

ОТЗЫВ

официального оппонента на диссертацию

Ватолина Алексея Сергеевича

«Обучение и оценивание мультязычных нейросетевых моделей семантического векторного представления научных текстов»,

представленную на соискание ученой степени кандидата технических наук по специальности 2.3.8 Информатика и информационные процессы

Актуальность темы. В последние годы наблюдается устойчивый рост объема научных публикаций в крупнейших международных и национальных библиографических базах данных. Количество ежегодно индексируемых статей постоянно увеличивается, а общий объем научных корпусов измеряется десятками миллионов документов. В этих условиях существенно повышаются требования к системам хранения, обработки и поиска научной информации. Эффективность исследовательской деятельности напрямую зависит от возможностей автоматизированного анализа больших коллекций научных текстов, включая определение тематической близости, извлечение релевантных материалов и анализ связей между научными работами.

Параллельно с ростом научного контента усложняются методы его обработки. Традиционные алгоритмы информационного поиска и статистические модели текстового анализа оказываются недостаточными для работы с современными научными текстами. Нейросетевые модели на основе архитектуры Transformer обеспечивают качественно новый уровень представления текстов, учитывая контекст, семантические связи и многоязычные особенности данных. Однако существующие решения в области векторизации текста преимущественно ориентированы на английский язык, в то время как специализированные инструменты для научных текстов на русском языке отсутствуют или находятся на ранних стадиях развития.

Кроме того, научный дискурс обладает специфическими особенностями: высокой частотой терминов, структурной регулярностью, информационной плотностью и сложными межтекстовыми связями, что требует разработки специализированных моделей и методов оценки, адаптированных именно к научным данным. Недостаток таких инструментов ограничивает возможности построения эффективных систем научного поиска и анализа.

Таким образом, создание компактных мультязычных моделей для векторного представления научных текстов, а также разработка специализированных бенчмарков для их объективной оценки на научных данных, является актуальной задачей, имеющей существенное значение как

для развития технологий обработки естественного языка, так и для поддержки научной инфраструктуры в целом.

Научная новизна. В диссертационной работе получены следующие наиболее важные научные результаты.

1. Разработаны компактные двуязычные модели SciRus, демонстрирующие соизмеримое или превосходящее качество по сравнению с крупными многоязычными моделями при значительно меньшем количестве параметров.
2. Создан специализированный двуязычный бенчмарк RuSciBench, объединяющий задачи классификации, регрессии, одно- и кроссязычного информационного поиска специально для научных данных.
3. Предложена методика формирования датасетов для проверки научных фактов на русском языке, сочетающая генерацию утверждений больших языковых моделей, самооценку модели и экспертную валидацию.
4. Создан первый русскоязычный набор данных для верификации научных фактов RuSciFact, предназначенный для оценки соответствующих моделей в научном контексте.

Практическая значимость полученных результатов подтверждается внедрением разработанной модели векторизации научных текстов SciRus-tiny в инфраструктуру крупнейшей российской научной электронной библиотеки eLIBRARY.ru. На основе модели реализован режим нейросетевого поиска, позволяющий находить тематически близкие публикации по тексту аннотации.

Дополнительную практическую ценность представляют созданные автором открытые наборы данных и методики их формирования. Разработанный набор данных RuSciBench был интегрирован в международный бенчмарк MTEB, что обеспечивает стандартизированную и воспроизводимую оценку качества моделей векторизации научных текстов. Исползованные в нем данные также легли в основу отдельных задач в мультязычном наборе AIRBench и русскоязычном тестовом наборе LIBRA, что свидетельствует о востребованности предложенных подходов в международном и отечественном научных сообществах.

Достоверность и обоснованность научных результатов. В диссертационной работе корректно применяются нейросетевые модели,

методы машинного обучения и средства автоматической обработки текстов. Экспериментальные исследования моделей и разработанной методики проведены на нескольких независимых тестовых наборах с использованием общепринятых метрик и воспроизводимых процедур оценки. Результаты сравниваются с показателями актуальных моделей, что позволяет объективно продемонстрировать преимущества предложенных решений.

Основные результаты диссертации обсуждались на международных научных конференциях. Содержание исследования достаточно полно изложено в пяти печатных изданиях, четыре из которых индексируются Web of Science и Scopus.

Содержание работы.

Диссертация включает введение, 4 главы, заключение, список литературы и приложения.

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследования, показаны научная новизна и практическая значимость работы.

В первой главе приводится обзор современных методов векторизации текстов – от статистических моделей до подходов на основе архитектуры Transformer. Автор подробно рассматривает процесс токенизации, архитектуру Transformer, методики предобучения и особенности контрастивных схем дообучения. Отдельное внимание уделено анализу моделей, ориентированных на научные тексты, и существующих бенчмарков для их оценки. Глава формирует теоретическую основу исследования и демонстрирует глубокое понимание предметной области.

Во второй главе представлена методология построения моделей SciRus, включая выбор обучающих данных, архитектурные решения и стратегии двухэтапного обучения (маскированное языковое моделирование и контрастивное дообучение). Описаны проведенные эксперименты по оценке качества и сравнению с существующими моделями, анализ влияния архитектурных решений и подходов к обучению, а также практическая апробация модели SciRus-tiny в реальной информационной системе. Глава содержит оригинальные исследовательские результаты и демонстрирует значительный вклад автора в развитие подходов векторизации научных текстов.

Третья глава посвящена созданию многоаспектного тестового набора RuSciBench. Автор описывает методологию отбора и подготовки данных, включающих задачи классификации, регрессии, моно- и кросс-язычного поиска. Приводятся результаты тестирования различных моделей, включая

оценку степени языковой специализации. Значимым результатом является интеграция RuSciBench в международный бенчмарк МТЕВ, что подтверждает высокое качество и востребованность разработанного набора данных.

В четвертой главе рассматривается проблема верификации научных фактов на русском языке и предлагается новая методика формирования датасетов, включающая генерацию утверждений больших языковых моделей, многоступенчатую фильтрацию и экспертную проверку. Автор описывает процесс построения бенчмарка RuSciFact, анализирует характеристики итогового датасета и проводит экспериментальную оценку моделей в задачах поиска и классификации. Глава завершается анализом сложности задач по научным дисциплинам.

В заключении представлены основные результаты работы.

Замечания по диссертации.

1. В работе подробно описана методика создания нового набора RuSciFact, однако не обсуждается альтернативный подход – перевод англоязычного бенчмарка SciFact и его адаптация для русского языка. Такой вариант мог бы существенно уменьшить объем трудозатрат и обеспечить сопоставимость результатов с международными исследованиями.

2. Валидация данных RuSciFact выполняется с привлечением экспертной разметки, однако в работе не приведены показатели согласованности разметчиков (например, Cohen's kappa или Krippendorff's alpha).

3. Внедрение модели SciRus-tiny в eLIBRARY.ru является существенным достижением, однако столь сильная связь практических экспериментов с одним конкретным ресурсом вызывает вопросы о воспроизводимости результатов в других условиях.

4. В работе представлена методика разработки моделей SciRus с использованием собственного предобучения, однако не обсуждаются альтернативы, такие как дообучение существующих базовых моделей.

5. Результаты разработанных моделей на RuSciFact оказываются сравнительно низкими, однако причины этого эффекта не анализируются, что затрудняет интерпретацию сложности задачи или ограничений моделей.

Указанные замечания не влияют на общую положительную оценку диссертации.

Заключение. В целом, считаю, что диссертационная работа Ватолина Алексея Сергеевича соответствует требованиям пунктов 9–14 Положения о присуждении ученых степеней, утвержденного постановлением

Правительства РФ от 24 сентября 2013 года № 842, предъявляемым к кандидатским диссертациям, а соискатель достоин присуждения учёной степени кандидата технических наук по специальности 2.3.8 Информатика и информационные процессы.

Котельников Евгений Вячеславович
доктор технических наук, доцент,
профессор Автономной некоммерческой
образовательной организации высшего образования
«Европейский университет в Санкт-Петербурге»
e.kotelnikov@eu.spb.ru

Kon

19.11.2025 г.

Подпись Котельникова Е.В. заверяю.

Специальность отзв. /присл. А.В. /

