

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доктора физико-математических наук Кузнецова Сергея Олеговича
на диссертационную работу Ватолина Алексея Сергеевича на тему: «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.

Стремительный рост объемов научной информации требует создания эффективных инструментов для ее автоматической обработки. Хотя современные нейросетевые модели показывают высокую точность, передовые решения либо ориентированы исключительно на английский язык, либо существующие мультязычные аналоги требуют значительных вычислительных ресурсов. Это создает барьер для их применения при анализе русскоязычного научного контента.

В связи с этим возникает актуальная задача разработки моделей с небольшим числом параметров и высокой скоростью применения, в том числе на CPU, работающих с русскими и английскими научными текстами. Данная диссертационная работа направлена на решение этой проблемы путем создания не только таких моделей, но и открытых инструментов для их объективной оценки, что определяет ее научную и практическую значимость.

Научная новизна диссертации заключается в разработке новых инструментов для семантического анализа научных текстов. Был предложен набор этапов подготовки данных и обучения двуязычных моделей (SciRus), который позволяет достигать качества, сопоставимого с моделями, значительно превосходящими их по размеру. Кроме того, разработан и опубликован мультизадачный двуязычный бенчмарк RuSciBench для оценки качества векторизации научных текстов. Этот инструментарий, включающий задачи классификации, регрессии и поиска, предоставляет исследователям возможность проводить объективную и воспроизводимую оценку моделей, учитывая специфику русскоязычного научного дискурса, и был интегрирован в международный фреймворк MTEB.

Теоретическая значимость диссертации заключается в развитии методических основ обучения и оценивания мультязычных нейросетевых моделей семантического векторного представления научных текстов с учётом доменной и языковой специфики. В работе систематизировано и экспериментально проанализировано влияние доменно-специфического корпуса, контрастивного обучения на парах «заголовок–аннотация», сигнала из графа цитирований, размера модели и степени её языковой специализации. Полученные результаты уточняют границы применимости универсальных мультязычных моделей к научным текстам, а предложенная методология формирования RuSciFact позволяет отдельно оценивать способность моделей к семантическому поиску и логическому сопоставлению научного утверждения с текстом аннотации.

Практическая значимость исследования определяется разработкой семейства моделей SciRus с небольшим количеством параметров, которые демонстрируют высокое качество семантического представления при значительно меньших требованиях к

аппаратным ресурсам. Это снижает барьеры для внедрения современных технологий семантического поиска в инфраструктуру научных библиотек, университетов и издательств. Ключевым подтверждением практической ценности работы является успешное внедрение модели SciRus-tiny в промышленную эксплуатацию в составе информационно-поисковой системы крупнейшей российской научной электронной библиотеки eLibrary.ru. Публикация в открытом доступе весов разработанных моделей и исходного кода бенчмарков RuSciBench и RuSciFact обеспечивает возможность их свободного использования и дальнейшего развития научным и инженерным сообществом.

Достоверность и обоснованность результатов подтверждаются использованием апробированных методов машинного обучения, проведением экспериментов на репрезентативных массивах реальных научных данных и всесторонним сравнительным анализом разработанных моделей с ведущими мировыми аналогами на общепринятых бенчмарках.

Воспроизводимости исследования и объективности оценки способствуют публикация в открытом доступе исходного кода бенчмарков RuSciBench и RuSciFact, весов обученных моделей SciRus и интеграция оценочного инструментария в международный фреймворк MTEB.

Основные положения и результаты диссертационной работы докладывались на международных конференциях и 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.

Структура и содержание диссертации

Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложения.

Первая глава носит обзорно-теоретический характер. В ней рассматриваются основы построения семантических векторных представлений, начиная от базовых статистических подходов и заканчивая современными моделями на архитектуре трансформер. Детально анализируются ключевые этапы: предобучение моделей (в частности, маскированное языковое моделирование) и их последующее дообучение для задач семантической близости с использованием контрастивных подходов. Глава завершается выводом о том, что существующие передовые решения ориентированы на английский язык, что формирует научную задачу диссертации.

Во **второй главе** описывается разработка, обучение и оценка семейства двуязычных моделей SciRus, предназначенных для векторного представления научных текстов на русском и английском языках. В главе описывается двухэтапная методология обучения, включающая маскированное языковое моделирование на большом корпусе научных статей и последующее контрастивное дообучение на парах «заголовок-аннотация» и данных о цитированиях. Приводятся результаты, демонстрирующие высокую конкурентоспособность моделей SciRus на англоязычных бенчмарках, и описывается их успешное практическое внедрение в поисковую систему eLibrary.ru.

В **третьей главе** решается задача создания инструментария для объективной оценки моделей. В главе представлена разработка и апробация RuSciBench — мультизадачного и двуязычного бенчмарка для оценки качества векторного представления научных текстов. Бенчмарк включает в себя набор задач классификации, регрессии и информационного поиска. Представлены результаты масштабного тестирования, которые показывают, что разработанные в диссертации доменно-адаптированные модели SciRus превосходят как большинство универсальных мультиязычных моделей, так и англоязычные модели для научного домена на англоязычной части бенчмарка.

В **четвертой главе** исследуется более сложная задача верификации научных фактов на русском языке и создается для ее оценки бенчмарк RuSciFact. Описывается оригинальная полуавтоматическая методология создания набора данных, сочетающая генерацию утверждений с помощью больших языковых моделей и последующую экспертную валидацию. Экспериментальные результаты показывают, что, в отличие от поиска релевантного контекста, логическое определение поддержки или опровержения факта представляет существенную сложность для исследованных моделей векторизации, что указывает на необходимость дальнейшего развития соответствующих методов.

Замечания:

1. В работе продемонстрировано преимущество моделей SciRus, однако было бы целесообразно более детально проанализировать факторы этого успеха. Более глубокое раскрытие влияния доменной адаптации обучающих данных и особенностей методологии дообучения усилило бы теоретическую ценность работы.
2. При описании задач классификации в бенчмарке RuSciBench в качестве основной метрики выбрана точность (Accuracy). Хотя автор указывает на предварительную балансировку классов, было бы полезно детальнее обосновать выбор именно этой метрики в сравнении с альтернативами, такими как F1-мера.

Следует отметить, что указанные замечания не снижают значимости и научной ценности работы, не затрагивают её основные выводы.

Общее заключение. Диссертация Ватолина Алексея Сергеевича на тему «Обучение и оценивание мультиязычных нейросетевых моделей семантического векторного представления научных текстов» является завершённой научно-квалификационной работой, в которой решается актуальная научная задача, имеющая важное научное и практическое значение.

Представленные в работе результаты и выводы корректно обоснованы, обладают научной новизной и достоверностью. Основные результаты работы опубликованы 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.

Таким образом, диссертация соответствует критериям, установленным пунктами 9–14 Положения о присуждении учёных степеней, утверждённого постановлением

Правительства Российской Федерации от 24 сентября 2013 г. № 842, а её автор заслуживает присуждения учёной степени кандидата технических наук по специальности 2.3.8 «Информатика и информационные процессы».

Официальный оппонент
доктор физико-математических наук,
профессор, руководитель
Департамента анализа данных и искусственного интеллекта
Федерального государственного
автономного образовательного
учреждения высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»
Адрес: 109028, г. Москва, Покровский бульвар, д. 11
Телефон: +7 (495) 772-95-90 доб. 27318
e-mail: skuznetsov@hse.ru

Кузнецов С. О.

27.07.2026

