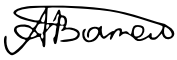
На правах рукописи



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 — 2025

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 «____» _____ 2025 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) уступают трансформерным моделям в семантическом поиске по научным текстам (Salton, 1975; Robertson, 2009; Vaswani, 2017). Современные нейросетевые модели, особенно архитектуры трансформер (Vaswani, 2017), демонстрируют преимущество над традиционными подходами, поскольку 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.

Разработка и предоставление научному сообществу открытых моделей и инструментариев для оценки, таких как SciRus и RuSciBench, отвечает целям и задачам развития искусственного интеллекта в Российской Федерации, перечисленным в «Национальной стратегии развития искусственного интеллекта на период до 2030 года» (с изменениями от 15 февраля 2024 г.) (Президент Российской Федерации, 2024). Это направление деятельности полностью соответствует концепции открытой науки, предполагающей свободный доступ к исследовательским данным, инструментам и результатам, что способствует ускорению научного прогресса и повышению прозрачности исследовательской деятельности.

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 двуязычной модели для векторного представления научных текстов на русском и английском языках. Исследовать подходы к обучению, основанные на доступных данных из мультиязычных корпусов, без дополнительной разметки. При этом модель должна обеспечивать высокую скорость работы на центральном процессоре.
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, при обучении используются пары «цитирующая статья — цитируемая статья», основанные на однонаправленной связи из графа цитирований. На бенчмарке SciDocs модели достигают качества, сравнимого с лучшими англоязычными моделями при вдвое меньшем числе параметров. На RuSciBench по рейтингу Борда занимают 1–2 места на русском языке и входят в топ-10 среди моделей на английском.
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 бенчмарк 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 бенчмарк RuSciFact, который позволяет оценить способность моделей векторизации решать задачу проверки научных фактов.

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 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», без использования графа цитирований.
2. Разработаны модели векторизации научных текстов с поддержкой русского и английского языков и высокой скоростью работы на центральном процессоре.
3. Разработан первый набор для оценки качества работы моделей с научными данными на русском и английском языках, состоящий из

различных типов задач. На его основе проведено сравнительное исследование широкого спектра современных моделей векторизации в задачах на научных данных. Благодаря симметричной двуязычной структуре бенчмарка впервые количественно установлена зависимость производительности моделей от языка задачи и выявлена степень языковой специализации каждой модели.

4. 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 в первую очередь подтверждается внедрением модели векторизации научных текстов SciRus-tiny на российском научном портале elibrary.ru. Был разработан новый режим «нейропоиск», который позволяет находить тематически близкие научные публикации, используя в качестве запроса аннотацию статьи. Это внедрение упрощает анализ научной информации для широкого круга исследователей и специалистов, работающих с научной библиотекой elibrary.

Кроме того, 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. Данные, на основе которых был создан бенчмарк RuSciBench, послужили основой для одной из задач в мультиязычном наборе для оценки AIRBench (Yang, 2024), а также для одной из задач в русскоязычном тестовом наборе LIBRA (Churin, 2024), что подтверждает интерес научного сообщества к результатам работы.

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: [1] - предобучение маленькой версии модели (SciRus-tiny) с помощью маскированного языкового моделирования, дообучение обеих версий модели (SciRus-tiny и SciRus-small) на парах заголовок-аннотация, а также на парах «цитирующая статья — цитируемая статья», валидация моделей на наборе данных SciDocs. [2] - 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. Также реализация исходного кода инструментария для оценки, валидация моделей, интеграция бенчмарка в международный бенчмарк MTEB. [3] - вклад соискателя является определяющим. [5] — в рамках работы над созданием и расширением 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.

Публикации. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 2 — в периодических научных журналах, индексируемых Web of Science и Scopus, 3 — в периодических научных журналах, индексируемых Scopus.

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 представлен анализ теоретических основ и современных методов построения семантических векторных представлений текстов, служащих фундаментом для последующих глав работы. В главе формализуется задача векторизации, рассматривается эволюция подходов и детально излагается технологический стек, использованный при разработке моделей в данной диссертации.

В **разделах 1.1 и 1.2** дается формальная постановка задачи семантической векторизации как построения отображения $f(x, \alpha)$ из пространства текстов

в многомерное векторное пространство $\mathbb{R}^d$. Подчеркивается ключевое требование к такому отображению: геометрическая близость векторов должна отражать семантическую близость исходных текстов. Приводится краткий исторический экскурс в развитие методов, от статистических подходов (TF-IDF) и матричных разложений (LSA) до нейросетевых моделей, способных генерировать контекстуализированные представления. Далее рассматривается этап предобработки текста — токенизация. Обосновывается переход от токенизации по словам, имеющей проблему «неизвестных слов», к методам токенизации на уровне фрагментов слов. В качестве основного подхода описывается алгоритм Byte-Pair Encoding (BPE) (Sennrich, 2015), который позволяет эффективно обрабатывать тексты с богатой морфологией, формируя словарь из наиболее частотных символьных последовательностей.

Раздел 1.3 посвящен архитектуре трансформер-кодировщика (Vaswani, 2017), лежащей в основе всех современных моделей. В рамках данной работы используется архитектура, основанная на модели BERT. Процесс получения контекстуализированных векторных представлений для последовательности токенов x_i начинается со слоя векторизации f_e, где формируются начальные представления токенов:

$$\mathbf{h}^{(0)} = f_e(x_i, \alpha_e) = E(x_i) + P_i.$$

На этом этапе каждому токену из входной последовательности сопоставляется его векторное представление из обучаемой матрицы $E \in \mathbb{R}^{|\mathcal{V}| \times d}$, где $|\mathcal{V}|$ — размер словаря, а d — размерность векторного пространства. К полученным векторам прибавляются соответствующие позиционные эмбеддинги из матрицы P, кодирующие информацию о порядке токенов. Полученная матрица скрытых состояний $\mathbf{h}^{(0)}$ последовательно обрабатывается L идентичными трансформер-блоками. Работа l-го блока ($l = 1, \ldots, L$) описывается следующими преобразованиями:

$$\mathbf{z}^{(l)} = \mathrm{LayerNorm}\big(\mathbf{h}^{(l-1)} + \mathrm{MHAtt}^{(l)}(\mathbf{h}^{(l-1)})\big),$$
$$\mathbf{h}^{(l)} = \mathrm{LayerNorm}\big(\mathbf{z}^{(l)} + \mathrm{FF}^{(l)}(\mathbf{z}^{(l)})\big).$$

Каждый блок состоит из двух ключевых подслоев: механизма многоголового внимания ($\mathrm{MHAtt}^{(l)}$) и двухслойной полносвязной нейронной сети ($\mathrm{FF}^{(l)}$). Механизм внимания, состоящий из H параллельных «голов», позволяет модели улавливать сложные контекстуальные зависимости, вычисляя взвешенную сумму векторов всех токенов в последовательности. После каждого из двух подслоев применяются остаточное соединение (skip connection) и нормализация по слою ($\mathrm{LayerNorm}$) для стабилизации процесса обучения. В результате, на выходе последнего трансформер-блока формируется матрица контекстуализированных векторов токенов $\mathbf{h}^{(L)}$.

В **разделе 1.4** излагается методология предобучения трансформер-кодировщика, исходно предложенная в работе **BERT** (Devlin, 2019). Однако

в обзоре основной акцент делается на конфигурацию **RoBERTa** (Liu, 2019) как более современную, где ключевой задачей при обучении служит маскированное языковое моделирование (MLM). Для каждой входной последовательности токенов $x_i = (w_{i,1}, \ldots, w_{i,n_i})$ случайным образом выбирается подмножество позиций для маскирования. Индикатор маскирования для j-й позиции определяется как случайная величина $m_{i,j} \sim \mathrm{Bernoulli}(p_m)$, где вероятность маскирования $p_m = 0.15$. На основе этих индикаторов формируется модифицированная последовательность $x'_i = (w'_{i,1}, \ldots, w'_{i,n_i})$, где токен в маскируемой позиции заменяется согласно следующему правилу:

$$w'_{i,j} = \begin{cases} [\mathrm{MASK}], & \text{если } m_{i,j} = 1 \text{ и } r < 0.8, \\ u, & \text{если } m_{i,j} = 1 \text{ и } 0.8 \le r < 0.9, \\ w_{i,j}, & \text{в остальных случаях,} \end{cases}$$

где $r \sim \mathcal{U}(0,1)$ — случайная величина, равномерно распределенная на отрезке $[0,1)$, а $u \sim \mathcal{U}(\mathcal{V})$ — случайный токен, выбранный из словаря $\mathcal{V}$. На этапе предобучения к выходам модели-кодировщика $f(\cdot, \alpha)$ добавляется дополнительный полносвязный слой, предназначенный для предсказания распределения вероятностей по всему словарю $\mathcal{V}$ для каждой токенизированной позиции. Таким образом, полная модель для решения задачи MLM, обозначим ее g_{MLM}, преобразует выходное представление j-го токена $\mathbf{h}_{i,j} = f(x'_i, \alpha)_j$ в вектор вероятностей $\mathbf{p}_{i,j}$:

$$\mathbf{p}_{i,j} = g_{\mathrm{MLM}}(x'_i; \alpha, \beta)_j = \mathrm{softmax}(W f(x'_i, \alpha)_j + \mathbf{b}), \tag{1}$$

где параметры $\beta = \{W, \mathbf{b}\}$ соответствуют матрице весов $W \in \mathbb{R}^{|\mathcal{V}| \times d}$ и вектору смещений $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$ данного классификационного слоя. Оптимизация совокупности параметров (α, β) производится путём минимизации функции потерь перекрестной энтропии, которая вычисляется только по маскированным позициям:

$$\mathcal{L}_{\mathrm{MLM}}(\alpha, \beta) = -\sum_{i=1}^{B} \sum_{j \in M_i} \log \mathbf{p}_{i,j}[w_{i,j}] \to \min_{\alpha,\beta},$$

где B — размер мини-выборки, $M_i = \{j \mid m_{i,j} = 1\}$ — множество индексов маскированных токенов для i-й последовательности, а $\mathbf{p}_{i,j}[w_{i,j}]$ — предсказанная моделью вероятность истинного токена $w_{i,j}$.

В исходной работе BERT (Devlin, 2019) дополнительно предлагалась задача предсказания следующего предложения (NSP), однако последующие исследования показали отсутствие устойчивого выигрыша от её использования, в RoBERTa от этой задачи отказались.

В конфигурации RoBERTa были тщательно пересмотрены ключевые аспекты процесса обучения: от статического маскирования перешли к динамическому (маска генерируется заново в каждой эпохе), исключили задачу NSP и использовали значительно большие мини-выборки и объёмы данных. Именно на

этой усовершенствованной методологии основываются модели, представленные в данной работе.

В **разделе 1.5** рассматриваются методы дообучения (fine-tuning) предобученных трансформер-кодировщиков для задач оценки семантической близости текстов. Показано, что без специальной адаптации стандартные выходы моделей типа BERT не позволяют эффективно решать задачи семантического поиска. В разделе противопоставляются два подхода. Первый — архитектура перекрестного кодировщика (cross-encoder), которая обрабатывает пару текстов совместно и выдает оценку их близости. Несмотря на высокую точность этого метода, его вычислительная сложность препятствует применению на больших коллекциях.

Второй, практически применимый подход, основан на сиамской архитектуре (bi-encoder), которая генерирует независимое семантическое векторное представление $\mathbf{v} = f(x, \alpha)$ для каждого текста, что позволяет реализовать эффективный поиск. Описаны ключевые функции потерь, включая Triplet Loss (Schroff, 2015), использующую тройки примеров (якорь, положительный, отрицательный):

$$\mathcal{L}_{\text{triplet}}(\alpha) = \sum_{i=1}^{B} \max\bigl(0, d(\mathbf{v}_{a,i}, \mathbf{v}_{p,i}) - d(\mathbf{v}_{a,i}, \mathbf{v}_{n,i}) + \epsilon\bigr) \to \min_{\alpha},$$

и функцию потерь InfoNCE (van den Oord, 2018), которая использует только положительные пары «якорь - положительный пример» и формирует отрицательные примеры динамически из других элементов мини-выборки (in-batch negatives):

$$\mathcal{L}_{\text{InfoNCE}}(\alpha) = -\sum_{i=1}^{B} \log \frac{\exp(s(\mathbf{v}_{a,i}, \mathbf{v}_{p,i})/\tau)}{\sum_{j=1}^{B} \exp(s(\mathbf{v}_{a,i}, \mathbf{v}_{p,j})/\tau)} \to \min_{\alpha},$$

где $s(\cdot, \cdot)$ — косинусная близость, а τ — гиперпараметр температуры.

В **разделе 1.6** рассматривается применение описанных методов в научной области. Обосновывается необходимость доменной адаптации: научные тексты характеризуются специализированной терминологией, строго регламентированной структурой и сложными семантическими связями через цитирования, что создает доменный сдвиг для моделей общего назначения. Рассматриваются специализированные модели:

- **SPECTER** (Cohan, 2020), использующая граф цитирований для формирования обучающих троек
- **SPECTER2**, предлагающая мультиформатное обучение для генерации специализированных векторов под различные типы задач
- **SciNCL** (Ostendorff, 2022), применяющая непрерывную меру близости из специально обученной графовой модели вместо дискретного сигнала цитирования.

В **разделе 1.7** представлен обзор инструментов оценки. Бенчмарк **SciDocs** включает семь задач в четырех категориях (классификация, предсказание цитирований, анализ пользовательской активности, рекомендации). Бенчмарк

SciRepEval расширяет подход до 24 задач в четырех форматах с разделением на группы In-Train и Out-of-Train для оценки обобщающей способности. Бенчмарк **SciFact** фокусируется на верификации научных фактов, требующей логического вывода для определения поддержки или опровержения утверждений. Универсальный бенчмарк **MTEB** объединяет и стандартизирует множество наборов данных, предоставляя единый интерфейс и таблицу лидеров для сравнения моделей.

В **заключительном разделе 1.8** делается вывод, что все передовые решения и инструменты оценки ориентированы исключительно на английский язык. Это формирует научную задачу диссертации — разработку эффективных двуязычных семантических представлений для научных документов и инструментария для их оценки.

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** формализуется постановка задачи векторизации научных текстов.

Дано: коллекция документов $\mathcal{D} = \{x_i\}_{i=1}^{N}$.

Найти: параметризованное отображение $f(\cdot, \alpha)$, которое сопоставляет каждому документу x_i вектор в евклидовом пространстве:

$$\mathbf{v}_i = f(x_i, \alpha) \in \mathbb{R}^d, \quad i = 1, \ldots, N.$$

Ключевое требование к этому отображению заключается в том, чтобы геометрическая близость векторов отражала семантическую близость исходных документов.

Качество и универсальность искомого отображения определяются тремя основными **критериями**, соответствующими этапам его построения и применения. Первый критерий связан с этапом предобучения, на котором модель обучается на задаче маскированного языкового моделирования, что позволяет ей выучить контекстные векторные представления токенов:

$$\sum_{i=1}^{N} \mathcal{L}_{\text{MLM}}(x_i, \alpha) \to \min_{\alpha}.$$

Второй критерий относится к этапу контрастивного дообучения, целью которого является формирование семантического векторного пространства. Это достигается за счет минимизации контрастивной функции потерь $\mathcal{L}_{\text{Contr}}$ на обучающих парах семантически близких документов $(x_{a,i}, x_{p,i})$.

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

Третий критерий относится к этапу применения модели для решения прикладных задач. Для этого используется новый малый набор данных $\mathcal{D}' = \{x'_i\}_{i=1}^{M}$, где $M \ll N$. Параметры α кодировщика f фиксируются, а обучается только модель $g'(\cdot, \beta')$:

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

В **разделе 2.2** описываются наборы данных для обучения. Для обучения моделей использовались два крупных источника научных текстов.

Первый источник — международный архив **Semantic Scholar Academic Graph Dataset (S2AG)** (Kyle, 2020), исходно содержащий метаданные более 200 миллионов публикаций. Из него была сформирована выборка объемом 30.5 млн пар «заголовок–аннотация»; такой объем был выбран для уменьшения итоговой доли английского языка в объединенном корпусе, поскольку S2AG является преимущественно англоязычным (83.3%). Помимо текстовых метаданных, набор данных включает граф цитирований S2ORC, содержащий более 51.9 миллионов ребер, что используется для формирования дополнительных обучающих пар на этапе контрастивного дообучения.

Второй источник — данные российской научной библиотеки **eLibrary.ru**. Эта часть состоит из 17.4 млн пар «заголовок–аннотация» и включает 8.6 млн русскоязычных и 8.8 млн англоязычных документов. Ключевой особенностью этого набора данных является наличие около 5.2 миллионов русскоязычных статей, имеющих параллельные англоязычные версии аннотаций, что представляет собой ценный параллельный корпус для формирования единого семантического пространства для русского и английского языков. Граф цитирований на основе данных eLibrary.ru содержит около 40 миллионов ребер.

Таким образом, итоговый текстовый корпус для обучения модели содержит около 48.2 миллионов заголовков и аннотаций научных статей, 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 S2AG 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 векторного представления документа из матрицы контекстуализированных векторов токенов $\mathbf{h}^{(L)} \in \mathbb{R}^{n \times d}$, полученной на выходе трансформер-кодировщика, был добавлен слой усреднения (Mean Pooling). Этот слой вычисляет итоговый вектор документа $\mathbf{v}$ как среднее арифметическое векторов всех токенов в последовательности:

$$\mathbf{v} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{h}_j, \tag{2}$$

где n — длина последовательности токенов, а $\mathbf{h}_j$ — векторное представление j-го токена на выходе последнего слоя кодировщика.

Обучение на данном этапе производилось путём минимизации функции потерь InfoNCE (van den Oord, 2018), параметр τ был установлен равным 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 разница составляет около 3 процентных пунктов, что свидетельствует о высокой эффективности обучения на текстовых метаданных. Этот результат имеет важное практическое значение: пары заголовков и аннотаций значительно более доступны, чем данные о цитированиях, которые требуют наличия полного графа научных публикаций. Таким образом, даже при отсутствии информации о цитированиях возможно обучение моделей векторизации с приемлемым качеством.

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.

Практическая значимость работы подтверждается в **разделе 2.7**, где описывается внедрение модели `SciRus-tiny` в информационно-поисковую систему

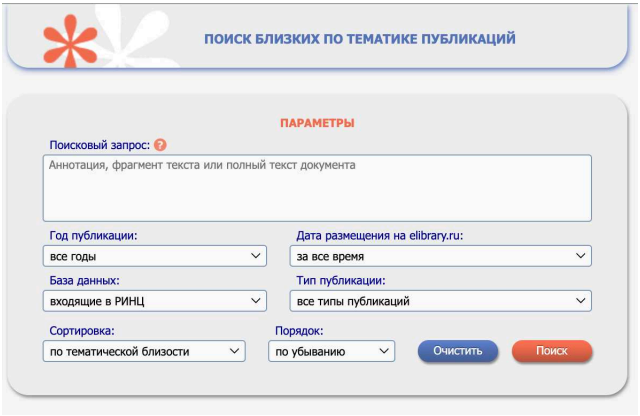


Рисунок 1 — Интерфейс режима «нейропоиск» на портале eLibrary.ru, использующего модель SciRus-tiny.

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 решается задача создания стандартизированного инструментария для оценки качества моделей векторизации научных текстов. Обосновывается необходимость такого инструмента, поскольку существующие бенчмарки общего назначения не учитывают специфику научного дискурса, а специализированные научные бенчмарки для русского языка отсутствуют. Для устранения этого пробела был разработан **RuSciBench** [2] — мультизадачный двуязычный бенчмарк, основанный на корпусе из 182 264 научных статей из электронной библиотеки **eLibrary.ru**, прошедших предварительную фильтрацию и предобработку. Примеры отобраны таким образом, чтобы для всех заголовков и аннотаций имелся перевод.

Каждая задача в бенчмарке представлена на русском и английском языках, что позволяет проводить всестороннюю оценку моделей. В диссертации подробно описываются 9 наборов данных, разработанных лично соискателем: две задачи классификации, одна задача регрессии и одна задача информационного поиска, каждая из которых представлена на обоих языках, а также одна задача кросс-языкового поиска. Всего бенчмарк состоит из 18 задач. При оценке модели на каждой из задач ее параметры $f(x, \alpha)$ остаются неизменными, а на полученных векторах $\mathbf{v}_i = f(x_i, \alpha)$ обучается модель $g'(\cdot, \beta')$ ($\dim(\beta') \ll \dim(\alpha)$).

В **разделе 3.1** описываются задачи классификации. Для коллекции документов $\mathcal{D} = \{x_i\}_{i=1}^N$ и множества меток $C = \{c_1, \ldots, c_K\}$ строится модель, сопоставляющая каждому документу x_i его класс $y_i \in C$. Векторные представления $\mathbf{v}_i = f(x_i, \alpha)$, полученные от замороженного кодировщика, используются в качестве признаков для обучения модели логистической регрессии $g'(\mathbf{v}, \beta')$.

Вероятность принадлежности документа x_i к классу c_k моделируется с помощью функции softmax:

$$p_{ik} = P(y_i = c_k|\mathbf{v}_i, \beta') = \frac{\exp(\mathbf{v}_i^T W_k + b_k)}{\sum_{j=1}^{K} \exp(\mathbf{v}_i^T W_j + b_j)},$$

где W_k — вектор весов, а b_k — свободный член (смещение) для класса c_k. Полный набор обучаемых параметров классификатора — $\beta' = \{(W_k, b_k)\}_{k=1}^{K}$. Процесс обучения сводится к минимизации кросс-энтропийной функции потерь с L2-регуляризацией весовых векторов:

$$\mathcal{L}(\beta') = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}[y_i = c_k]\log p_{ik} + \lambda\sum_{k=1}^{K}||W_k||_2^2 \longrightarrow \min_{\beta'}.$$

В связи с дисбалансом классов была применена процедура балансировки выборок, что позволило использовать точность (Accuracy) в качестве основной меры качества. Бенчмарк включает две задачи этого типа:

– **Классификация по рубрикатору ГРНТИ** - 29 классов.

– **Классификация по типу публикации** - 4 класса.

Раздел 3.2 посвящен задаче регрессии в бенчмарке RuSciBench. В рамках данной постановки для каждого документа x_i необходимо предсказать соответствующее ему действительное число $y_i \in \mathbb{R}$. Аналогично задачам классификации, параметры кодировщика $f(x, \alpha)$ остаются неизменными, а на полученных векторах $\mathbf{v}_i$ обучается регрессионная модель $g'(\mathbf{v}, \beta')$. В качестве такой модели используется линейная регрессия:

$$g'(\mathbf{v}_i, \beta') = \mathbf{v}_i^T W + b,$$

где $\beta' = \{W, b\}$ — обучаемые параметры: вектор весов W и свободный член b. Обучение заключается в минимизации среднеквадратичной ошибки (MSE) по параметрам β':

$$\mathcal{L}(\beta') = \frac{1}{N}\sum_{i=1}^{N}(y_i - g'(\mathbf{v}_i, \beta'))^2 \longrightarrow \min_{\beta'}.$$

Для оценки качества используется коэффициент ранговой корреляции Кендалла (вариант τ_b), который вычисляется по формуле:

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T_{\hat{y}})(P + Q + T_y)}}, \tag{3}$$

где P — число согласованных пар объектов (ранги истинных и предсказанных значений совпадают), Q — число несогласованных пар, а T_y и $T_{\hat{y}}$ — число пар с совпадающими значениями (связками) в истинных и предсказанных рангах соответственно. Выбор этой меры качества обусловлен ее устойчивостью к выбросам

и нелинейным зависимостям, что особенно важно для таких сильно скошенных распределений, как число цитирований. Итоговое значение меры качества принимается равным $\max(0, \tau_b)$ для обеспечения единой шкалы с другими задачами бенчмарка. Бенчмарк включает одну задачу этого типа - предсказание числа цитирований. В данной задаче оценивается способность модели улавливать в тексте аннотации сигналы, коррелирующие с научной значимостью и влиятельностью публикации, а также возрастом публикации, так как наблюдается корреляция между годом публикации и количеством цитат.

В **разделе 3.3** описывается задача информационного поиска в бенчмарке RuSciBench, а именно задача поиска прямых цитирований, в рамках которой для заданной статьи-запроса, представленной ее заголовком и аннотацией, необходимо найти в общем корпусе научных публикаций те работы, которые она цитирует.

Процедура оценки следует стандартной парадигме информационного поиска. На первом этапе все запросы и документы корпуса векторизуются с помощью оцениваемой модели $f(x, \alpha)$. Затем для каждого вектора запроса $\mathbf{v}_q$ документы корпуса ранжируются по убыванию косинусной близости:

$$\text{similarity}(\mathbf{v}_q, \mathbf{v}_d) = \frac{\mathbf{v}_q \cdot \mathbf{v}_d}{||\mathbf{v}_q|| \cdot ||\mathbf{v}_d||}.$$

Качество полученного списка оценивается с помощью нормализованного дисконтированного совокупного выигрыша на первых 10 позициях (NDCG@10). Данная мера качества определяется как отношение дисконтированного совокупного выигрыша (DCG) к его идеальному значению (IDCG), которое достигается при наилучшем возможном ранжировании:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}. \tag{4}$$

Поскольку релевантность в данной задаче бинарна, дисконтированный совокупный выигрыш DCG@k вычисляется как сумма, где каждая релевантная работа вносит вклад, обратно пропорциональный логарифму ее позиции в списке:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{[rel_i = 1]}{\log_2(i+1)},$$

где $rel_i = 1$, если документ на i-й позиции является процитированным (то есть релевантным), и 0 в противном случае.

Раздел 3.4 посвящен задаче кросс-языкового поиска, предназначенной для оценки качества выравнивания семантических пространств разных языков. Задача состоит в том, чтобы для каждой аннотации на русском языке найти ее точный перевод в параллельном корпусе аннотаций на английском языке.

Поиск выполняется путём нахождения ближайшего соседа в векторном пространстве по косинусной близости. Предсказанным переводом $\hat{d}_j$ для запроса q_j считается документ:

$$\hat{d}_j = \arg\max_{d_i \in \mathcal{D}_{en}} \frac{\mathbf{v}_{q,j} \cdot \mathbf{v}_{d,i}}{||\mathbf{v}_{q,j}|| \cdot ||\mathbf{v}_{d,i}||}.$$

В качестве меры качества используется точность (Accuracy) — доля правильно найденных переводов:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}_{ru}|} \sum_{q_j \in \mathcal{D}_{ru}} [\hat{d}_j = d_{j,\text{true}}],$$

где $d_{j,\text{true}}$ — истинный перевод для запроса q_j.

В **разделе 3.5** представлены итоговые результаты оценки 29 моделей, включая ряд передовых решений, опубликованных уже после выхода бенчмарка. Для агрегирования результатов по 18 задачам бенчмарка используется метод Борда, который вычисляет итоговый балл B_i для каждой модели на основе суммы ее рангов по всем задачам:

$$B_i = \sum_{j=1}^{m} (n - r_{ij}),$$

где m — число задач, n — число моделей, r_{ij} — ранг i-й модели в j-й задаче. 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.

Раздел 3.6 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 3.7** 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:

$$e^* = \underset{e_i \in D}{\operatorname{argmax}}\, s(f(c, \alpha), f(e_i, \alpha)), \tag{5}$$

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:

$$\text{MRR@1} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} [\text{rank}_i = 1], \tag{6}$$

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:

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

$$\mathcal{L}(\beta') = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \to \min_{\beta'}. \tag{7}$$

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

$$F_1 = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{8}$$

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

Разделы 4.2 и 4.3 посвящены методологии и многоэтапному конвейеру формирования набора данных `RuSciFact`. Ключевым элементом подхода стало использование большой языковой модели `Meta-Llama-3.1-405B-Instruct`, выбор которой был обоснован ее лидирующими позициями среди моделей с открытыми весами на многозадачном русскоязычном бенчмарке MERA (Fenogenova, 2024). Это обеспечило как высокое качество генерации, так и воспроизводимость исследования. Весь процесс был построен с целью создания примеров, проверка которых требует от моделей именно логического вывода, а не поверхностного лексического сопоставления.

Процесс генерации был разделен на две независимые ветви для создания подтверждающих и противоречащих утверждений.

Генерация подтверждающих утверждений включала несколько этапов фильтрации для повышения сложности и валидности данных:

- **Отбор информативных аннотаций.** Языковой модели посредством промпта была поставлена задача либо сгенерировать научный факт, строго следующий из аннотации, либо сообщить о невозможности сделать это. На этом этапе в качестве информативных аннотаций было выбрано 42% аннотаций из начального набора.
- **Фильтрация по лексическому сходству.** Чтобы исключить тривиальные примеры, представляющие собой прямое цитирование, все сгенерированные пары «утверждение–аннотация» проходили фильтрацию на основе меры лексического сходства, исключающую примеры, где утверждение содержит большую долю слов, совпадающих с текстом аннотации.
- **Отбор по уровню сложности.** На заключительном этапе автоматической фильтрации LLM классифицировала каждое утверждение как «простое», «среднее» или «сложное». Для итогового набора данных отбирались только утверждения средней и высокой сложности, что позволило сфокусировать бенчмарк на нетривиальных научных фактах, которые не являются общеизвестными.

Генерация противоречащих утверждений представляла собой более сложную задачу и требовала особого подхода:

- **Создание семантических противоречий.** Ключевым требованием к модели был запрет на использование прямого синтаксического отрицания (например, добавления частицы «не»). Вместо этого модель должна была сформулировать утверждение, которое является антонимичным по смыслу аннотации.
- **Автоматизированный контроль качества.** Для фильтрации артефактов, возникающих при генерации, была применена парадигма «LLM как судья» (*LLM-as-a-judge*) (Zheng, 2023). Языковая модель оценивала каждую сгенерированную пару по двум шкалам: релевантность утверждения теме аннотации и степень поддержки утверждения текстом (от полного противоречия до полной поддержки). Для итогового набора

отбирались только те пары, которые получили высокие оценки за релевантность и низкие — за поддержку.

В **разделе 4.4** описывается этап экспертной валидации. Весь сгенерированный корпус был независимо размечен двумя асессорами, которые имеют большой опыт работы с научными текстами. 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
LaBSE-en-ru	129 млн	0.53
rubert-tiny	12 млн	0.09

В задаче **классификации** (Таблица 5) картина принципиально иная. Результаты большинства моделей оказываются хуже, чем в задаче поиска, и сильно уплотнены в узком диапазоне значений. Это явление указывает на то, что логический вывод для различения подтверждения и семантического противоречия является существенно более сложной задачей для существующих векторных представлений. Модели семейства `SciRus` показывают результаты на уровне многих более крупных аналогов, что подтверждает их эффективность.

В **разделе 4.6** проведен анализ сложности задачи классификации в разрезе научных дисциплин путём вычисления усреднённой по моделям частоты ошибок для каждой области. Результаты показывают значительную вариативность сложности: некоторые области демонстрируют существенно более высокую частоту ошибок, что указывает на необходимость доменно-специфичных подходов для различных научных дисциплин.

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

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

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. Бенчмарк суммарно включает 18 задач, 9 из которых были разработаны лично, и основан на данных российской научной электронной библиотеки eLibrary.ru. 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. *Н. Герасименко, А. Ватолин, А. Янина, К. Воронцов*, SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 193-202 (РИНЦ, RSCI, Scopus)

2. *А. Ватолин, Н. Герасименко, А. Янина, К. Воронцов*, 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 (РИНЦ, 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. https://doi.org/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. https://doi.org/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., Thakur N., Dey M., Vasani D., Chitale P., Tedeschi S., Tai N., Snegirev A., Günther M., Xia M., Shi W., Lù X. H., Clive J., Krishnakumar G., Maksimova A., Wehrli S., Tikhonova M., Panchal H., Abramov A., Ostendorff M., Liu Z., Clematide S., Miranda L. J., Fenogenova A., Song G., Safi R. B., Li W.-D., Borghini A., Cassano F., Su H., Lin J., Yen H., Hansen L., Hooker S., Xiao C., Adlakha V., Weller O., Reddy S., Muennighoff N.* MMTEB: Massive Multilingual Text Embedding Benchmark // International Conference on Representation Learning (ICLR). 2025. pp. 101715–101771. https://openreview.net/forum?id=zl3pfz4VCV (Scopus)