой аргументативные функции: «поддержка» (support) или «атака» (attack). Центральное утверждение кодируется дополнительной функцией «cc» (central claim). Это единственная функция, которая может быть назначена узлу без родителя (корневому узлу).
   а) Дополнительный узел `root`, который является фиктивным родителем реального корня дерева, случайным образом кодируется в вектор $\mathbf{v}_0$.
   б) Матрица признаков $\mathbf{V} \in \mathbb{R}^{(n+1)\times d_{LM}} = [\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_n]$ пропускается через четыре полносвязных слоя для получения

скрытых представлений дуг и функций для родительского и зависимого узлов:

$$\begin{aligned} H^{(\text{arc-parent})} &= \text{FF}^{(\text{arc-parent})}(\mathbf{V}), \\ H^{(\text{arc-dep})} &= \text{FF}^{(\text{arc-dep})}(\mathbf{V}), \\ H^{(\text{func-parent})} &= \text{FF}^{(\text{func-parent})}(\mathbf{V}), \\ H^{(\text{func-dep})} &= \text{FF}^{(\text{func-dep})}(\mathbf{V}). \end{aligned} \tag{5.16}$$

в) Скрытые представления узлов для расчёта оценок дуг используются для оценки вероятности каждого возможного назначения родительского узла для каждого потомка с помощью билинейного внимания:

$$s_{i,j}^{(\text{arc})} = H_i^{(\text{arc-parent})}\, \mathbf{U}^{(\text{arc})} \left(H_j^{(\text{arc-dep})}\right)^\top + \mathbf{b}^{(\text{arc})}, \tag{5.17}$$

где $\mathbf{U}^{(\text{arc})}$ и $\mathbf{b}^{(\text{arc})}$ — обучаемые параметры модели.

г) После выбора наиболее вероятного родителя i для узла $j \neq 0$ выбираются соответствующие скрытые представления $H_i^{(\text{func-parent})}$ и $H_j^{(\text{func-dep})}$. Затем с помощью билинейной функции оцениваются вероятности аргументативных функций:

$$s_{i,j}^{(\text{func})} = H_i^{(\text{func-parent})}\, \mathbf{U}^{(\text{func})} \left(H_j^{(\text{func-dep})}\right)^\top + \mathbf{b}^{(\text{func})}, \tag{5.18}$$

$$P(\text{function}_{i,j}) = \text{softmax}\Big(s_{i,j}^{(\text{func})}\Big), \tag{5.19}$$

где $\mathbf{U}^{(\text{func})}$ и $\mathbf{b}^{(\text{func})}$ — обучаемые параметры.

3. Поскольку аргументативная роль высказывания напрямую связана с его функцией по отношению к родительскому узлу, она не предсказывается моделью непосредственно, а выводится из предсказанных зависимостей. Так как центральное утверждение по определению является утверждением пропонента, производится обход предсказанного дерева зависимостей с назначенными функциями, и для узла j с родителем i роль (“pro” = “$\overline{\text{“opp”}}$) определяется

следующим образом:

$$\text{role}_j = \begin{cases} \text{“pro”}, & \text{если function}_{j,0} = \text{“cc”}; \\ \overline{\text{role}_i}, & \text{если function}_{i,j} = \text{“attack”}; \\ \text{role}_i, & \text{в противном случае.} \end{cases} \tag{5.20}$$

С целью оценки эффективности риторического анализа в анализе структуры аргументации в данном исследовании предложено два варианта метода. Описанный алгоритм анализа зависимостей соответствует **Биаффинному анализатору аргументации (BAP)**. Его модификация **Биаффинный анализатор на основе анализа дискурса (DBAP)** дополнительно учитывает дискурсивные отношения в риторическом дереве.

Биаффинный анализатор аргументации на основе анализа дискурса (DBAP)

Для интеграции риторических структур в анализ аргументации оценки зависимостей (5.17) умножаются на дискурсивные коэффициенты $\boldsymbol{C}^{(\text{RST})}$ для соответствующих отношений из риторического дерева:

$$S^{(\text{arc})} = S^{(\text{arc})} \circ \boldsymbol{C}^{(\text{RST})}, \tag{5.21}$$

где $\circ$ обозначает поэлементное умножение. Дерево составляющих риторической структуры преобразуется в дерево зависимостей в соответствии с положением ядра в каждом отношении.

В простейшем случае дискурсивные коэффициенты предсказываются из бинарной матрицы смежности $A^{(\text{RST-adj})}$ размера $n \times n$, где $a_{i,j}^{(\text{RST-adj})} = 1$, если существует дискурсивная связь от дискурсивной единицы i к ядерной[5] дискурсивной единице j:

[5] В случае нескольких ядер — к левому ядру.

$$\boldsymbol{C}^{(\text{RST})} = \theta A^{(\text{RST-adj})} + \mathbf{b}^{(\text{RST})}, \tag{5.22}$$

где θ и $\mathbf{b}^{(\text{RST})}$ — обучаемые скалярные параметры, контролирующие влияние дискурсивных связей на оценки дуг.

При обучении дискурсивных коэффициентов также необходимо учитывать тип риторической связи между ДЕ, так как некоторые связи могут сигнализировать о наличии аргументации или о ее отсутствии. Для этого тип риторического отношения каждой возможной дуги кодируется вещественным значением в дополнительном обучаемом слое. Риторическое дерево зависимостей с указанием типов отношений представляется в виде трехмерной матрицы смежности $A^{(\text{RST-full})}$ размера $n \times n \times k$, где $a_{i,j}^{(\text{RST-full})}$ — разреженное представление риторической связи от дискурсивной единицы i к ядру j. Далее вычисляются дискурсивные коэффициенты:

$$\boldsymbol{c}_{i,j}^{(\text{RST})} = \text{FF}^{(\text{rst-arc})}\Big(a_{i,j}^{(\text{RST-full})}\Big) = \varphi\Big((a_{i,j}^{(\text{RST-full})})^{\top}\boldsymbol{\Theta} + \mathbf{b}^{(\text{RST})}\Big), \tag{5.23}$$

где $\boldsymbol{\Theta}$ содержит обучаемые веса каждого типа риторических связей, а φ — функция активации ReLU, гарантирующая $\boldsymbol{c}_{i,j}^{(\text{RST})} \geqslant 0$.

Важно учитывать, что для некоторых типов риторической связи направление, определяемое ядерностью, может противоречить направлению аргументации. Поэтому в модели также рассматриваются инвертированные риторические связи:

$$\boldsymbol{c}_{i,j}^{(\text{RST})} = \text{FF}^{(\text{rst-arc})}\Big(a_{i,j}^{(\text{RST-full})}\Big) + \text{FF}^{(\text{rst-inv})}\Big(a_{j,i}^{(\text{RST-full})}\Big). \tag{5.24}$$

Включение инвертированных риторических отношений позволяет не только штрафовать предсказания, противоречащие аргументативным риторическим связям, но и поощрять инверсию дискурсивных связей, которые по определению противоречат структуре аргументации (например, Preparation).

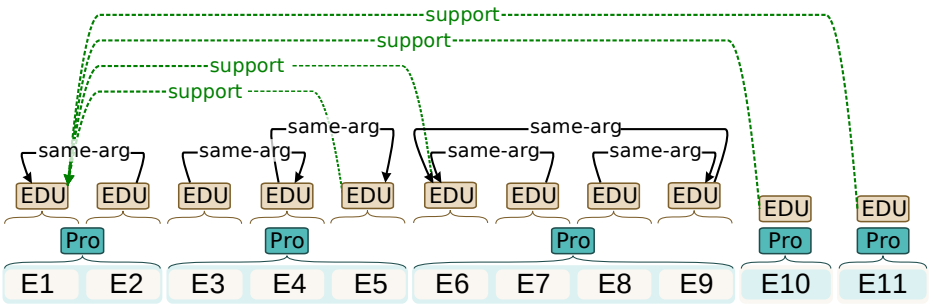


Рисунок 5.4 — Представление структуры аргументации в анализаторе с автоматической сегментацией, `micro_k002:En`. См. примеры на рисунках 5.5 и 5.7a.

Неоднозначность интерпретации дискурса приводит к высокой вариативности разметки риторических структур. В автоматическом анализе вариативность объясняется как непоследовательностью в экспертной разметке, так и внутренними ограничениями статистических моделей в понимании языка. Для получения вариантов дискурсивной структуры текстов-рассуждений с одинаковой структурой аргументации предлагается использовать перефразирования текстов на уровне АДЕ. В этом случае модель обучается на нескольких вариантах текстового и риторического изложения одной аргументационной структуры, что позволяет более объективно оценивать коэффициенты типов риторических отношений.

Предложенный подход позволяет более точно моделировать аргументативную структуру текста-рассуждения за счет использования информации о риторических структурах, что подтверждается результатами экспериментов.

Сегментация

В анализе структуры аргументации «с нуля» предлагается рассматривать как листья дерева аргументации элементарные дискурсивные единицы. Всякий раз, когда АДЕ соответствует риторическому поддереву из несколь-

ких ЭДЕ, мы сохраняем структуру дискурсивных отношений, присваивая каждому риторичскому отношению внутри АДЕ фиктивную функцию аргументации “same-arg” (рисунок 5.4). Добавление третьего класса функций не влияет на архитектуру модели.

5.4.2. Экспериментальное исследование метода в задаче анализа структуры аргументации

Продемонстрируем применение метода на примере простейшей структуры аргументации из корпуса Microtexts (рисунок 5.5), представленной в виде текста и его различных перефразировок в таблице 32. На первом шаге исходный текст и его переводы анализируются риторическим анализатором, в результате чего получаются различные варианты риторических структур (рисунок 5.6). Эти структуры используются для улучшения предсказаний аргументативных зависимостей в модели DBAP, учитывающей риторические связи между дискурсивными единицами.

Сбор и анализ данных

На риторическую структуру текста оказывают заметное влияние выбор стратегии перевода [132]. Для перефразирования аргументационных высказываний в текстах мы применяем метод обратного перевода на параллельном корпусе разметки аргументации.

В русскоязычной версии корпуса Microtexts АДЕ в обеих частях исходного корпуса были вручную переведены с английского на русский язык [197]. Это литературный перевод, который часто не соответствует оригиналу по количеству предложений и клауз в рамках АДЕ. Такие перефразирования

Таблица 32 — Пример перефразирования текста по АДЕ, `micro_k002`.

Данные	#	Текст
En	1	Actually, it would be justified if all German universities charged tuition fees.
	2	As long as it is ensured that the funds really benefit the universities directly, one can continue to regard this as social justice.
	3	Those who study later decide this early on, anyway.
	4	It's always possible to take out a student loan or to earn a scholarship.
	5	To oblige non-academics to finance others' degrees through taxes, however, is not just.
Ru→En	1	In fact, it would be justified if all German universities charged tuition fees.
	2	As long as it is guaranteed that the funds really benefit the universities directly, we can continue to regard it as social justice.
	3	In any case, the question of further training must be decided in advance.
	4	You can always take a student loan or get a scholarship.
	5	However, it is unfair to oblige people who do not belong to scientific circles to pay for someone else's education by collecting additional taxes.

вводят заметные различия в риторическую структуру по сравнению с оригиналом и вызывают существенные изменения в структуру дискурса.

Чтобы получить варианты пересказов для одних и тех же аргументативных структур, был выполнен машинный перевод с сохранением исходных границ АДЕ: *Экспертный Английский* → *Русский* и *Экспертный Русский* → *Английский*. Для обоих направлений перевода использовалась актуальная многоязычная модель NLLB `nllb-200-distilled-1.3B` [147]. Оценка качества автоматического перевода корпуса Microtexts относительно текстов, написанных носителями языка, достигла 31,6% BLEU с английского на русский язык и 29,2% BLEU в обратном направлении.

Таблица 32 иллюстрирует пример полученной перефразировки для простой аргументативной структуры. На рисунке 5.5 показано, что АДЕ #2–5 независимо поддерживают центральное утверждение (#1). Полуавтоматический обратный перевод помогает перефразировать отдельные высказывания внутри аргумента слегка (АДЕ #1, #2, #4) или значительно (АДЕ #3, #5).

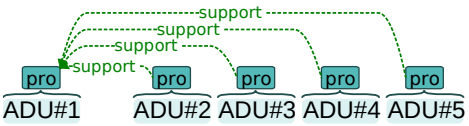
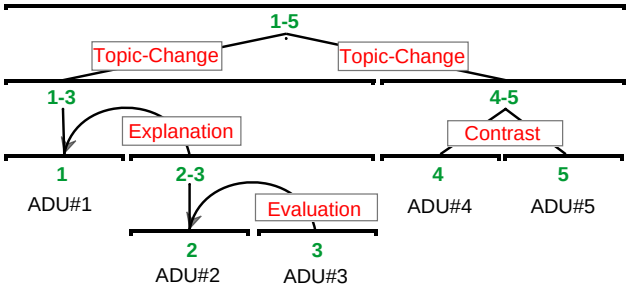
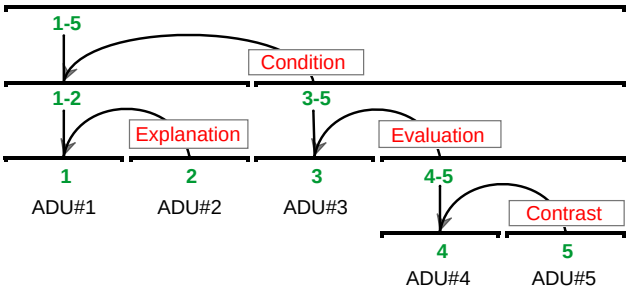


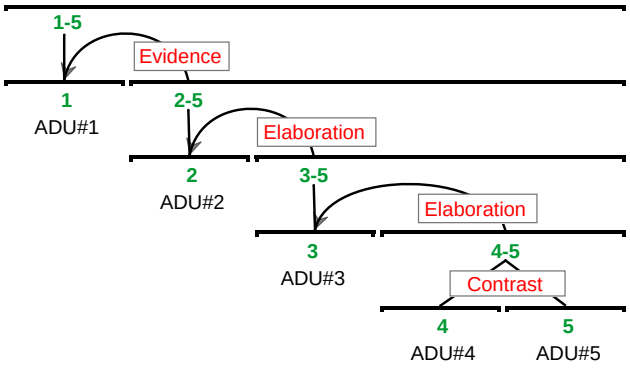
Рисунок 5.5 — Пример простой структуры аргументации, `micro_k002`.



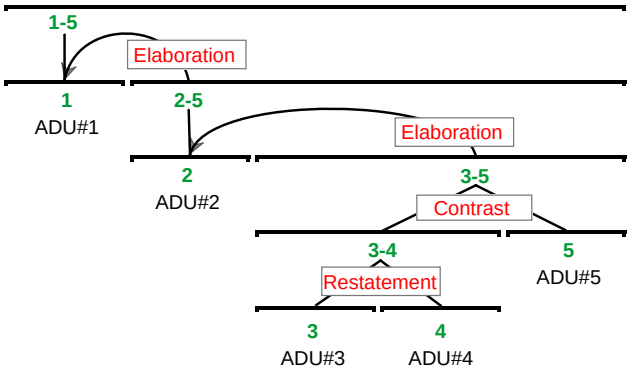
(а) Исходный текст на английском (`En`).



(б) Буквальный перевод с русского на английский (`Ru→En`).



(в) Исходный текст на русском `Ru→En`.



(г) Буквальный перевод `En` на русский.

Рисунок 5.6 — Четыре варианта структуры RST, предсказанные для документа `micro_k002`, сведенные к отношениям между аргументативными дискурсивными единицами.

В данном исследовании используются автоматические анализаторы дискурса для английского[6] [74] и русского (раздел 3.4.2) языков.

В первую очередь оценивается разнообразие риторических структур с опорой на стандарт сегментации АДЕ в корпусе. Рисунок 5.6 иллюстрирует вариации риторической структуры для перефразирований, полученных для примера в таблице 32, при условии, что листьями дискурсивного дерева являются аннотированные АДЕ. Ни одно из полученных риторических де-

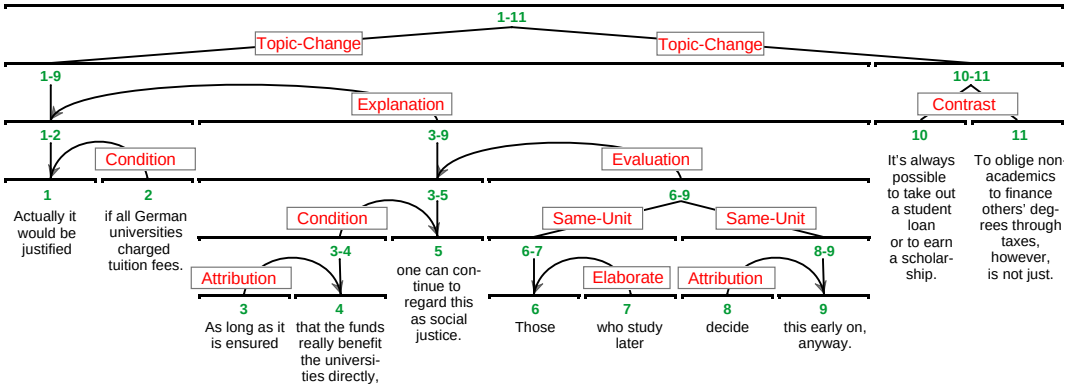
[6]Модели обучены на корпусе RST-DT.

ревьев не совпадает с экспертной разметкой аргументационной структуры (рис. 5.5). Тем не менее в каждом варианте наиболее центральная дискурсивная единица в риторическом дереве (АДЕ #1, по положению ядер) естественным образом соответствует центральному утверждению в аргументативной структуре.

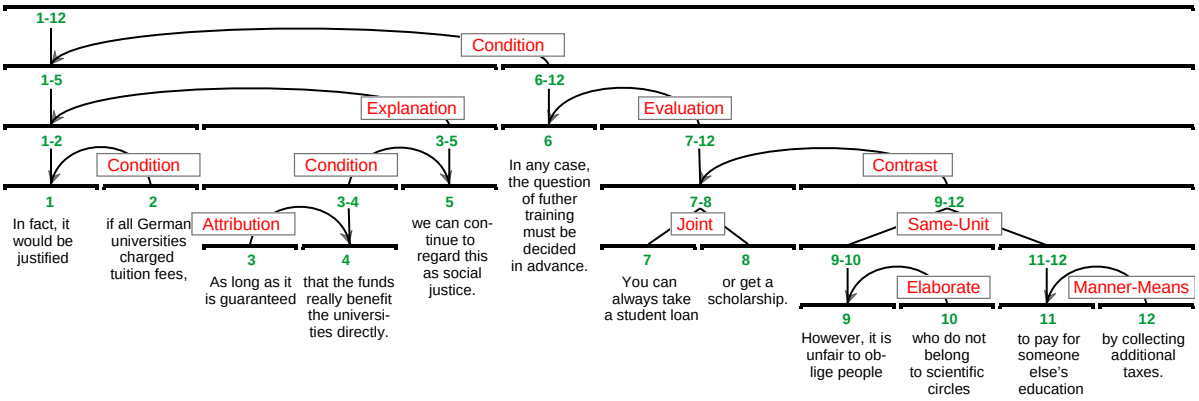
Оценка согласованности риторических структур по методу [93] показывает, что два варианта структуры, предсказанные одним и тем же парсером для английского языка (рис. 5.5а, 5.5б), имеют коэффициент согласованности 0,06 по ядерности и -0,04 по составляющим. Согласие по ядерности для двух вариаций, предсказанных одним и тем же парсером для русского языка (рис. 5.5в, 5.5г), составляет 0,6, а согласие по составляющим — 0,32. Значения согласия получены с помощью инструмента RST-Tace [198]. Оригинальные риторические структуры для примеров 5.6 с сохранением элементарных дискурсивных единиц представлены на рисунке 5.7; они иллюстрируют, как риторическая структура варьируется внутри отдельных АДЕ.

В таблице 33 приведены усредненные результаты оценки согласованности двух вариантов риторической разметки для каждого языка. Результаты показывают умеренную согласованность структуры составляющих и слабую согласованность ядерности и отношений. Согласованность по ядерности, признаку, непосредственно связанному с определением центральной идеи текста, в среднем наименьшая. Значение коэффициента Каппа для составляющих равно 1,0 в 22% пар текстов на английском и в 18% пар на русском языке. Полностью совпадающая риторическая структура обнаружена в 4% пар текстов на английском и в 8% на русском языке. Таким образом, выбранная стратегия перефразирования помогает получать риторические структуры с высокой вариативностью.

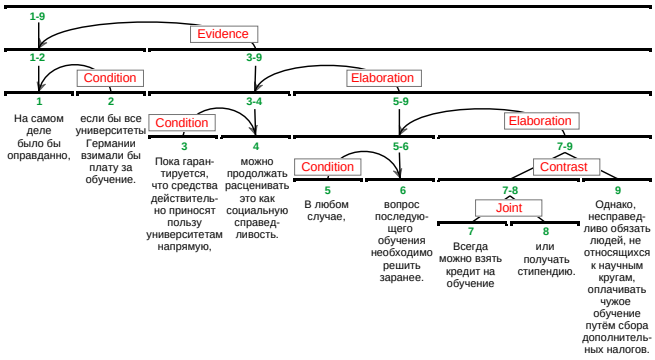
В экспериментах с методами BAP и DBAP перефразированные тексты и результаты их дискурсивного анализа используются для аугментации обучающих данных.



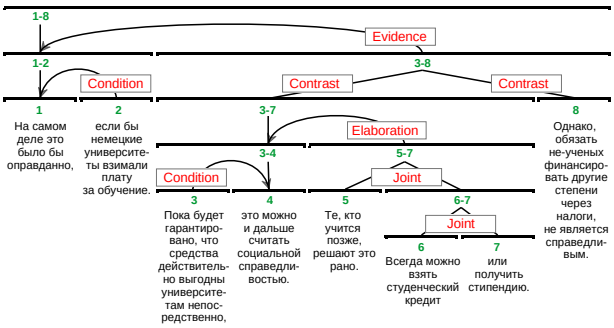
(а) Разбор текста на английском языке (`En`).



(б) Разбор перефразировки (`Рус.`→`Англ.`).



(в) Разбор текста на русском языке (`Ru`).



(г) Разбор перефразировки (`Англ.`→`Рус.`).

Рисунок 5.7 — Четыре варианта риторической структуры для структуры аргументации `micro_k002`.

Таблица 33 — Согласованность дискурсивного анализа между различными версиями одного текста на одном языке (среднее ± стандартное отклонение). Стандартная сегментация АДЕ.

Язык	Составляющие	Ядерность	Отношение	Среднее
En	0,56 ± 0,3	0,27 ± 0,5	0,35 ± 0,4	0,39 ± 0,3
Ru	0,50 ± 0,3	0,26 ± 0,4	0,29 ± 0,4	0,35 ± 0,3

Детали экспериментов

Эксперименты проводятся на оригинальных и дополненных вариантами риторических структур обучающих данных. Все эксперименты проводились на первых двух из десяти разбиений 5-кратной кросс-валидации из экспериментов, описанных для базового метода EG [195]. Поскольку в исходных разбиениях не предусмотренно валидационных данных, мы выбираем случайные 15% тренировочных данных в каждом разбиении для валидации. Обучающие данные дополнены второй частью корпуса, собранной с помощью краудсорсинга [199]. Так же, как и в связанных работах [157; 195; 199], мы используем упрощённый набор аргументативных функции, где функции «поддержка» («support»), «пример» («example») и «связь» («link») обобщаются функцией «поддержка» («support»), а функции «опровержение» («rebut») и «подрыв» («undercut») обобщаются функцией «атака» («attack»). Для извлечения признаков используется библиотека spaCy[7].

Все эксперименты, предполагающие использование предобученных языковых моделей, проводились с использованием модели `Microsoft/mDeBERTa_v3` [200] – многоязычной модели, подходящей для обоих языков.

Гиперпараметры настраиваются на валидационном подмножестве соответствующего разбиения. Для оптимизации используется Adam с параметром затухания веса 0.1 и коэффициентом отброса (dropout) 0,2; $\beta =$

[7] https://spacy.io/. Модели `en_core_web_lg` и `ru_core_news_lg`.

(0,9, 0,9). Для языковой модели скорость дообучения 2e-5, для случайно инициализированных слоёв — 2e-6. Дискурсивные коэффициенты обучаются со скоростью 2e-2. Размерность представления зависимостей составляет 100, а размерность представления тегов — 50. Максимальная длина последовательности установлена в 150 токенов, а размер батча — 4.

Для оценки анализа структуры аргументации, помимо показателей качества определения зависимостей (UAS, LAS), используются метрики, предложенные [195]. Таким образом, в результатах дополнительно указываются макроусреднённая F1 для обнаружения центрального утверждения (`cc`), назначения ролей (`ro`), классификации функций (`fu`), а также F1 для наличия связи (`at`).

Результаты оценки базового метода

В качестве базового метода используется классический метод EG [195]. В этом методе аргументативное дерево зависимостей на основе заданных АДЕ предсказывается используется при помощи таких признаков, как мешки слов и биграмм, мешки коннекторов и ассоциированных с ними дискурсивных отношений, части речи, знаки препинания, кластеры Брауна [201] для слов и биграмм, а также признак нахождения двух АДЕ в одном предложении. В таблице 34 показаны результаты базовой модели.

В экспериментах с двумя языками не использовались коннекторы или признаки, связанные с отношениями (`-Cues`). То же самое касается кластеров Брауна, которые недоступны для русского языка (`-BC`). Исключение этих признаков из исходной модели для английского языка в среднем приводит к снижению показателя F1 на 2,5% для обнаружения центральной единицы и назначения ролей, на 2,9% для классификации функций и на

Таблица 34 — Эффективность базового метода Evidence Graphs [195] на исходных и перефразированных данных (стандартная сегментация).

Данные	Признаки	cc	ro	fu	at	UAS	LAS
En	Все	87,3 ± 6,2	74,4 ± 4,9	75,9 ± 6,4	50,7 ± 4,3	56,3 ± 5,2	50,1 ± 5,1
En	-BC	85,9 ± 6,9	71,4 ± 5,7	75,0 ± 7,1	51,2 ± 4,2	56,3 ± 5,1	49,4 ± 5,1
En	-Cues	88,4 ± 7,1	71,1 ± 6,4	73,0 ± 7,1	50,9 ± 3,6	56,8 ± 6,2	49,2 ± 6,6
En	-BC, -Cues	84,8 ± 6,3	71,9 ± 5,2	73,0 ± 5,9	51,4 ± 4,9	56,1 ± 5,9	48,5 ± 5,6
Ru→En	-BC, -Cues	82,0 ± 7,0	72,5 ± 6,7	72,7 ± 4,8	52,9 ± 3,3	56,5 ± 4,3	50,4 ± 4,3
Ru	-BC, -Cues	85,3 ± 3,9	73,4 ± 7,0	74,5 ± 3,6	56,5 ± 4,3	60,4 ± 4,7	52,9 ± 4,6
En→Ru	-BC, -Cues	87,3 ± 6,8	73,8 ± 6,4	73,8 ± 6,3	57,5 ± 4,4	61,8 ± 5,7	54,7 ± 5,5

1,6% для LAS. Тем не менее, их исключение необходимо для стандартизации экспериментов с многоязычными данными.

Важным вопросом исследования с классическим методом EG является вопрос, нарушает ли машинный перевод структуру рассуждения. В таблице 34 дополнительно приведены результаты на перефразировках (`En`→`Ru`, `Ru`→`En`). Результаты на машинных переводах немного лучше, чем на экспертном переводе, за исключением определения центрального утверждения в данных на английском и функций в данных на русском языке. Однако показатели F1 для идентификации ролей и функций не отражают качество построения аргументативного дерева ввиду несбалансированности классов ролей и функций. Показатели LAS выше на парафразах. Наиболее вероятная причина заключается в том, что с каждым шагом перевода маркеры аргументации последовательно упрощаются. Результаты позволяют сделать вывод, что качество собранных дополнительных данных практически не уступает оригинальным.

Анализ структуры аргументации на основе экспертной сегментации АДЕ

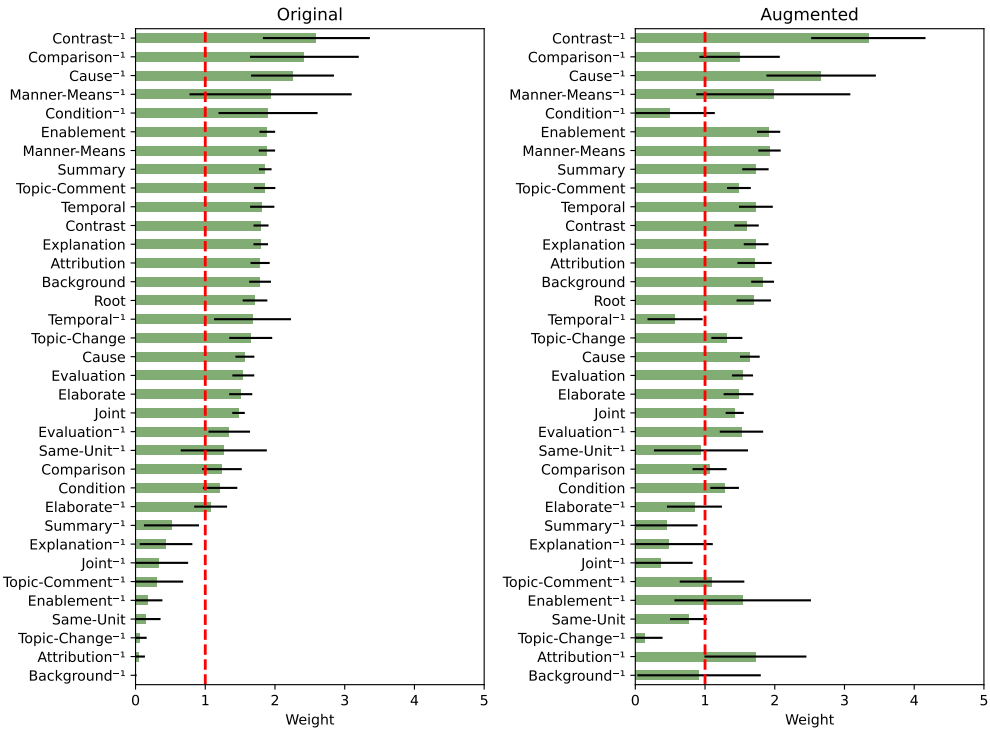
Результаты экспериментов с моделями анализа структуры на основе экспертной сегментации АДЕ представлены в таблице 35.

Таблица 35 — Оценки качества методов BAP и DBAP на корпусе с экспертной сегментацией АДЕ. Результаты, значимо отличающиеся от таковых без использования дополнительных вариантов структур, отмечены знаками * ($p < 0{,}05$) и ** ($p < 0{,}005$).

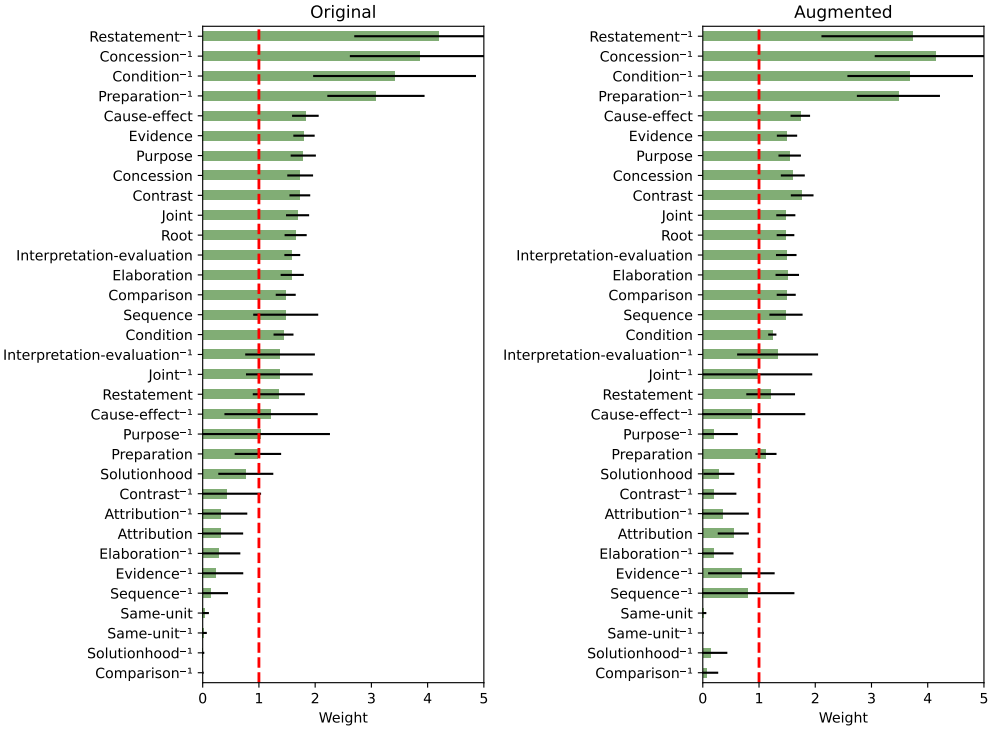
Метод	Доп. стр.	cc	ro	fu	at	UAS	LAS
Английский язык							
BAP	Нет	88,3 ± 4,9	**71,1 ± 5,7**	77,1 ± 4,6	53,8 ± 6,8	59,1 ± 6,8	52,9 ± 6,3
	Да	88,9 ± 4,7	69,2 ± 3,9	**78,3 ± 4,9**	56,2 ± 5,9	61,2 ± 5,8	55,1 ± 5,9
DBAP	Нет	**90,3 ± 3,3**	68,8 ± 6,9	77,3 ± 3,2	59,7 ± 7,4*	64,5 ± 6,6*	56,2 ± 5,3*
	Да	89,5 ± 4,3	68,8 ± 7,6	76,5 ± 3,1	**60,1 ± 4,3****	**64,6 ± 4,1***	**56,6 ± 3,2***
Русский язык							
BAP	Нет	**90,5 ± 5,7**	69,3 ± 7,8	78,9 ± 4,2	56,1 ± 6,3	61,7 ± 6,6	55,2 ± 6,7
	Да	90,3 ± 2,8	66,9 ± 6,9	77,5 ± 4,3	56,1 ± 5,1	61,6 ± 4,7	53,9 ± 5,7
DBAP	Нет	90,3 ± 5,7	68,9 ± 2,5	**79,8 ± 3,6**	59,8 ± 5,3	**64,6 ± 5,8**	**58,0 ± 3,6**
	Да	88,3 ± 6,4*	**69,9 ± 5,4**	77,2 ± 6,1	**60,6 ± 4,9***	**64,6 ± 5,8**	57,0 ± 5,8

Модели с использованием дискурса демонстрируют значительное улучшение качества относительно моделей, не использующих риторические структуры, при этом модели без дискурса демонстрируют лучшее качество, чем базовый метод EG (таблица 34). Хотя качество анализа структуры аргументации улучшается при добавлении текстовых перефразировок одних и тех же аргументов, введение вариаций риторической структуры при сохранении одинаковой сегментации может ухудшать результаты ввиду внесения шума в обучающие данные.

Дискурсивные коэффициенты $\boldsymbol{C}^{(\mathrm{RST})}$, оптимизированные при обучении моделей типа DBAP, иллюстрирует рисунок 5.8. Исходя из полученных



(а) Английский язык (набор отношений RST-DT).



(б) Русский язык (набор отношений RRT).

Рисунок 5.8 — Статистики коэффициентов $\boldsymbol{C}^{(\mathrm{RST})}$ для английского и русского вариантов корпуса.

результатов, риторические отношения можно разделить на четыре категории[8]:

I Сопровождающие аргументацию. Из набора отношений **RST-DT** таковыми являются следующие типы риторических отношений:

– Contrast, Contrast^{-1}. В RST-DT 17 общих типов риторических отношений соответствуют 78 типам детализированных отношений. В анализаторах на основе RST-DT, таким образом, детализированные отношения *Contrast*, *Concession* и *Antithesis* рассматриваются как единый тип Contrast. Простые случаи Contrast полностью согласуются с аргументативной структурой:

> [Компостирование помогает окружающей среде] $\xleftarrow{\text{Attack}}$ [Одним из недостатков компостирования является то, что не все материалы полезны для окружающей среды.] $_{\text{Contrast}\leftarrow}$

Положение риторического ядра и направление аргументации могут не совпадать, если подразумевается детализированное отношение, отличное от *Contrast*. Результаты для Contrast^{-1} согласуются с предыдущими исследованиями аргументативности отношения *Concession*, хотя предсказанное направление Сателлит→Ядро чаще всего противостоит направлению аргументации:

> [Верно, что социальные медиа очень полезны для поддержания связи на расстоянии,] $_{\text{Contrast}\rightarrow}$ $\xleftarrow{\text{Attack}}$ [но из-за отсутствия ограничений подростки часто проводят больше времени в виртуальном мире, чем в реальном.]

– Cause^{-1}, Cause. Несмотря на то, что причинно-следственные отношения сопутствуют той или иной аргументации, предсказанное положение ядра может противоречить направлению аргумента. Это может мешать анализу аргументации как разбору зависимостей на

[8]Примеры приводятся по русскоязычной версии корпуса Microtexts.

основе зависимостей в риторическом дереве, в особенности в случае глубокого и сложного дискурса.

> [Собачьи экскременты на тротуарах представляют реальную опасность.] $\xrightarrow{\text{Support}}$ [Поэтому увеличение штрафов — верный путь.] Cause←

Когда ядерность риторической связи соответствует причинно-следственной логике, дискурсивное и аргументативное отношения согласуются:

> [Супермаркетам и торговым центрам следует разрешить работу в любые воскресенья и праздники по их усмотрению.] $\xleftarrow{\text{Support}}$ [В этом случае воскресный шоппинг будет равномернее распределен в течение года.] Cause←

– Enablement, Manner-Means, Explanation. Риторические отношения аргументативной природы. Часто встречаются внутри предложения и характеризуются явными маркерами:

> [Супермаркеты должны взимать плату за пластиковые пакеты] $\xleftarrow{\text{Support}}$ [чтобы поощрять использование многоразовых сумок.] Enablement←

> [Переработка отходов помогает окружающей среде,] $\xleftarrow{\text{Support}}$ [удерживая вне её искусственные материалы.] Manner-Means←

– Summary. Как и Contrast, класс отношений Summary в анализаторах на основе RST-DT также включает другое детализированное отношение *Restatement*. В коротких текстах это отношение также может быть ошибочно присвоено примерам схожих неявных Elaborate и Explanation.

> [Жестокие игры заставляют людей реагировать неопределённым образом.] $\xleftarrow{\text{Support}}$ [Они вызывают стимуляцию, которая может провоцировать насилие.] Summary←

– Temporal (редкое отношение, встречается в 1% данных). Анализатор не обнаруживает в последовательности событий причинность, предполагаемую в аргументационном отношении:

[Маленькие дети играют в жестокие игры, считают наблюдаемое поведение нормой] $\xleftarrow{\text{Support}}$ [и переносят его в реальный мир.] Temporal←

– Joint, Elaborate. В Elaborate сателлит (всегда справа) предоставляет дополнительные детали о состоянии дел в ядре. В многоядерном Joint ядра независимы и равны по отношению к общей функции текста. Оба отношения являются наиболее частыми в текстах (38% разобранных документов на английском языке содержат по крайней мере одно Elaborate и 13% по крайней мере одно Joint). Риторические анализаторы склонны предсказывать эти отношения из-за дисбаланса классов в корпусах TPC-разметки.

[так как он имеет самое посредственное разрешение,] $\xleftarrow{\text{Support}}$ [и фото в темноте часто зашумлено.] Joint←

[Носороги вымирают.] $\xleftarrow{\text{Support}}$ [Браконьеры безжалостно убивают носорогов.] Elaborate←

Для набора отношений **RRT**:

– Restatement^{-1}. Одна из частей этого отношения (ядро или сателлит) может выступать в качестве поддерживающего аргумента для другой в аргументации:

[Они сделали общение доступнее,] $\xrightarrow{\text{Support}}$ [это означает, что семьи общаются даже тогда, когда раньше общение было невозможно.] Restatement←

В корпусе RST-DT отношение *Restatement* – часть типа Summary.

– Concession^{-1}. Этот класс описан выше как часть отношения Contrast в RST-DT.

[Хотя кажется, что такое поведение препятствует развитию самостоятельности ребенка,] $_{\text{Concession}\rightarrow}$ $\xleftarrow{\text{Attack}}$ [на самом деле ребенок сам знает, что его родитель всегда будет где-то рядом.]

– Condition^{-1}. Для автоматического анализа характерна ошибка, когда отношение Concession выражается коннектором «даже если»:

[Даже если можно подумать, что необходим дополнительный контроль за арендной платой помимо текущих мер по защите арендаторов,] $_{\text{Condition}\rightarrow}$ $\xleftarrow{\text{Attack}}$ [нельзя отбирать у давних домовладельцев возможность корректировать свой доход в соответствии с требованиями рынка.]

– Preparation^{-1}. Отсутствует как отдельное отношение в RST-DT. Preparation в RRT можно рассматривать как прямую противоположность Elaboration. В этом отношении спутник задаёт или вводит тему для ядра, но сам содержит минимальную информацию. Анализатор для русского языка склонен назначать это отношение первому утверждению в тексте:

[Да, атомная энергия безопасна.] $_{\text{Preparation}\rightarrow}$ $\xleftarrow{\text{Support}}$ [На электростанциях существуют специальные защитные меры для предотвращения аварий.]

– Cause-effect. Соответствует отношениям Cause/Cause^{-1}, описанным выше. Однако важно отметить, что в корпусе RRT нуклеарность причинно-следственных отношений определяется логикой описываемых событий, а не отношением к точке зрения автора. Такие риторические отношения лучше соответствуют логике аргументации; однако преобразование в риторическое зависимое дерево в этих обстоятельствах может нарушить согласованность структуры всего текста.

[Пока охота ведется сдержанно и в практических целях, окружающей среде ничего не угрожает.] $\text{CAUSE-EFFECT}\rightarrow \xrightarrow{\text{Support}}$ [Так что в охоте нет ничего страшного.]

- CONTRAST. Аналогично CONTRAST и CONTRAST^{-1} в RST-DT. В RRT отношение CONTRAST всегда многоядерное. Поэтому при преобразовании в зависимость главным считается высказывание слева. Результаты показывают, что такое определение CONTRAST согласуется с аргументативными функциями, в то время как обратные (слева направо) дуги (CONTRAST^{-1}) последовательно штрафуются (см. примеры для RST-DT).
- PURPOSE. См. ENABLEMENT в RST-DT.
- EVIDENCE. Входит в EXPLANATION в RST-DT.
- SEQUENCE. Входит в TEMPORAL в RST-DT. Описание и пример для последовательного TEMPORAL см. выше.
- JOINT и ELABORATION. См. JOINT, ELABORATE.

II Противоположные аргументации. Риторические отношения, наличие которых штрафует вероятность дуги аргументации. **RST-DT:** SUMMARY^{-1}, JOINT^{-1}, TOPIC-CHANGE^{-1}, SAME-UNIT. **RRT:** отношения, противоположные аргументации (PURPOSE^{-1}, CONTRAST^{-1}, ELABORATION^{-1}, JOINT^{-1}), а также неаргументативные отношения (SOLUTIONHOOD, SOLUTIONHOOD^{-1}, ATTRIBUTION, ATTRIBUTION^{-1}, SAME-UNIT, SAME-UNIT^{-1}).

III Слабо сопутствующие аргументации. Среднее значение коэффициента немного превышает единицу при высокой дисперсии. Эти риторические отношения часто ошибочно предсказываются в аргументативных текстах. **RST-DT:** COMPARISON^{-1}, MANNER-MEANS^{-1}, SOLUTIONHOOD^{-1}, ENABLEMENT^{-1}, ATTRIBUTION^{-1}. **RRT:** $\text{INTERPRETATION-EVALUATION}^{-1}$.

IV Слабо противоположные. Значения около или ниже единицы при высокой дисперсии. **RST-DT:** Condition^{-1}, Temporal^{-1}, Same-Unit^{-1}, Explanation^{-1}, Background^{-1}, Topic-Comment^{-1}. **RRT:** Cause-effect^{-1}, Evidence^{-1}, Sequence^{-1}.

Анализ структуры аргументации на основе предсказанной сегментации ЭДЕ

В таблице 36 приведены результаты оценки полного анализа структуры аргументации на тех же тестовых данных. В качестве терминальных узлов рассматриваются ЭДЕ. Для корректного сравнения моделей BAP и DBAP фиктивная функция “same-arg” исключена из оценки. Результаты для BAP в этом случае можно рассматривать как базовый бесструктурный порог оценки.

Таблица 36 — Оценки качества полного анализа структуры аргументации. Результаты, значимо отличающиеся от результатов для базового метода, отмечены знаками * ($p < 0{,}05$) и ** ($p < 0{,}005$).

Метод	Доп. стр.	cc	ro	fu	at	UAS	LAS
Английский язык							
BAP	Нет	**86,8 ± 6,1**	60,3 ± 4,2	**40,0 ± 2,7**	39,2 ± 5,0	40,8 ± 8,0	23,1 ± 6,8
	Да	86,3 ± 4,8	**64,5 ± 5,0***	39,3 ± 3,0	42,0 ± 5,6	40,0 ± 6,7	25,5 ± 5,7
DBAP	Нет	85,8 ± 5,5	60,4 ± 6,2	39,3 ± 2,7	65,7 ± 2,9**	59,0 ± 4,7**	23,0 ± 5,2
	Да	84,8 ± 5,0	62,9 ± 5,7	39,1 ± 2,8	**66,7 ± 4,2****	**62,2 ± 3,8****	**26,3 ± 7,1**
Русский язык							
BAP	Нет	88,6 ± 6,9	60,6 ± 5,8	42,3 ± 2,6	42,5 ± 6,4	45,4 ± 7,9	28,6 ± 6,4
	Да	**90,2 ± 5,7**	58,9 ± 3,5	**43,6 ± 2,5**	43,5 ± 5,5	47,2 ± 8,0	30,7 ± 6,9
DBAP	Нет	86,5 ± 5,2	59,7 ± 5,5	42,9 ± 2,9	60,7 ± 4,9**	**59,6 ± 8,6****	**31,7 ± 5,8**
	Да	87,5 ± 5,1	**60,7 ± 4,9**	41,9 ± 3,5	**61,0 ± 3,8****	58,2 ± 4,2**	29,9 ± 7,4

При добавлении второго варианта риторической структуры в обучающих данных увеличивается разнообразие представлений связей между

одними и теми же листовыми узлами. Это помогает обнаружить более общие дискурсивные шаблоны, что приводит к улучшению качества анализа на оригинальных тестовых данных на английском языке. На данных на русском языке не наблюдалось значительного улучшения качества, что подчёркивает различия между интерпретациями нуклеарности в двух корпусах риторической разметки. В результате объединения нескольких отношений в одно определение нуклеарности для некоторых отношений в корпусе экспертной разметки RRT для русского языка отличается от оригинального теоретического определения в TPC. В отношениях Cause-effect и Purpose ядро всегда подразумевает логический эффект, независимо от намерения автора. Это влияет на адекватность получаемого при конвертации составляющих дискурсивного дерева риторических зависимостей.

5.5. Выводы

Методы анализа дискурсивной структуры позволяют улучшить решение прикладных задач обработки естественного языка. Классификаторы текстов, рассматривающие документ как последовательность токенов, демонстрируют высокую эффективность при работе с короткими текстами, однако их качество снижается при анализе сложных рассуждений, где важен учёт иерархической организации дискурса. Для вычисления более точных векторных представлений сложного текста в диссертации предложен метод, сочетающий преимущества предобученной языковой модели и гибридного анализатора риторической структуры.

Результаты экспериментальных исследований показали, что использование признаков расстояния между сущностями в риторической структуре существенно улучшает качество разрешения кореференции, особенно на крупных корпусах. Это подчеркивает важность интеграции дискурсив-

ных признаков при решении задач кореференции, особенно при работе с текстами сложной дискурсивной структуры. При этом стоит учитывать адаптацию риторических анализаторов к текстам разных жанров и объемов. Предложен метод разрешения кореференции на русском языке, использующий признаки расстояний между упоминаниями в риторической структуре.

Показано, что использование в качестве исходных данных автоматически предсказанных риторических структур позволяет обучить на небольшом корпусе модель полного анализа структуры аргументации в тексте-рассуждении. Анализ поощряемых и штрафуемых анализатором аргументации риторических отношений демонстрирует множество соответствий между двумя моделями описания текста. Результаты экспериментальных исследований демонстрируют, что анализ вариантов риторических структур способствует более точному моделированию аргументации, что важно для построения надежных и интерпретируемых систем анализа рассуждений.

Полученные в рамках диссертационного исследования результаты показывают, что учёт риторических признаков позволяет повысить качество классификации текстов с рассуждениями, разрешения кореференции и анализа структуры аргументации. Результаты, полученные в главе 5, представлены в публикациях «Влияние признаков иерархического дискурса на разрешение кореференции в русском языке» [4], “Discourse-aware text classification for argument mining” [9], “Light Coreference Resolution for Russian with Hierarchical Discourse Features” [10], “End-to-End Argument Mining over Varying Rhetorical Structures” [1].

Представленные в Главе 5 результаты в части метода разрешения кореференции с использованием риторических признаков 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».

Заключение

Основные результаты, полученные в дисертационном исследовании:

1. Разработаны и реализованы методы анализа риторических структур в текстах на русском языке. Проведены экспериментальные исследования методов анализа риторических отношений, поверхностного и полнотекстового риторического анализа текстов на русском языке.
2. Исследованы возможности кросс-языковой адаптации дискурсивного анализа на материале большого параллельного корпуса дискурсивной разметки.
3. Разработан метод риторического анализа, позволяющий достичь качества полнотекстового анализа риторической структуры текста на русском и английском языках, превышающего качество предыдущих систем.
4. Разработан метод реализации риторического анализа при помощи глубокого обучения на материалах разнородной риторической разметки.
5. Разработан метод классификации текстов с учетом риторических структур, показана 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.

Список сокращений и условных обозначений

АДЕ	Аргументационная дискурсивная единица
ДЕ	Дискурсивная единица
ТРС	Теория риторических структур
ЭДЕ	Элементарная дискурсивная единица
ЯМ	Языковая модель
BiMPM	Bilateral Multi-Perspective Matching, многостороннее симметричное сопоставление
CRF	Conditional random field, условные случайные поля
GRU	Gated Recurrent Units, управляемые рекуррентные блоки
LAS	Labeled attachment score
SVM	Support vector machine, метод опорных векторов
LSTM	Long short-term memory, долгосрочная краткосрочная память
UAS	Unlabeled attachment score

Список литературы

1. *Chistova Elena.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023. — Toronto, Canada: Association for Computational Linguistics, 2023. — Pp. 3376–3391.

2. *Chistova Elena.* Bilingual Rhetorical Structure Parsing with Large Parallel Annotations // Findings of the Association for Computational Linguistics: ACL 2024. — Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024. — Pp. 9689–9706.

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

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

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

6. *Chistova Elena et al.* Classification models for RST discourse parsing of texts in Russian // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”. — No. 18. — 2019. — Pp. 163–176.

7. *Chistova Elena et al.* Towards the Data-driven System for Rhetorical Parsing of Russian Texts // Proceedings of the Workshop on Discourse Relation

Parsing and Treebanking 2019. — Minneapolis, MN: Association for Computational Linguistics, 2019. — Pp. 82–87.

8. *Chistova Elena et al.* RST discourse parser for Russian: an experimental study of deep learning models // Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020 / Springer. — Moscow, Russia: 2021. — Pp. 105–119.

9. *Chistova Elena, Smirnov Ivan.* Discourse-aware text classification for argument mining // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — No. 2022. — 2022. — Pp. 93–105.

10. *Chistova Elena, Smirnov Ivan.* Light Coreference Resolution for Russian with Hierarchical Discourse Features // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — No. 22. — 2023. — Pp. 34–41.

11. *Mikolov Tomas et al.* Distributed representations of words and phrases and their compositionality // *Advances in neural information processing systems.* — 2013. — Vol. 26.

12. *Pennington Jeffrey, Socher Richard, Manning Christopher.* GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Doha, Qatar: Association for Computational Linguistics, 2014. — Pp. 1532–1543.

13. *Sennrich Rico, Haddow Barry, Birch Alexandra.* Neural Machine Translation of Rare Words with Subword Units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Berlin, Germany: Association for Computational Linguistics, 2016. — Pp. 1715–1725.

14. *Kudo Taku, Richardson John.* SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Brussels, Belgium: Association for Computational Linguistics, 2018. — Pp. 66–71.

15. Deep Contextualized Word Representations / Matthew E. Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — New Orleans, Louisiana: Association for Computational Linguistics, 2018. — Pp. 2227–2237.

16. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota: Association for Computational Linguistics, 2019. — Pp. 4171–4186.

17. *Apidianaki Marianna.* From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation // *Computational Linguistics.* — 2023. — Vol. 49, no. 2. — Pp. 465–523.

18. The Penn Discourse TreeBank 2.0. / Rashmi Prasad, Nikhil Dinesh, Alan Lee et al. // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). — Marrakech, Morocco: European Language Resources Association (ELRA), 2008.

19. Annotating Subordinators in the Turkish Discourse Bank / Deniz Zeyrek, Umit Deniz Turan, Cem Bozsahin et al. // Proceedings of the Third Linguistic Annotation Workshop (LAW III). — Suntec, Singapore: Association for Computational Linguistics, 2009. — Pp. 44–47.

20. *Mirzaei Azadeh, Safari Pegah.* Persian Discourse Treebank and coreference corpus // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

21. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style / Deniz Zeyrek, Amália Mendes, Yulia Grishina et al. // *Language Resources and Evaluation.* — 2020. — Vol. 54, no. 2. — Pp. 587–613.

22. Announcing the Prague Discourse Treebank 3.0 // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia: ELRA and ICCL, 2024. — Pp. 1270–1279.

23. *Prasertsom Ponrawee, Jaroonpol Apiwat, Rutherford Attapol T.* The Thai Discourse Treebank: Annotating and Classifying Thai Discourse Connectives // *Transactions of the Association for Computational Linguistics.* — 2024. — Vol. 12. — Pp. 613–629.

24. *Zhou Yuping, Xue Nianwen.* The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations // *Language Resources and Evaluation.* — 2015. — Vol. 49. — Pp. 397–431.

25. The Utility of Discourse Parsing Features for Predicting Argumentation Structure / Freya Hewett, Roshan Prakash Rane, Nina Harlacher, Manfred Stede // Proceedings of the 6th Workshop on Argument Mining.

— Florence, Italy: Association for Computational Linguistics, 2019. — Pp. 98–103.

26. Using discourse signals for robust instructor intervention prediction / Muthu Kumar Chandrasekaran, Carrie Epp, Min-Yen Kan, Diane Litman // Proceedings of the AAAI Conference on Artificial Intelligence. — Vol. 31. — 2017.

27. Unsupervised extraction of semantic relations using discourse cues / Juliette Conrath, Stergos Afantenos, Nicholas Asher, Philippe Muller // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. — Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. — Pp. 2184–2194.

28. DiscoLQA: zero-shot discourse-based legal question answering on European Legislation / Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, Vittoria Pistone // *Artificial Intelligence and Law*. — 2024. — Pp. 1–37.

29. *Wolf Florian, Gibson Edward*. Representing Discourse Coherence: A Corpus-Based Study // *Computational Linguistics*. — 2005. — Vol. 31, no. 2. — Pp. 249–287.

30. eRST: A Signaled Graph Theory of Discourse Relations and Organization / Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu et al. // *Computational Linguistics*. — 2024. — Pp. 1–50.

31. *Mann William C, Thompson Sandra A*. Rhetorical structure theory: Toward a functional theory of text organization // *Text-interdisciplinary Journal for the Study of Discourse*. — 1988. — Vol. 8, no. 3. — Pp. 243–281.

32. *Ji Yangfeng, Smith Noah A*. Neural Discourse Structure for Text Categorization // Proceedings of the 55th Annual Meeting of the Association

for Computational Linguistics (Volume 1: Long Papers). — 2017. — Pp. 996–1005.

33. *Lee Kangwook et al.* A discourse-aware neural network-based text model for document-level text classification // *Journal of Information Science*. — 2018. — Vol. 44, no. 6. — Pp. 715–735.

34. *Goyal Naman, Eisenstein Jacob.* A Joint Model of Rhetorical Discourse Structure and Summarization // Proceedings of the Workshop on Structured Prediction for NLP. — Austin, TX: Association for Computational Linguistics, 2016. — Pp. 25–34.

35. *Pu Dongqi, Wang Yifan, Demberg Vera.* Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Toronto, Canada: Association for Computational Linguistics, 2023. — Pp. 5574–5590.

36. *Pu Dongqi, Demberg Vera.* RST-LoRA: A Discourse-Aware Low-Rank Adaptation for Long Document Abstractive Summarization // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). — Mexico City, Mexico: Association for Computational Linguistics, 2024. — Pp. 2200–2220.

37. *Tu Mei, Zhou Yu, Zong Chengqing.* A Novel Translation Framework Based on Rhetorical Structure Theory // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Sofia, Bulgaria: Association for Computational Linguistics, 2013. — Pp. 370–374.

38. Discourse Structure in Machine Translation Evaluation / Shafiq Joty, Francisco Guzmán, Lluís Màrquez, Preslav Nakov // *Computational Linguistics*. — 2017. — Vol. 43, no. 4. — Pp. 683–722.

39. Modeling Discourse Structure for Document-level Neural Machine Translation / Junxuan Chen, Xiang Li, Jiarui Zhang et al. // Proceedings of the First Workshop on Automatic Simultaneous Translation. — Seattle, Washington: Association for Computational Linguistics, 2020. — Pp. 30–36.

40. *Galitsky Boris, Ilvovsky Dmitry, Goncharova Elizaveta*. Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). — Held Online: INCOMA Ltd., 2021. — Pp. 444–453.

41. Structure-Discourse Hierarchical Graph for Conditional Question Answering on Long Documents / Haowei Du, Yansong Feng, Chen Li et al. // Findings of the Association for Computational Linguistics: ACL 2023. — Toronto, Canada: Association for Computational Linguistics, 2023. — Pp. 6282–6293.

42. Parallel Discourse Annotations on a Corpus of Short Texts / Manfred Stede, Stergos Afantenos, Andreas Peldszus et al. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — European Language Resources Association (ELRA), 2016. — Pp. 1051–1058.

43. *Accuosto Pablo, Saggion Horacio*. Transferring Knowledge from Discourse to Arguments: A Case Study with Scientific Abstracts // Proceedings of the 6th Workshop on Argument Mining. — Florence, Italy: Association for Computational Linguistics, 2019. — Pp. 41–51.

44. Discourse Structure-Aware Prefix for Generation-Based End-to-End Argumentation Mining / Yang Sun, Guanrong Chen, Caihua Yang et al. // Findings of the Association for Computational Linguistics ACL 2024. — Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024. — Pp. 11597–11613.

45. *Aldawsari Mohammed, Finlayson Mark.* Detecting Subevents using Discourse and Narrative Features // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy: Association for Computational Linguistics, 2019. — Pp. 4780–4790.

46. Distinguishing Between Foreground and Background Events in News / Mohammed Aldawsari, Adrian Perez, Deya Banisakher, Mark Finlayson // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. — Pp. 5171–5180.

47. *Dachkovsky Svetlana, Stamp Rose, Sandler Wendy.* Mapping the body to the discourse hierarchy in sign language emergence // *Language and Cognition.* — 2023. — Vol. 15, no. 1. — Pp. 53–85.

48. Towards building a discourse-annotated corpus of Russian / Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva et al. // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"(2017). — 2017. — Pp. 201–212.

49. *Marcu Daniel.* The theory and practice of discourse parsing and summarization. — MIT press, 2000.

50. *Miller George A.* WordNet: a lexical database for English // *Communications of the ACM.* — 1995. — Vol. 38, no. 11. — Pp. 39–41.

51. *Carlson Lynn, Marcu Daniel, Okurovsky Mary Ellen.* Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory //

Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. — 2001.

52. *Marcus Mitchell P., Santorini Beatrice, Marcinkiewicz Mary Ann.* Building a Large Annotated Corpus of English: The Penn Treebank // *Computational Linguistics.* — 1993. — Vol. 19, no. 2. — Pp. 313–330.

53. *Li Jinfen, Xiao Lu.* Neural-based RST Parsing And Analysis In Persuasive Discourse // Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). — Online: Association for Computational Linguistics, 2021. — Pp. 274–283.

54. *Guz Grigorii, Carenini Giuseppe.* Coreference for Discourse Parsing: A Neural Approach // Proceedings of the First Workshop on Computational Approaches to Discourse. — Online: Association for Computational Linguistics, 2020. — Pp. 160–167.

55. *Guz Grigorii, Huber Patrick, Carenini Giuseppe.* Unleashing the Power of Neural Discourse Parsers - A Context and Structure Aware Approach Using Large Scale Pretraining // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. — Pp. 3794–3805.

56. A Simple and Strong Baseline for End-to-End Neural RST-style Discourse Parsing / Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito et al. // Findings of the Association for Computational Linguistics: EMNLP 2022. — Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. — Pp. 6725–6737.

57. *Sagae Kenji.* Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing // Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09). — Paris, France: Association for Computational Linguistics, 2009. — Pp. 81–84.

58. *Hernault Hugo et al.* HILDA: A discourse parser using support vector machine classification // *Dialogue & Discourse.* — 2010. — Vol. 1, no. 3. — Pp. 1–33.

59. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis / Shafiq Joty, Giuseppe Carenini, Raymond Ng, Yashar Mehdad // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Sofia, Bulgaria: Association for Computational Linguistics, 2013. — Pp. 486–496.

60. *Feng Vanessa Wei, Hirst Graeme.* A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Baltimore, Maryland: Association for Computational Linguistics, 2014. — Pp. 511–521.

61. *Braud Chloé, Plank Barbara, Søgaard Anders.* Multi-view and multi-task training of RST discourse parsers // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — Osaka, Japan: The COLING 2016 Organizing Committee, 2016. — Pp. 1903–1913.

62. *Joty Shafiq, Carenini Giuseppe, Ng Raymond T.* CODRA: A Novel Discriminative Framework for Rhetorical Analysis // *Computational Linguistics.* — 2015. — Vol. 41, no. 3. — Pp. 385–435.

63. *Ji Yangfeng, Eisenstein Jacob.* Representation Learning for Text-level Discourse Parsing // Proceedings of the 52nd Annual Meeting of the

Association for Computational Linguistics (Volume 1: Long Papers). — Baltimore, Maryland: Association for Computational Linguistics, 2014. — Pp. 13–24.

64. *Wang Yizhong, Li Sujian, Wang Houfeng.* A Two-Stage Parsing Method for Text-Level Discourse Analysis // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Vancouver, Canada: Association for Computational Linguistics, 2017. — Pp. 184–188.

65. *Li Jiwei, Li Rumeng, Hovy Eduard.* Recursive Deep Models for Discourse Parsing // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Doha, Qatar: Association for Computational Linguistics, 2014. — Pp. 2061–2069.

66. *Li Qi, Li Tianshi, Chang Baobao.* Discourse Parsing with Attention-based Hierarchical Neural Networks // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. — Austin, Texas: Association for Computational Linguistics, 2016. — Pp. 362–371.

67. Modeling discourse cohesion for discourse parsing via memory network / Yanyan Jia, Yuan Ye, Yansong Feng et al. // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Melbourne, Australia: Association for Computational Linguistics, 2018. — Pp. 438–443.

68. *Yu Nan, Zhang Meishan, Fu Guohong.* Transition-based Neural RST Parsing with Implicit Syntax Features // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. — Pp. 559–570.

69. Prompting Implicit Discourse Relation Annotation / Frances Yung, Mansoor Ahmad, Merel Scholman, Vera Demberg // Proceedings of The 18th

Linguistic Annotation Workshop (LAW-XVIII). — St. Julians, Malta: Association for Computational Linguistics, 2024. — Pp. 150–165.

70. Can we obtain significant success in RST discourse parsing by using Large Language Models? / Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). — St. Julian's, Malta: Association for Computational Linguistics, 2024. — Pp. 2803–2815.

71. A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure / Longyin Zhang, Yuqing Xing, Fang Kong et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online: Association for Computational Linguistics, 2020. — Pp. 6386–6395.

72. *Kobayashi Naoki et al.* Top-Down RST Parsing Utilizing Granularity Levels in Documents // *Proceedings of the AAAI Conference on Artificial Intelligence*. — 2020. — Vol. 34, no. 05. — Pp. 8099–8106.

73. *Koto Fajri, Lau Jey Han, Baldwin Timothy.* Top-down Discourse Parsing via Sequence Labelling // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. — Online: Association for Computational Linguistics, 2021. — Pp. 715–726.

74. *Zhang Longyin, Kong Fang, Zhou Guodong.* Adversarial Learning for Discourse Rhetorical Structure Parsing // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Online: Association for Computational Linguistics, 2021. — Pp. 3946–3957.

75. RST Parsing from Scratch / Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, Xiaoli Li // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 2021. — Pp. 1613–1625.

76. *Liu Zhengyuan, Shi Ke, Chen Nancy.* DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing // Proceedings of the 2nd Workshop on Computational Approaches to Discourse. — Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, 2021. — Pp. 154–164.

77. *Zeldes Amir.* The GUM corpus: Creating multilayer resources in the classroom // *Language Resources and Evaluation.* — 2017. — Vol. 51, no. 3. — Pp. 581–612.

78. *Iruskieta Mikel et al.* The RST Basque TreeBank: an online search interface to check rhetorical relations // 4th workshop RST and discourse studies. — 2013. — Pp. 40–49.

79. *Cardoso Paula Christina Figueira, Maziero Erick Galani.* CSTNews — a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese // 3rd RST Brazilian Meeting. — 2011.

80. *Collovini Sandra et al.* // *Proceedings of TIL.* — 2007. — Vol. 121.

81. *Pardo Thiago Alexandre Salgueiro, Seno Eloize Rossi Marques.* Rhetalho: um corpus de referência anotado retoricamente // *Anais do V Encontro de Corpora.* — 2005. — Pp. 24–25.

82. *Pardo Thiago Alexandre Salgueiro, Nunes Maria das Graças Volpe.* A construção de um corpus de textos científicos em português do brasil e sua marcação retórica. — 2003.

83. *da Cunha Iria, Torres-Moreno Juan-Manuel, Sierra Gerardo.* On the Development of the RST Spanish Treebank // Proceedings of the 5th Linguistic Annotation Workshop. — Portland, Oregon, USA: Association for Computational Linguistics, 2011. — Pp. 1–10.

84. *Cao Shuyuan, da Cunha Iria, Iruskieta Mikel.* The RST Spanish-Chinese Treebank // Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. — Pp. 156–166.

85. *Peng Siyao, Liu Yang Janet, Zeldes Amir.* GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing // Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). — Online only: Association for Computational Linguistics, 2022. — Pp. 382–391.

86. *Stede Manfred, Neumann Arne.* Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. // LREC. — 2014. — Pp. 925–929.

87. Multi-layer discourse annotation of a Dutch text corpus / Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet et al. // *age.* — 2012. — Vol. 1. — P. 2.

88. *Shahmohammadi Sara, Veisi Hadi, Darzi Ali.* Persian Rhetorical Structure Theory // *arXiv preprint arXiv:2106.13833.* — 2021.

89. *Stede Manfred.* Disambiguating rhetorical structure // *Research on Language and Computation.* — 2008. — Vol. 6. — Pp. 311–332.

90. *Iruskieta Mikel, Braud Chloé.* EusDisParser: improving an under-resourced discourse parser with cross-lingual data // Proceedings of the Workshop on

Discourse Relation Parsing and Treebanking 2019. — Minneapolis, MN: Association for Computational Linguistics, 2019. — Pp. 62–71.

91. *Liu Zhengyuan, Shi Ke, Chen Nancy.* Multilingual Neural RST Discourse Parsing // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. — Pp. 6730–6738.

92. *Morey Mathieu, Muller Philippe, Asher Nicholas.* How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark: Association for Computational Linguistics, 2017. — Pp. 1319–1324.

93. *Iruskieta Mikel, Da Cunha Iria, Taboada Maite.* A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora // *Language resources and evaluation.* — 2015. — Vol. 49, no. 2. — Pp. 263–309.

94. *Braud Chloé, Coavoux Maximin, Søgaard Anders.* Cross-lingual RST Discourse Parsing // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. — Valencia, Spain: Association for Computational Linguistics, 2017. — Pp. 292–304.

95. *Zeyrek Deniz, Mendes Amália, Kurfalı Murathan.* Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

96. *Soricut R., Marcu D.* Sentence level discourse parsing using syntactic and lexical information // Proceedings of the 2003 Conference of the North

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. — 2003. — Pp. 149–156.

97. *Sagae Kenji.* Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing // Proceedings of the 11th International Conference on Parsing Technologies. — 2009. — Pp. 81–84.

98. *Lin Z., Kan M. Y., Ng H. T.* Recognizing implicit discourse relations in the Penn Discourse Treebank // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. — 2009. — Pp. 343–351.

99. *Feng V. W., Hirst G.* Text-level discourse parsing with rich linguistic features // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vol. 1. — 2012. — Pp. 60–68.

100. *Feng Vanessa Wei, Hirst Graeme.* A linear-time bottom-up discourse parser with constraints and post-editing // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2014. — Pp. 511–521.

101. *Li J., Li R., Hovy E.* Recursive deep models for discourse parsing // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. — 2014. — Pp. 2061–2069.

102. Implicit Discourse Relation Recognition using Neural Tensor Network with Interactive Attention and Sparse Learning / Fengyu Guo, Ruifang He, Di Jin et al. // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. — Pp. 547–558.

103. Using entity features to classify implicit discourse relations / Annie Louis, Aravind Joshi, Rashmi Prasad, Ani Nenkova // Proceedings of the SIGDIAL 2010 Conference. — Tokyo, Japan: Association for Computational Linguistics, 2010. — Pp. 59–62.

104. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution / W. Lei, Y. Xiang, Y. Wang et al. // Thirty-Second AAAI Conference on Artificial Intelligence. — 2018.

105. Implicit discourse relation classification via multi-task neural networks / Yang Liu, Sujian Li, Xiaodong Zhang, Zhifang Sui // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. — AAAI'16. — AAAI Press, 2016. — P. 2750–2756.

106. *Demberg Vera, Scholman Merel CJ, Asr Fatemeh Torabi.* How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations // *Dialogue & Discourse.* — 2019. — Vol. 10, no. 1. — Pp. 87–135.

107. *Salton G., Buckley Ch.* Term-weighting approaches in automatic text retrieval // *Information processing & management.* — 1988. — Vol. 24, no. 5. — Pp. 513–523.

108. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // ICLR Workshop. — 2013.

109. *Boser B.E., Guyon I.M., Vapnik V.N.* A training algorithm for optimal margin classifiers // Proceedings of the fifth annual workshop on Computational learning theory / ACM. — 1992. — Pp. 144–152.

110. Lightgbm: A highly efficient gradient boosting decision tree / G. Ke, Q. Meng, T. Finley et al. // Advances in Neural Information Processing Systems. — 2017. — Pp. 3146–3154.

111. *Dorogush A. V., Ershov V., Gulin A.* CatBoost: gradient boosting with categorical features support // *arXiv preprint arXiv:1810.11363*. — 2018.

112. SMOTE: synthetic minority over-sampling technique / Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, W Philip Kegelmeyer // *Journal of artificial intelligence research*. — 2002. — Vol. 16. — Pp. 321–357.

113. Руководство по разметке текстов (на основе Теории риторических структур) [Электронный ресурс]. — 2019. — URL: https://docs.google.com/document/d/1wd-sgGyIo5AQq2IPj6jWa_QmU0fUohXj48qsfVDgcBs (дата обращения: 04.02.2025).

114. *Muller Philippe, Braud Chloé, Morey Mathieu.* ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. — Minneapolis, MN: Association for Computational Linguistics, 2019. — Pp. 115–124.

115. The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification / Chloé Braud, Yang Janet Liu, Eleni Metheniti et al. // Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023). — Toronto, Canada: The Association for Computational Linguistics, 2023. — Pp. 1–21.

116. *Li Jing, Sun Aixin, Joty Shafiq.* SegBot: a generic neural text segmentation model with pointer network // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. — 2018. — Pp. 4166–4172.

117. *Wang Yizhong, Li Sujian, Yang Jingfeng.* Toward Fast and Accurate Neural Discourse Segmentation // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium: Association for Computational Linguistics, 2018. — Pp. 962–967.

118. The DISRPT 2019 Shared Task on Elementary Discourse Unit Segmentation and Connective Detection / Amir Zeldes, Debopam Das, Erick Galani Maziero et al. // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. — Minneapolis, MN: Association for Computational Linguistics, 2019. — Pp. 97–104.

119. Deep contextualized word representations / Matthew E Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of NAACL-HLT. — 2018. — Pp. 2227–2237.

120. *Wang Zhiguo, Hamza Wael, Florian Radu.* Bilateral multi-perspective matching for natural language sentences // Proceedings of the 26th International Joint Conference on Artificial Intelligence. — 2017. — Pp. 4144–4150.

121. On the Importance of Word and Sentence Representation Learning in Implicit Discourse Relation Classification / Xin Liu, Jiefu Ou, Yangqiu Song, Xin Jiang // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. — International Joint Conferences on Artificial Intelligence Organization, 2020. — Pp. 3830–3836.

122. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches / Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, Yoshua Bengio // Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. — Doha, Qatar: Association for Computational Linguistics, 2014. — Pp. 103–111.

123. *Soricut Radu, Marcu Daniel.* Sentence Level Discourse Parsing using Syntactic and Lexical Information // Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. — 2003. — Pp. 228–235.

124. *Straka Milan, Straková Jana.* Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe // Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — 2017. — Pp. 88–99.

125. Enriching Word Vectors with Subword Information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // *Transactions of the Association for Computational Linguistics.* — 2017. — Vol. 5. — Pp. 135–146.

126. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian / Anna Rogers, Alexey Romanov, Anna Rumshisky et al. // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. — Pp. 755–763.

127. Clustering-based undersampling in class-imbalanced data / Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, Jing-Shang Jhang // *Information Sciences.* — 2017. — Vol. 409. — Pp. 17–26.

128. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure / Yancui Li, Wenhe Feng, Jing Sun et al. // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Doha, Qatar: Association for Computational Linguistics, 2014. — Pp. 2105–2114.

129. *Cao Shuyuan.* How does discourse affect Spanish-Chinese Translation? A case study based on a Spanish-Chinese parallel corpus // Proceedings of

the First Workshop on Computational Approaches to Discourse. — Online: Association for Computational Linguistics, 2020. — Pp. 1–10.

130. *Vinyals Oriol, Fortunato Meire, Jaitly Navdeep.* Pointer networks // *Advances in neural information processing systems.* — 2015. — Vol. 28.

131. *Liu Shikun, Johns Edward, Davison Andrew J.* End-to-end multi-task learning with attention // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2019. — Pp. 1871–1880.

132. *Da Cunha Iria, Iruskieta Mikel.* Comparing rhetorical structures in different languages: The influence of translation strategies // *Discourse Studies.* — 2010. — Vol. 12, no. 5. — Pp. 563–598.

133. *Joty Shafiq, Carenini Giuseppe, Ng Raymond.* A Novel Discriminative Framework for Sentence-Level Discourse Analysis // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. — Jeju Island, Korea: Association for Computational Linguistics, 2012. — Pp. 904–915.

134. *Nejat Bita, Carenini Giuseppe, Ng Raymond.* Exploring Joint Neural Model for Sentence Level Discourse Parsing and Sentiment Analysis // Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. — Saarbrücken, Germany: Association for Computational Linguistics, 2017. — Pp. 289–298.

135. A Unified Linear-Time Framework for Sentence-Level Discourse Parsing / Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, M Saiful Bari // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy: Association for Computational Linguistics, 2019. — Pp. 4190–4200.

136. *Zhang Ying, Kamigaito Hidetaka, Okumura Manabu.* A Language Model-based Generative Classifier for Sentence-level Discourse Parsing //

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. — Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. — Pp. 2432–2446.

137. *Liu Yang Janet, Zeldes Amir.* Why Can't Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. — Dubrovnik, Croatia: Association for Computational Linguistics, 2023. — Pp. 3112–3130.

138. Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online: Association for Computational Linguistics, 2020. — Pp. 8440–8451.

139. *Kurfalı Murathan, Östling Robert.* Probing Multilingual Language Models for Discourse // Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). — Online: Association for Computational Linguistics, 2021. — Pp. 8–19.

140. DisCut and DiscReT: MELODI at DISRPT 2023 / Eleni Metheniti, Chloé Braud, Philippe Muller, Laura Rivière // Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023). — Toronto, Canada: The Association for Computational Linguistics, 2023. — Pp. 29–42.

141. *Surdeanu Mihai, Hicks Tom, Valenzuela-Escárcega Marco Antonio.* Two Practical Rhetorical Structure Theory Parsers // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. — Denver, Colorado: Association for Computational Linguistics, 2015. — Pp. 1–5.

142. *Hayashi Katsuhiko, Hirao Tsutomu, Nagata Masaaki.* Empirical comparison of dependency conversions for RST discourse trees // Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. — Los Angeles: Association for Computational Linguistics, 2016. — Pp. 128–136.

143. Neural Generative Rhetorical Structure Parsing / Amandla Mabona, Laura Rimell, Stephen Clark, Andreas Vlachos // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China: Association for Computational Linguistics, 2019. — Pp. 2284–2295.

144. RST Discourse Parsing with Second-Stage EDU-Level Pre-training / Nan Yu, Meishan Zhang, Guohong Fu, Min Zhang // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland: Association for Computational Linguistics, 2022. — Pp. 4269–4280.

145. *Zmitrovich Dmitry et al.* A family of pretrained transformer language models for Russian // *arXiv preprint arXiv:2309.10931.* — 2023.

146. *Kuratov Yuri, Arkhipov Mikhail.* Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — 2019. — Pp. 333–339.

147. *Costa-jussà Marta R et al.* No language left behind: Scaling human-centered machine translation // *arXiv preprint arXiv:2207.04672.* — 2022.

148. *Voll Kimberly, Taboada Maite.* Not all words are created equal: Extracting semantic orientation as a function of adjective relevance //

Australasian Joint Conference on Artificial Intelligence / Springer. — 2007. — Pp. 337–346.

149. *Hogenboom Alexander et al.* Using rhetorical structure in sentiment analysis // *Communications of the ACM*. — 2015. — Vol. 58, no. 7. — Pp. 69–77.

150. *Zirn Cäcilia et al.* Fine-grained sentiment analysis with structural features // Proceedings of 5th International Joint Conference on Natural Language Processing. — 2011. — Pp. 336–344.

151. *Kavuluru Ramakanth et al.* Classification of helpful comments on online suicide watch forums // Proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics. — 2016. — Pp. 32–40.

152. *Liu Xingyun, Liu Xiaoqian.* Online suicide identification in the framework of rhetorical structure theory (RST) // Healthcare / MDPI. — Vol. 9. — 2021. — P. 847.

153. *Rubin Victoria L, Lukoianova Tatiana.* Truth and deception at the rhetorical structure level // *Journal of the Association for Information Science and Technology*. — 2015. — Vol. 66, no. 5. — Pp. 905–917.

154. Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs / Zae Myung Kim, Kwang Lee, Preston Zhu et al. // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Bangkok, Thailand: Association for Computational Linguistics, 2024. — Pp. 5449–5474.

155. Under the Surface: Tracking the Artifactuality of LLM-Generated Data / Debarati Das, Karin De Langis, Anna Martin et al. // *arXiv preprint arXiv:2401.14698*. — 2024.

156. *Chernyavskiy Alexander, Ilvovsky Dmitry, Nakov Preslav.* Unleashing the Power of Discourse-Enhanced Transformers for Propaganda Detection // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). — St. Julian's, Malta: Association for Computational Linguistics, 2024. — Pp. 1452–1462.

157. *Peldszus Andreas, Stede Manfred.* Rhetorical structure and argumentation structure in monologue text // Proceedings of the Third Workshop on Argument Mining (ArgMining2016). — Berlin, Germany: Association for Computational Linguistics, 2016. — Pp. 103–112.

158. *Bhatia Parminder, Ji Yangfeng, Eisenstein Jacob.* Better Document-level Sentiment Analysis from RST Discourse Parsing // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Lisbon, Portugal: Association for Computational Linguistics, 2015. — Pp. 2212–2218.

159. Long Short-term Memory Network over Rhetorical Structure Theory for Sentence-level Sentiment Analysis / Xianghua Fu, Wangwang Liu, Yingying Xu et al. // Proceedings of The 8th Asian Conference on Machine Learning / PMLR. — 2016. — Pp. 17–32.

160. *Tai Kai Sheng, Socher Richard, Manning Christopher D.* Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Beijing, China: Association for Computational Linguistics, 2015. — Pp. 1556–1566.

161. *Huber Patrick, Carenini Giuseppe.* From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation // Proceedings of the

28th International Conference on Computational Linguistics. — 2020. — Pp. 185–197.

162. *Kraus Mathias, Feuerriegel Stefan.* Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees // *Expert Systems with Applications.* — 2019. — Vol. 118. — Pp. 65–79.

163. *Koto Fajri, Lau Jey Han, Baldwin Timothy.* Discourse Probing of Pretrained Language Models // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 2021. — Pp. 3849–3864.

164. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions / Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan et al. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China: Association for Computational Linguistics, 2019. — Pp. 2933–2943.

165. *Eckle-Kohler Judith, Kluge Roland, Gurevych Iryna.* On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Lisbon, Portugal: Association for Computational Linguistics, 2015. — Pp. 2236–2242.

166. A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations / Elena Musi, Manfred Stede, Leonard Kriese et al. // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

167. *Akiba Takuya et al.* Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. — 2019. — Pp. 2623–2631.

168. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution / Sam Wiseman, Alexander M. Rush, Stuart Shieber, Jason Weston // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Beijing, China: Association for Computational Linguistics, 2015. — Pp. 1416–1426.

169. *Wiseman Sam, Rush Alexander M., Shieber Stuart M.* Learning Global Features for Coreference Resolution // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — San Diego, California: Association for Computational Linguistics, 2016. — Pp. 994–1004.

170. *Moosavi Nafise Sadat, Strube Michael.* Use Generalized Representations, But Do Not Forget Surface Features // Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017). — Valencia, Spain: Association for Computational Linguistics, 2017. — Pp. 1–7.

171. *Kahardipraja Patrick, Vyshnevska Olena, Loáiciga Sharid.* Exploring Span Representations in Neural Coreference Resolution // Proceedings of the First Workshop on Computational Approaches to Discourse. — Online: Association for Computational Linguistics, 2020. — Pp. 32–41.

172. *Dobrovolskii Vladimir.* Word-Level Coreference Resolution // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. — Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. — Pp. 7670–7675.

173. *Grenander Matt, Cohen Shay B., Steedman Mark.* Sentence-Incremental Neural Coreference Resolution // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. — Pp. 427–443.

174. *Kibrik Andrej A.* Cognitive inferences from discourse observations: reference and working memory // Discourse studies in cognitive linguistics. Proceedings of the 5th International cognitive linguistics conference. — 1999. — Pp. 29–52.

175. *Khosla Sopan, Fiacco James, Rosé Carolyn.* Evaluating the Impact of a Hierarchical Discourse Representation on Entity Coreference Resolution Performance // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 2021. — Pp. 1645–1651.

176. *Ri Ryokan, Yamada Ikuya, Tsuruoka Yoshimasa.* mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland: Association for Computational Linguistics, 2022. — Pp. 7316–7330.

177. End-to-end Neural Coreference Resolution / Kenton Lee, Luheng He, Mike Lewis, Luke Zettlemoyer // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark: Association for Computational Linguistics, 2017. — Pp. 188–197.

178. *Lee Kenton, He Luheng, Zettlemoyer Luke.* Higher-Order Coreference Resolution with Coarse-to-Fine Inference // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New

Orleans, Louisiana: Association for Computational Linguistics, 2018. — Pp. 687–692.

179. *Kirstain Yuval, Ram Ori, Levy Omer.* Coreference Resolution without Span Representations // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). — Online: Association for Computational Linguistics, 2021. — Pp. 14–19.

180. Scaling Within Document Coreference to Long Texts / Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar et al. // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. — Online: Association for Computational Linguistics, 2021. — Pp. 3921–3931.

181. NARC – Norwegian Anaphora Resolution Corpus / Petter Mæhlum, Dag Haug, Tollef Jørgensen et al. // Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference. — Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022. — Pp. 48–60.

182. *Dobrovolskii Vladimir, Michurina Mariia, Ivoylova Alexandra.* RuCoCo: a new Russian corpus with coreference annotation // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". — 2022.

183. *Moosavi Nafise Sadat, Strube Michael.* Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Berlin, Germany: Association for Computational Linguistics, 2016. — Pp. 632–642.

184. Ru-eval-2019: Evaluating anaphora and coreference resolution for Russian / A. E. Budnikov, S. Yu. Toldova, D. S. Zvereva et al. //

Computational Linguistics and Intellectual Technologies-Supplementary Volume. — 2019.

185. A Model-Theoretic Coreference Scoring Scheme / Marc Vilain, John Burger, John Aberdeen et al. // Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. — 1995.

186. *Bagga Amit, Baldwin Breck*. Algorithms for scoring coreference chains // The first international conference on language resources and evaluation workshop on linguistics coreference / Citeseer. — Vol. 1. — 1998. — Pp. 563–566.

187. *Luo Xiaoqiang*. On Coreference Resolution Performance Metrics // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. — Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005. — Pp. 25–32.

188. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes / Sameer Pradhan, Lance Ramshaw, Mitchell Marcus et al. // Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. — Portland, Oregon, USA: Association for Computational Linguistics, 2011. — Pp. 1–27.

189. *Gutnik Gleb*. Experiments on Adaptation of End-to-end Coreference Resolution Models For Russian. — Dialogue 2022 Student Session. — 2022.

190. *Azar Moshe*. Argumentative text as rhetorical structure: An application of rhetorical structure theory // *Argumentation*. — 1999. — Vol. 13, no. 1. — Pp. 97–114.

191. *Villalba Maria Paz Garcia, Saint-Dizier Patrick.* Some Facets of Argument Mining for Opinion Analysis. // *COMMA.* — 2012. — Vol. 245. — Pp. 23–34.

192. *Green Nancy L.* Representation of argumentation in text with rhetorical structure theory // *Argumentation.* — 2010. — Vol. 24, no. 2. — Pp. 181–196.

193. The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error / Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, Malihe Alikhani // Findings of the Association for Computational Linguistics: ACL 2022. — Dublin, Ireland: Association for Computational Linguistics, 2022. — Pp. 824–845.

194. *Peldszus Andreas, Stede Manfred.* An annotated corpus of argumentative microtexts // Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon. — Vol. 2. — 2015. — Pp. 801–815.

195. *Peldszus Andreas, Stede Manfred.* Joint prediction in MST-style discourse parsing for argumentation mining // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — Lisbon, Portugal: Association for Computational Linguistics, 2015. — Pp. 938–948.

196. *Dozat Timothy, Manning Christopher D.* Deep biaffine attention for neural dependency parsing // *arXiv preprint arXiv:1611.01734.* — 2016.

197. *Fishcheva Irina, Kotelnikov Evgeny.* Cross-lingual argumentation mining for Russian texts // International Conference on Analysis of Images, Social Networks and Texts. — 2019. — Pp. 134–144.

198. RST-Tace A tool for automatic comparison and evaluation of RST trees / Shujun Wan, Tino Kutschbach, Anke Lüdeling, Manfred Stede // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019.

— Minneapolis, MN: Association for Computational Linguistics, 2019. — Pp. 88–96.

199. *Skeppstedt Maria, Peldszus Andreas, Stede Manfred.* More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing // Proceedings of the 5th Workshop on Argument Mining. — Brussels, Belgium: Association for Computational Linguistics, 2018. — Pp. 155–163.

200. *He Pengcheng, Gao Jianfeng, Chen Weizhu.* Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing // *arXiv preprint arXiv:2111.09543.* — 2021.

201. Class-based n-gram models of natural language / Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza et al. // *Computational linguistics.* — 1992. — Vol. 18, no. 4. — Pp. 467–480.

Приложение А. Результаты интеллектуальной деятельности

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024618391

Программа для анализа дискурсивной риторической структуры текстов

Правообладатель: ***Общество с ограниченной ответственностью «РИ Технологии» (RU)***

Автор(ы): ***Чистова Елена Викторовна (RU)***

Заявка № **2024616947**
Дата поступления **04 апреля 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***11 апреля 2024 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 429b6a0fe3853164baf96f83b73b4aa7
Владелец Зубов Юрий Сергеевич
Действителен с 10.05.2023 по 02.08.2024

Ю.С. Зубов

Приложение Б. Справка об использовании

Общество с ограниченной ответственностью
«РИ ТЕХНОЛОГИИ»
121205, Москва, инновационный Центр «Сколково»,
Большой бульвар, д. 42, стр.1.
Тел./факс: +7(499) 135-51-45 www.ritech.ru

Исх. № 120-25 от 06.02.2025г.
на № _____ от _____

СПРАВКА

об использовании результатов диссертации
Чистовой Елены Викторовны
на соискание ученой степени кандидата технических наук на тему
«Методы анализа риторической структуры текстов на русском языке»

Настоящая справка подтверждает, что результаты диссертационной работы, полученные Чистовой Еленой Викторовной, были использованы в ООО «РИ ТЕХНОЛОГИИ» при разработке программных систем анализа текстов.

Разработанные в диссертации методы анализа риторической структуры текстов на русском языке использовались в ООО «РИ ТЕХНОЛОГИИ» при создании системы лингво-статистического анализа коллекций текстов и позволили реализовать в системе новые поисковые и аналитические функции, связанные с обработкой верхнеуровневых представлений текстов.

Предложенные в диссертации методы анализа риторической структуры текстов на русском языке реализованы Чистовой Еленой Викторовной в программе для анализа дискурсивной риторической структуры текстов, которая используется в разрабатываемых ООО «РИ ТЕХНОЛОГИИ» программных средствах психолингвистического анализа текстов.

Планируется использование результатов диссертации Чистовой Елены Викторовны в интеллектуальной цифровой платформе агрегации и анализа научно-технической информации SciApp.

Генеральный директор Е.А. Рыбкина