На правах рукописи

Ватолин Алексей Сергеевич

Обучение и оценивание мультиязычных нейросетевых моделей семантического векторного представления научных текстов

Специальность 2.3.8 —
«Информатика и информационные процессы»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2026

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН).

Научный руководитель: доктор физико-математических наук, доцент **Абгарян Каринэ Карленовна**

Официальные оппоненты: **Котельников Евгений Вячеславович,** доктор технических наук, доцент, профессор Школы вычислительных социальных наук Автономной некоммерческой образовательной организации высшего образования «Европейский университет в Санкт-Петербурге»

Кузнецов Сергей Олегович, доктор физико-математических наук, профессор Национального исследовательского университета «Высшая школа экономики»

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук

Защита состоится «____» _____ 20___г. в _____:_____ на заседании диссертационного совета 24.1.224.03 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу: 119333, Россия, Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук и на сайте http://www.frccsc.ru.

Автореферат разослан «____» _____ 2026 года.

Ученый секретарь
диссертационного совета
24.1.224.03,
к.т.н. Рейер Иван Александрович

Общая характеристика работы

Актуальность темы. Современная наука характеризуется стремительным увеличением объемов публикуемой информации, что создает значительные трудности для исследователей в поиске и анализе релевантных данных. Эта тенденция подтверждается динамикой наполнения крупнейших библиометрических баз данных. Так, согласно анализу базы данных Scopus, количество ежегодно индексируемых научных статей выросло с приблизительно 921 тысячи в 2000 году до более чем 2,57 миллиона в 2020 году, что свидетельствует о почти трехкратном увеличении за два десятилетия (Thelwall, 2022). Общий объем публикаций в Scopus к 2020 году превысил 56 миллионов единиц. Аналогичные тенденции наблюдаются и в других международных наукометрических системах, таких как Web of Science. Российская научная электронная библиотека eLibrary.ru, которая представляет материалы как на русском, так и на других языках, также демонстрирует значительный рост: количество публикаций в год увеличилось с 45,7 тысяч в 2000 году до более чем 4,76 миллиона в 2020 году (ООО Научная электронная библиотека, 2025). Количество публикаций увеличивается не только на английском языке. Этот информационный поток делает задачу поиска релевантной информации для исследователей всё более трудной, а также ставит новые вызовы по эффективной обработке и анализу постоянно растущих текстовых массивов, что требует применения высокопроизводительных подходов, в том числе методов параллельной обработки данных (Бажанов, 2021).

Статистические методы (VSM, BM25) остаются широко используемыми методами информационного поиска, однако в задачах семантического анализа научных текстов все большую роль играют нейросетевые подходы (Salton, 1975) (Robertson, 2009) (Vaswani, 2017). Архитектура трансформер лежит в основе значительной части современных нейросетевых моделей обработки текста (Vaswani, 2017). Такие модели способны учитывать контекст, улавливать семантические связи между терминами, работать с синонимией и осуществлять эффективный многоязычный и кросс-язычный поиск. Внедрение таких моделей в научно-информационные системы позволяет существенно повысить качество и релевантность поиска, а также открывает новые возможности для анализа содержания научных текстов (Jin, 2023) (Wang, 2022). В данной диссертации представлены нейросетевые модели SciRus [1], которые

позволяют упростить работу с научными публикациями. Одна из разработанных моделей, SciRus-tiny, была внедрена на сервисе eLibrary.ru.

Для объективной оценки и совершенствования моделей обработки естественного языка используют бенчмарки — специализированные наборы данных и задач. В последние годы для русского языка появились такие универсальные инструменты оценки, как RuSentEval (Mikhailov, 2021) и encodechka (Dale, 2022). Несмотря на это, наблюдается дефицит инструментов для русскоязычного научного домена. Научные тексты обладают рядом специфических характеристик, таких как информационная плотность и сложность текста, узкоспециализированная терминология, особая структура и стиль изложения. Из-за этого оценка на универсальных наборах для оценки не дает понимания о качестве работы модели на научных данных.

Отсутствие специализированных русскоязычных научных наборов данных для оценки затрудняет сравнительный анализ как существующих, так и разрабатываемых моделей векторного представления текстов в данной предметной области. Как следствие, это сдерживает дальнейшее развитие алгоритмов для анализа русскоязычного научного контента. Еще одна задача таких наборов — помочь исследователям подобрать подходящую модель для своей задачи, связанной с научными текстами.

Открытые модели SciRus и бенчмарк RuSciBench дают исследователям готовые средства для анализа русско- и англоязычных научных текстов и сопоставимой оценки новых моделей.

Целью данной работы является разработка, исследование и апробация инструментов и моделей, предназначенных для решения задач эффективной обработки, анализа и оценки качества представления научных текстов на русском языке.

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

1. Разработать методику обучения легковесной двуязычной модели для эффективного векторного представления научных текстов на русском и английском языках. Исследовать подходы к обучению, основанные на доступных данных из мультиязычных корпусов, без дополнительной разметки.
2. Разработать методологию и на ее основе создать инструментарий для оценки качества векторных представлений научных текстов на русском и английском языках. Данный инструментарий должен учитывать специфику научного дискурса и охватывать

разнообразные задачи, используя данные из российской научной среды.

3. Исследовать проблему верификации научных фактов на русском языке. Разработать и апробировать методологию полуавтоматизированного формирования русскоязычного набора данных, включающую генерацию научных утверждений на основе аннотаций с использованием больших языковых моделей и их последующую экспертную валидацию. Разработать тестовый набор на основе данного набора для оценки способности моделей определять соответствие или противоречие утверждений.

Основные положения, выносимые на защиту:

1. Предложены компактные двуязычные модели SciRus-tiny (23 млн параметров) и SciRus-small (61 млн параметров) для представления научных текстов в векторном пространстве. Обучение проводится в два этапа: сначала модель обучается с помощью маскированного языкового моделирования, затем с помощью контрастивного дообучения на парах «заголовок-аннотация». Дополнительно, для формирования более сильного обучающего сигнала, при обучении используются пары «цитирующая статья — цитируемая статья», основанные на однонаправленной связи из графа цитирований. Эксперименты на русскоязычных и англоязычных научных наборах для оценки показали, что при числе параметров, в 24 и 9 раз меньшем по сравнению с multilingual-e5-large-instruct (560 млн параметров), предлагаемые модели демонстрируют сопоставимый, а на русскоязычных задачах — превосходящий уровень качества.
2. Разработан мультизадачный двуязычный бенчмарк RuSciBench, включающий задачи классификации, регрессии, моно- и кросс-языкового поиска на научных данных. Этот тестовый набор обеспечивает воспроизводимую процедуру тестирования, интегрированную в международный лидерборд MTEB.
3. Предложена полуавтоматическая методика формирования наборов данных для проверки научных фактов на русском языке, сочетающая генерацию утверждений с помощью LLM, многоступенчатую самооценку модели и экспертную верификацию. На её основе создан первый русскоязычный набор данных для проверки научных фактов RuSciFact.

Методы исследования. В диссертационном исследовании использованы известные, достоверные и хорошо зарекомендовавшие себя на практике методы. В модели используется архитектура трансформер, она обучается с помощью маскированного языкового моделирования и с помощью контрастивного дообучения с применением функции потерь InfoNCE. В наборах для оценки используются распространенные критерии оценки качества, такие как Accuracy, F1-мера, NDCG@k, MRR@k, коэффициент корреляции Кендалла.

Научная новизна:

1. Показана эффективность контрастивного дообучения модели векторизации текста на парах «заголовок-аннотация» и на парах цитирующая-цитируемая статья.
2. Разработаны легковесные модели векторизации научных текстов.
3. Разработан первый набор для оценки качества работы моделей с научными данными, состоящий из различных типов задач.
4. Впервые предложена полуавтоматизированная многоступенчатая методика формирования наборов данных для проверки научных фактов на русском языке, совмещающая генерацию утверждений с помощью LLM, самокритичную оценку модели-генератора и экспертную валидацию.
5. Разработан и опубликован первый набор данных для проверки научных фактов на русском языке RuSciFact.

Практическая значимость обусловлена разработкой открытых научных наборов для оценки, которые были внедрены в авторитетный международный бенчмарк MTEB (Massive Text Embedding Benchmark), что подтверждает их актуальность, а также существенно упрощает их использование для разработчиков моделей. Кроме того, данные RuSciBench были использованы при формировании одной из задач мультиязычного набора для оценки AIRBench (Chen, 2024) и одной из задач русскоязычного тестового набора LIBRA (Churin, 2024). Это свидетельствует о востребованности полученных результатов в научном сообществе.

Практическая значимость работы также подтверждается внедрением модели векторизации научных текстов SciRus-tiny на российском научном портале eLibrary.ru. На основе этой модели был разработан режим «нейропоиск», предназначенный для поиска тематически близких научных публикаций по аннотации статьи. Такое внедрение расширяет

возможности поиска и анализа научной информации для исследователей и специалистов, работающих с научной библиотекой eLibrary.ru.

Теоретическая значимость работы заключается в развитии методических основ обучения и оценивания мультиязычных нейросетевых моделей семантического векторного представления научных текстов в условиях доменной и языковой специфики. В работе систематизированы и экспериментально исследованы факторы, влияющие на качество таких моделей в русско- и англоязычном научном домене: использование доменно-специфического корпуса, контрастивного обучения на парах «заголовок–аннотация», обучающего сигнала из графа цитирований, размера модели и степени ее языковой специализации. На основе разработанного двуязычного мультизадачного бенчмарка количественно оценены различия в качестве работы современных моделей на русском и английском языках, что уточняет границы применимости универсальных мультиязычных трансформерных моделей к специализированным научным текстам. Дополнительный теоретический вклад состоит в предложенной методологии формирования и анализа наборов данных для верификации научных фактов на русском языке, позволяющей разделять способность моделей к семантическому поиску и способность к логическому сопоставлению научного утверждения с текстом аннотации.

Достоверность полученных результатов подтверждается следующим:

1. докладами и обсуждениями результатов на международных конференциях
2. публикациями результатов в рецензируемых научных изданиях, рекомендованных ВАК
3. открытым исходным кодом и воспроизводимостью результатов.

Апробация работы. Основные результаты работы докладывались на:

1. А. С. Ватолин. Сравнительный анализ современных мультиязычных моделей для векторизации текста на русском языке. *Международная научно-практическая конференция «Информационные технологии, искусственный интеллект, большие данные: актуальные тенденции, перспективные исследования»*, 2024

2. А. С. Ватолин. ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian. *Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог»*, 2025
3. А. С. Ватолин. Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024. *Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог»*, 2025

Личный вклад соискателя в работах с соавторами заключается в следующем. В работе [1] выполнены предобучение маленькой версии модели (SciRus-tiny) с помощью маскированного языкового моделирования, контрастивное дообучение моделей SciRus-tiny и SciRus-small в двух режимах, только на парах «заголовок–аннотация» и на объединении таких пар с парами «цитирующая статья–цитируемая статья», а также валидация моделей на наборе данных SciDocs. В работе [2] выполнены сбор датасетов для классификации по ГРНТИ, по типу публикации, для поиска цитирований, для регрессии по количеству цитат, для поиска английского перевода по тексту на русском языке, реализация исходного кода инструментария для оценки, валидация моделей и интеграция бенчмарка в международный бенчмарк MTEB. В работе [3] вклад соискателя является определяющим. В работе [5] в рамках создания и расширения международного многоязычного бенчмарка MMTEB соискателем проведена работа по добавлению новых задач на различных языках, включая задачи, разработанные им самостоятельно. Также выполнен значительный вклад в обеспечение качества и корректности данных во всех задачах, вошедших в итоговый набор бенчмарка. В публикации [4] соискатель является единственным автором.

Содержание диссертации и положения, выносимые на защиту, отражают персональный вклад автора в опубликованных работах. Все представленные результаты получены лично автором либо при его определяющем участии.

Публикации. Основные результаты по теме диссертации изложены в 5 печатных изданиях: 2 — в рецензируемых научных журналах, индексируемых в Scopus и РИНЦ, 3 — в полнотекстовых статьях в

рецензируемых трудах международных конференций, индексируемых в Scopus.

Соответствие диссертации паспорту научной специальности. Основные результаты диссертации соответствуют направлениям исследований паспорта научной специальности 2.3.8 «Информатика и информационные процессы», указанным в пунктах 4, 7 и 12.

Объем и структура работы. Диссертация состоит из введения, 4 глав, заключения и 1 приложения. Полный объём диссертации составляет 134 страницы, включая 9 рисунков и 23 таблицы. Список литературы содержит 76 наименований.

Содержание работы

Во введении обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, формулируются научная новизна и практическая значимость представляемой работы.

В первой главе диссертации систематизированы подходы к построению семантических векторных представлений текстов и выделены ограничения, определяющие постановку дальнейшего исследования. Анализ сфокусирован на том, какие свойства моделей необходимы для научного домена: устойчивость к терминологической насыщенности, переносимость между дисциплинами и способность учитывать связи публикаций, не сводимые к прямому лексическому совпадению.

В **разделах 1.1–1.4** показана эволюция методов от статистических представлений и матричных разложений к трансформер-кодировщикам, предобучаемым на больших корпусах. Для диссертации существенен не сам общеизвестный механизм BERT/RoBERTa, а его роль как базовой архитектуры, допускающей доменную адаптацию: предобучение на неразмеченных текстах формирует языковые представления, но без последующего специализированного обучения такие представления недостаточно пригодны для семантического поиска и сопоставления научных документов.

В **разделе 1.5** обоснован выбор биэнкодерной схемы, позволяющей заранее индексировать научный корпус и использовать один кодировщик для поиска, классификации, регрессии и кросс-языкового сопоставления. Такой подход согласуется с прикладной задачей работы, поэтому для

формирования семантического пространства используется контрастивное обучение.

В **разделе 1.6** проанализированы специализированные модели для научной области, включая **SPECTER**, **SPECTER2** и **SciNCL**. Эти решения используют структуру научной коммуникации, прежде всего графы цитирований, и демонстрируют преимущество доменно адаптированных представлений над универсальными языковыми моделями. Одновременно выявлено их принципиальное ограничение для задач настоящей работы: передовые специализированные модели и связанные с ними обучающие сигналы ориентированы преимущественно на англоязычный корпус и не обеспечивают полноценной поддержки русскоязычных и русско-английских научных коллекций.

В **разделе 1.7** рассмотрены инструменты оценки качества семантических представлений: **SciDocs**, **SciRepEval**, **SciFact** и **MTEB**. Их анализ показывает, что существующая инфраструктура хорошо покрывает англоязычные и общеязыковые сценарии, но не дает стандартизированной проверки качества моделей на русскоязычных научных текстах, на параллельных русско-английских метаданных и на задачах, характерных для российской научной информационной среды.

В **заключительном разделе 1.8** сформулирован научный пробел, закрываемый диссертацией: отсутствуют компактные двуязычные модели семантической векторизации научных документов, обученные с учетом русско-английского корпуса и структурных связей публикаций, а также отсутствует открытый инструментарий для их воспроизводимой оценки на русскоязычных и кросс-языковых научных задачах.

Во второй главе диссертации описывается процесс разработки, обучения и оценки семейства двуязычных моделей SciRus, предназначенных для построения семантических векторных представлений научных текстов. Основной целью работы является преодоление двух ключевых ограничений существующих решений: англоязычности передовых специализированных моделей (SPECTER, SciNCL) и высокой вычислительной ресурсоемкости универсальных моделей.

В **разделе 2.1** задача векторизации научных текстов задана формально. Для коллекции документов $\mathcal{D}$ требуется построить параметризованный кодировщик $f(\cdot,\alpha)$, сопоставляющий каждому документу плотное представление в общем векторном пространстве:

$$\mathbf{v}_i = f(x_i,\alpha) \in \mathbb{R}^d, \quad x_i \in \mathcal{D}.$$

Практическая ценность такого отображения определяется не самим фактом получения векторов, а их пригодностью для поиска, классификации, регрессии и кросс-языкового сопоставления. Поэтому центральный этап обучения задаётся сближением положительных пар научных документов и отделением их от нерелевантных примеров:

$$\sum_{i=1}^{N} \mathcal{L}_{\text{Contr}}\big(f(x_{a,i},\alpha),f(x_{p,i},\alpha)\big) \to \min_{\alpha}.$$

На этапе оценки параметры кодировщика фиксируются, а для конкретной прикладной задачи обучается только малый набор параметров β':

$$\sum_{i=1}^{M} \mathcal{L}'_i\big(g'(f(x'_i,\alpha),\beta'),y'_i\big) \to \min_{\beta'}, \quad \dim(\beta') \ll \dim(\alpha).$$

В **разделе 2.2** описываются наборы данных для обучения моделей SciRus: международный корпус **Semantic Scholar Open Research Corpus (S2ORC)** (Lo, 2020) и данные российской научной библиотеки **eLibrary.ru**. Из S2ORC были использованы мультиязычные пары «заголовок–аннотация» и граф цитирований, служивший источником структурной информации для контрастивного дообучения.

Данные eLibrary.ru дополнили корпус большим массивом русскоязычных и англоязычных научных текстов, включая параллельные русско-английские аннотации для формирования единого семантического пространства двух языков. Граф цитирований eLibrary.ru также использовался при построении обучающих пар. Итоговый обучающий корпус объединил около 48.2 миллионов пар «заголовок–аннотация», что соответствует примерно 15 миллиардам токенов.

Раздел 2.3 посвящен архитектуре разработанных моделей. За основу взята архитектура RoBERTa, детально описанная в предыдущей главе. В рамках работы были созданы две конфигурации: **SciRus-tiny** (23 млн параметров, размерность вектора $d = 312$) и **SciRus-small** (61 млн параметров, $d = 768$). Обе модели используют $L = 3$ трансформер-блока, $H = 12$ голов внимания и общий токенизатор BPE со словарем на 50265 токенов.

В **разделе 2.4** описывается процесс обучения моделей SciRus, состоявший из двух последовательных этапов: предобучение с использованием задачи маскированного языкового моделирования (MLM) и

последующее контрастивное дообучение для формирования семантического векторного пространства.

Первый этап, подробно изложенный в **подразделе 2.4.1**, заключался в предобучении моделей с нуля. Такой подход был выбран ввиду отсутствия готовых моделей сопоставимого размера с поддержкой русского и английского языков, а также поскольку токенизаторы существующих моделей были обучены на неспециализированных данных и не были оптимальными для научных текстов. Модели инициализировались случайными весами, а в качестве обучающих данных использовался объединенный текстовый корпус, состоящий из заголовков и аннотаций из наборов данных S2ORC и eLibrary.ru. Обучение проводилось с использованием задачи маскированного языкового моделирования (MLM) (Devlin, 2019). Процесс предобучения продолжался в течение двух эпох, и контроль сходимости осуществлялся на валидационной выборке. По завершении этого этапа линейный слой с параметрами β отбрасывается.

Второй этап, описанный в **подразделе 2.4.2**, состоял в дообучении моделей с использованием контрастивного подхода. Для формирования обучающих примеров использовались данные двух типов. Первый тип основан на парах «заголовок–аннотация» и исходит из предположения, что заголовок научной статьи семантически близок к ее собственной аннотации, но не близок к аннотации случайно выбранной статьи. Второй тип данных основан на графах цитирования S2ORC и eLibrary.ru, при этом цитирующая статья считается близкой по содержанию к цитируемой, но далекой от случайной статьи из корпуса. На основе этих двух предположений формировались положительные пары $(x_{a,i}, x_{p,i})$. Также применялась кросс-языковая стратегия формирования пар: опорный пример $x_{a,i}$ и положительный пример $x_{p,i}$ могли быть представлены на разных языках (например, русскоязычный заголовок и англоязычная аннотация), если такие данные доступны, что способствовало формированию единого семантического пространства для русского и английского языков.

Для получения единого вектора документа применялось усреднение токеновых представлений последнего слоя кодировщика. Контрастивное обучение выполнялось с функцией потерь InfoNCE (van den Oord, 2018) при температурном коэффициенте $\tau = 0.01$.

В результате были получены две линейки моделей. Модели, обученные исключительно на парах «заголовок–аннотация», именуются

`SciRus-tiny` и `SciRus-small`. Модели, при обучении которых дополнительно использовались данные о цитированиях, получили суффикс `-cite`: `SciRus-tiny-cite` и `SciRus-small-cite`.

В **разделе 2.5** представлены результаты оценки качества разработанных моделей на общепринятом англоязычном бенчмарке **SciDocs** (Cohan, 2020) в сравнении с ведущими мировыми аналогами. В таблице 1 приведена выдержка из результатов. Модели сравниваются по среднему значению меры качества по 6 задачам бенчмарка, как было предложено в оригинальной публикации.

Таблица 1 — Сравнение моделей на бенчмарке SciDocs (среднее значение мер качества по всем задачам)

Модель	**Количество параметров**	**Языки**	**Среднее**
SciRus-tiny	23 млн	русско-английский	86.53
SciRus-small	61 млн	русско-английский	86.89
SciRus-tiny-cite	23 млн	русско-английский	89.55
SciRus-small-cite	**61 млн**	**русско-английский**	**90.02**
multilingual-e5-small	118 млн	мультиязычная	86.23
SPECTER	110 млн	английский	89.10
multilingual-e5-large-instruct	560 млн	мультиязычная	89.18
GritLM-7B	7.24 млрд	мультиязычная	89.70
SciNCL	110 млн	английский	90.84
GIST-large-Embedding-v0	335 млн	английский	**91.26**

Анализ результатов показывает, что модели SciRus, обученные с использованием данных о цитированиях, демонстрируют высокую конкурентоспособность. Модель `SciRus-small-cite` (61 млн параметров) достигает качества, сопоставимого со специализированной англоязычной моделью `SciNCL` (110 млн), и превосходит как более раннюю модель `SPECTER`, так и значительно более крупную модель общего назначения `GritLM-7B`.

Модели, обученные только на парах «заголовок-аннотация» без использования графа цитирований (`SciRus-tiny` и `SciRus-small`), показывают более низкое качество, однако сохраняют приемлемый уровень результатов. Этот результат имеет важное практическое значение: пары заголовков и аннотаций значительно более доступны, чем данные о цитированиях, которые требуют наличия полного графа научных публикаций. Дополнительное использование пар «цитирующая статья–цитируемая статья» исследовано отдельно: на бенчмарке SciDocs оно повышает среднее качество примерно на 3 процентных пункта.

Важно подчеркнуть, что в отличие от англоязычных моделей `SPECTER` и `SciNCL`, модели SciRus являются двуязычными.

Раздел 2.6 посвящен оценке вычислительной эффективности. Были проведены замеры скорости векторизации текстов на центральном процессоре. Результаты показывают, что модель `SciRus-small` обрабатывает запросы примерно в 15 раз быстрее, чем `SciNCL`, и в 20 раз быстрее, чем `SPECTER`, что является критически важным для применения в промышленных системах.

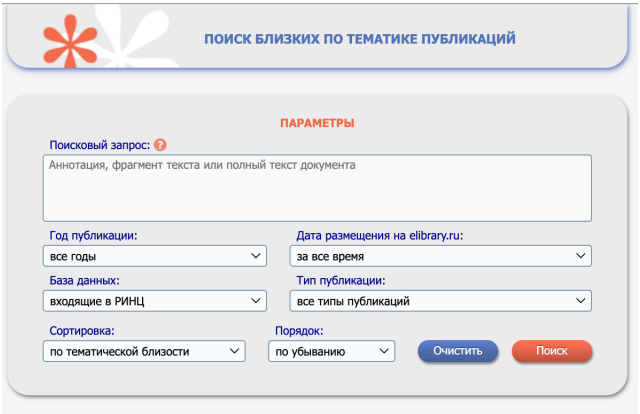

Рисунок 1 — Интерфейс режима «нейропоиск» на портале eLibrary.ru, использующего модель SciRus-tiny.

Практическая значимость работы подтверждается в **разделе 2.7**, где описывается внедрение модели `SciRus-tiny` в информационно-поисковую систему научной электронной библиотеки **eLibrary.ru**. На

основе разработанной модели был запущен сервис семантического поиска «нейропоиск», позволяющий пользователям находить релевантные публикации (см. рис. 1).

В третьей главе диссертации представлен **RuSciBench** [2] — мультизадачный двуязычный бенчмарк для оценки моделей векторизации научных текстов. Он основан на корпусе из 182 264 статей eLibrary.ru, для которых доступны русско-английские пары заголовков и аннотаций.

Полная версия RuSciBench включает 18 сочетаний «задача–язык/направление» для классификации, регрессии, информационного поиска и кросс-языкового поиска. В диссертации подробно описаны разработанные лично компоненты: две задачи классификации, одна задача регрессии, одна задача информационного поиска в русской и английской версиях, а также поиск английского перевода по русскому тексту.

В **разделах 3.1 и 3.2** описано, как корпус после очистки, дедупликации и проверки языка используется для оценки моделей: параметры кодировщика фиксируются, а полученные векторные представления подаются в модели малой сложности.

В **подразделе 3.2.1** описаны задачи классификации, в которых качество векторных представлений проверяется через обучение логистической регрессии поверх замороженного кодировщика. Для устранения влияния дисбаланса классов применялась балансировка выборок, а основной мерой качества была точность (Accuracy). Бенчмарк включает две задачи этого типа:

– **Классификация по рубрикатору ГРНТИ** - 29 классов.
– **Классификация по типу публикации** - 4 класса.

Подраздел 3.2.2 посвящен регрессионной задаче предсказания числа цитирований по векторному представлению аннотации. Параметры кодировщика при оценке не изменяются, а поверх его выходов обучается линейная регрессионная модель. Качество измеряется коэффициентом ранговой корреляции Кендалла τ_b, поскольку распределение цитирований резко скошено, содержит выбросы и требует оценки не абсолютной ошибки, а способности модели воспроизводить относительный порядок публикаций по признакам научной влиятельности.

Подраздел 3.2.3 описывает задачу информационного поиска прямых цитирований: по заголовку и аннотации статьи-запроса необходимо найти в общем корпусе публикации, на которые она ссылается. Запросы и документы векторизуются оцениваемой моделью, ранжируются по косинусной близости, а качество списка измеряется NDCG@10, что позволяет

учитывать не только факт обнаружения релевантной публикации, но и ее позицию в выдаче.

Подраздел 3.2.4 посвящен кросс-языковому поиску точного перевода аннотации: для каждого русскоязычного запроса модель должна найти соответствующую англоязычную аннотацию в параллельном корпусе. Эта задача проверяет не тематическую близость внутри одного языка, а согласованность русского и английского подпространств. Качество измеряется долей правильно найденных переводов.

В **разделе 3.3** представлены итоговые результаты оценки 29 моделей, включая ряд передовых решений, опубликованных уже после выхода бенчмарка. Для агрегирования результатов по 18 задачам используется метод Борда, основанный на суммировании рангов моделей по отдельным задачам. В таблице 2 представлен сокращенный итоговый рейтинг, включающий первые восемь моделей, а также все модели семейства SciRus для наглядного сопоставления.

Таблица 2 — Сокращенный сводный рейтинг моделей на RuSciBench

Модель	**Количество параметров**	**Ранг Борда**	**Среднее**
GritLM-7B	7.24 млрд	1	0.4687
Linq-Embed-Mistral	7.00 млрд	2	0.4574
SFR-Embedding-Mistral	7.00 млрд	3	0.4557
SFR-Embedding-2_R	7.11 млрд	4	0.4604
gte-Qwen2-7B-instruct	7.00 млрд	5	0.4593
SciRus-small-cite	61 млн	6	0.4545
multilingual-e5-large-instruct	560 млн	7	0.4395
SciRus-tiny-cite	23 млн	8	0.4490
…	…	…	…
SciRus-small	61 млн	12	0.4334
…	…	…	…
SciRus-tiny	23 млн	17	0.4299

Анализ результатов показывает, что, несмотря на общую тенденцию роста качества с увеличением числа параметров, доменная адаптация играет решающую роль. Компактные модели **SciRus-small-cite** и **SciRus-tiny-cite** демонстрируют высокую производительность, опережая многие значительно более крупные модели общего назначения, особенно на русскоязычных задачах.

Подраздел 3.3.1 посвящен анализу языковой специализации моделей. Благодаря полностью симметричной структуре бенчмарка, где

каждой русскоязычной задаче соответствует англоязычный аналог, становится возможным количественно оценить смещение производительности моделей в сторону одного из языков. Для этого отдельно вычислялся ранг Борда на подмножестве задач на русском языке и на подмножестве задач на английском языке, что позволяет выявить языковую направленность каждой модели. Кроме того, рассчитывался прирост среднего значения мер качества на всех задачах на русском языке по отношению к задачам на английском. В таблице 3 представлены результаты этого анализа.

Таблица 3 — Оценка степени языковой специализации моделей

Модель	**Ранг (англ.)**	**Ранг (рус.)**	**Прирост (рус.), %**
GritLM-7B (7.24 млрд)	1	3	-7.26
SciRus-small-cite (61 млн)	6	**1**	**+0.89**
multilingual-e5-large-instruct (560 млн)	10	8	-0.23
SciRus-tiny-cite (23 млн)	9	**2**	**+0.61**
NV-Embed-v2 (7 млрд)	3	13	-10.65
Giga-Embeddings-instruct (2 млрд)	22	10	14.73
GIST-large-Embedding-v0 (335 млн)	13	29	-48.43

Данные показывают, что многие ведущие модели общего назначения, такие как GritLM-7B и NV-Embed-v2, являются сильно англоцентричными, демонстрируя значительное падение качества на русском языке. В то же время модели семейства **SciRus**, специально адаптированные для русскоязычного научного домена, не только занимают первые места в рейтинге по русскому языку, но и показывают практически идеальный языковой баланс. Интересно отметить, что наблюдается и обратная ситуация: модель от русскоязычных авторов Giga-Embeddings-instruct демонстрирует существенное падение качества при переходе на английский язык (прирост на русском +14.73%), что свидетельствует о ее преимущественной ориентации на русский язык.

В **разделе 3.4** отмечается, что для обеспечения воспроизводимости и стандартизации оценки бенчмарк RuSciBench был интегрирован в ведущий международный фреймворк **Massive Multilingual Text Embedding Benchmark (MMTEB)**. Бенчмарк RuSciBench, включающий 9 русскоязычных задач, составляет существенную долю от общего числа задач в MMTEB на русском языке, насчитывавшего на момент интеграции 73 задачи, что подчёркивает значимость данного вклада в развитие инструментов оценки. Это позволяет любому исследователю легко воспроизвести полученные результаты и оценить новые модели в рамках единой экосистемы.

Если в предыдущих главах оценка моделей проводилась на задачах, связанных со структурой и метаданными научных публикаций, то **четвертая глава** переходит к существенно более сложной задаче — оценке способности моделей к логическому выводу на основе содержания текста. Для этого была формализована задача верификации научных фактов и создан специализированный русскоязычный бенчмарк `RuSciFact`. В главе детально описывается методология его полуавтоматического создания, основанная на генерации утверждений с помощью больших языковых моделей и последующей экспертной валидации, а также приводятся результаты комплексного тестирования моделей, выявляющие их сильные и слабые стороны.

В **разделе 4.1** формализуется задача верификации научных фактов, которая декомпозируется на два последовательных этапа.

– **Информационный поиск**. Для заданного утверждения c в корпусе аннотаций D необходимо найти наиболее релевантный документ-источник e^*. Задача решается путём ранжирования документов по мере близости их векторных представлений:

$$e^* = \underset{e_i \in D}{\operatorname{argmax}}\, s(f(c, \alpha), f(e_i, \alpha)), \tag{1}$$

где $f(\cdot, \alpha)$ — модель-кодировщик, а $s(\cdot, \cdot)$ — косинусная близость. Качество решения оценивается мерой MRR@1, вычисляемой как доля запросов, для которых релевантный документ оказался на первой позиции:

$$\text{MRR@1} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} [\text{rank}_i = 1], \tag{2}$$

где Q — множество запросов (утверждений), rank_i — ранг правильного документа для i-го запроса.

– **Классификация**. Для найденной пары (c, e) определяется метка отношения $l \in \{\texttt{подтверждает}, \texttt{опровергает}\}$. Задача решается обучением классификатора $g'(\cdot, \beta')$ поверх замороженных векторов, минимизируя функцию потерь бинарной перекрестной энтропии:

$$p_i = g'([f(c,\alpha), f(e_i, \alpha)], \beta')$$

$$\mathcal{L}(\beta') = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \to \min_{\beta'}. \tag{3}$$

Качество оценивается F1-мерой, которая является гармоническим средним точности и полноты:

$$F_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{4}$$

где TP, FP и FN — число истинно-положительных, ложно-положительных и ложно-отрицательных решений соответственно.

Разделы 4.2 и 4.3 посвящены конвейеру формирования `RuSciFact`. Утверждения генерировались моделью `Llama-3.1-405B-Instruct`, выбранной по результатам бенчмарка MERA (Fenogenova, 2024). Цель конвейера состояла в том, чтобы получить примеры, проверка которых требует логического вывода, а не лексического сопоставления. Генерация велась по двум независимым направлениям.

Формирование подтверждающих утверждений включало несколько этапов отбора:

- **Отбор информативных аннотаций.** Модель генерировала научный факт, строго следующий из аннотации, либо сообщала о невозможности это сделать. Информативными признано 42% аннотаций из исходного набора.
- **Отбор по лексическому сходству.** Из набора исключались пары, в которых утверждение содержало большую долю слов, совпадающих с текстом аннотации.
- **Отбор по уровню сложности.** LLM присваивала каждому утверждению одну из меток: «простое», «среднее» или «сложное». В итоговый набор включались только утверждения средней и высокой сложности.

Формирование противоречащих утверждений проводилось по следующей схеме:

- **Построение семантического противоречия.** Модели запрещалось использовать синтаксическое отрицание. Утверждение должно было быть противоположным по смыслу аннотации.
- **Проверка качества.** По схеме *LLM-as-a-judge* (Zheng, 2023) каждая пара оценивалась по двум критериям: релевантность теме аннотации и степень поддержки утверждения ее текстом. В итоговый набор отбирались пары с высокой релевантностью и низкой поддержкой.

В **разделе 4.4** описывается заключительный этап экспертной валидации: сгенерированные пары размечались асессорами с опытом работы

с научными текстами, после чего из корпуса исключались неоднозначные и дефектные примеры. Итоговый корпус содержит 1128 пар «утверждение–аннотация» с дисбалансом классов примерно 2:1 в пользу подтверждающих примеров и охватывает широкий спектр научных дисциплин.

Раздел 4.5 представляет результаты экспериментальной оценки широкого спектра моделей на бенчмарке `RuSciFact`. В задаче **информационного поиска** (Таблица 4) наблюдается доминирование крупномасштабных моделей, которые достигают значений MRR@1, близких к идеальным. Это свидетельствует о том, что задача поиска релевантного контекста для современных моделей практически решена. Разработанные модели семейства `SciRus` демонстрируют конкурентоспособные результаты в своем классе, значительно опережая базовые многоязычные модели сопоставимого размера.

Таблица 4 — Результаты оценки моделей в задаче информационного поиска на RuSciFact (MRR@1)

Название модели	Количество параметров	MRR@1
GritLM-7B	7.24 млрд	0.95
SFR-Embedding-2_R	7.11 млрд	0.94
multilingual-e5-large-instruct	560 млн	0.93
BERTA	128 млн	0.92
SciRus-small-cite	61 млн	0.75
SciRus-tiny-cite	23 млн	0.70
bm25	-	0.62
LaBSE-en-ru	129 млн	0.53
rubert-tiny	12 млн	0.09

В **подразделе 4.5.2** рассмотрена задача **классификации** (Таблица 5), в которой структура результатов меняется. Максимальное значение F1 достигает 0.87 у `gte-Qwen2-7B-instruct`, однако значительная часть моделей с разным числом параметров и разной специализацией сосредоточена в узком диапазоне 0.67–0.70. Такое «плато производительности» показывает, что многие кодировщики фиксируют тематическую близость утверждения и аннотации, но их векторные представления имеют ограниченную разрешающую способность для различения отношений поддержки и семантического противоречия. Модели семейства `SciRus`

находятся внутри этого диапазона и достигают уровня ряда существенно более крупных моделей.

Таблица 5 — Результаты оценки моделей в задаче классификации на RuSciFact (F1-мера)

Название модели	Количество параметров	F1
gte-Qwen2-7B-instruct	7 млрд	0.87
SFR-Embedding-2_R	7.11 млрд	0.82
multilingual-e5-large-instruct	560 млн	0.77
GritLM-7B	7.24 млрд	0.73
SciRus-small-cite	61 млн	0.68
LaBSE-en-ru	129 млн	0.68
FRIDA	823 млн	0.67
SciRus-tiny-cite	23 млн	0.67
GIST-large-Embedding-v0	335 млн	0.58

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

1. С помощью двухэтапной методологии обучения разработаны двуязычные модели для векторного представления научных текстов на русском и английском языках. Показана применимость использования пар «заголовок-аннотация» на этапе контрастивного обучения. Эксперименты на бенчмарках RuSciBench и SciDocs показали, что предложенные модели, несмотря на значительно меньшее число параметров, демонстрируют качество, сопоставимое с гораздо более крупными моделями, и обладают высокой вычислительной эффективностью.
2. Разработан и апробирован мультизадачный русско-английский бенчмарк RuSciBench, предназначенный для оценки качества моделей векторного представления научных текстов. Полная версия бенчмарка основана на данных российской научной электронной библиотеки eLibrary.ru и включает 18 сочетаний «задача–язык/направление»; в диссертации подробно описаны компоненты, разработанные лично автором. RuSciBench обеспечивает стандартизированную и воспроизводимую процедуру тестирования и интегрирован в международный лидерборд MTEB. На основе данного бенчмарка впервые проведено

масштабное сравнительное исследование широкого спектра современных моделей векторизации на задачах, связанных с научными текстами. Симметричная двуязычная структура бенчмарка позволила количественно оценить влияние языка задачи на производительность моделей и выявить степень языковой специализации каждой из них.

3. Предложена и экспериментально проверена полуавтоматическая методика формирования наборов данных для задачи верификации научных фактов на русском языке. Данная методика сочетает генерацию научных утверждений на основе аннотаций с использованием больших языковых моделей (LLM), многоэтапную фильтрацию и оценку сгенерированных утверждений самой моделью, а также последующую экспертную валидацию. На основе этой методики создан и опубликован RuSciFact - первый русскоязычный бенчмарк для оценки способности моделей определять, подтверждается ли научное утверждение текстом аннотации или противоречит ему. Проведена оценка современных моделей векторизации на данном бенчмарке.

Публикации автора по теме диссертации

1. *Герасименко Н., Ватолин А., Янина А., Воронцов К.* SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 216–227 (RSCI, Scopus)

2. *Ватолин А., Герасименко Н., Янина А., Воронцов К.* RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 284–294 (RSCI, Scopus)

3. *Vatolin A., Gerasimenko N., Loukachevitch N., Ianina A., Vorontsov K.* ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference ”Dialogue”. 2025. V. 23. pp. S435–S459. DOI: 10.28995/2075-7182-2025-23-435-459 (Scopus)

4. *Vatolin A.* Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024 // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference ”Dialogue”. 2025. V. 23. pp. S416–S434. DOI: 10.28995/2075-7182-2025-23-416-434 (Scopus)

5. *Enevoldsen K., Chung I., Kerboua I., Kardos M., Mathur A., Stap D., Gala J., Siblini W., Krzemiński D., Winata G. I., Sturua S., Utpala S., Ciancone M., Schaeffer M., Sequeira G., Misra D., Dhakal S., Rystrøm J., Solomatin R., Çağatan Ö., Kundu A., Bernstorff M., Xiao S., Sukhlecha A., Auras D., Plüster B., Harries J. P., Magne L., Mohr I., Hendriksen M., Zhu D., Gisserot-Boukhlef H., Aarsen T., Kostkan J., Wojtasik K., Lee T., Šuppa M., Zhang C., Rocca R., Hamdy M., Michail A., Yang J., Faysse M., Vatolin A. [u ∂p.]* MMTEB: Massive Multilingual Text Embedding Benchmark // International Conference on Learning Representations (ICLR). 2025. pp. 101715–101771. https://proceedings.iclr.cc/paper_files/paper/2025/file/fc0e3f908a2116ba529ad0a1530a3675-Paper-Conference.pdf (Scopus)