

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. Предложен набор этапов подготовки данных и обучения двуязычных моделей (SciRus), который позволяет достигать качества, сопоставимого с моделями, значительно превосходящими их по размеру. Разработан и опубликован первый русскоязычный тест 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 работы заключается в развитии методов построения семантических векторных представлений для научных текстов, особенно в контексте двуязычных корпусов. В диссертации предложена и экспериментально обоснована эффективность двухэтапной методики обучения компактных моделей, которая включает предобучение на задаче маскированного языкового моделирования и последующее контрастивное дообучение. Исследование показывает, что использование легкодоступных пар «заголовок-аннотация» позволяет создавать модели, сопоставимые по качеству с более сложными и ресурсоемкими аналогами, что вносит вклад в теорию доменной адаптации и обучения нейросетевых моделей. Кроме того, разработка методологии создания специализированных тестов, таких как RuSciBench и RuSciFact, и анализ языковой специализации моделей расширяют представления о методах оценки качества.

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 тестов RuSciBench и RuSciFact обеспечивает возможность их свободного использования и дальнейшего развития научным и инженерным сообществом.

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

Воспроизводимость исследования и объективность оценки гарантируются публикацией в открытом доступе исходного кода разработанных тестов RuSciBench и RuSciFact, весов обученных моделей SciRus, а также интеграцией оценочного инструментария во фреймворк MTEB.

Основные положения и результаты диссертационной работы докладывались на международных конференциях и изложены в рецензируемых научных изданиях, индексируемых в базах Scopus и Web of Science.

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 тесте 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 основных выводов.

Общее заключение. Диссертация Ватолина Алексея Сергеевича на тему «Обучение и оценивание мультязычных 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 трудах, индексируемых в базах Scopus и Web of Science, докладывались автором на международных конференциях.

Таким образом, диссертация соответствует требованиям ВАК РФ («Положение о присуждении ученых степеней», утвержденное постановлением правительства РФ от 24

сентября 2013 года № 842), а ее автор — Ватолин Алексей Сергеевич — заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.8 — «Информатика и информационные процессы».

Официальный оппонент

доктор физико-математических наук,

профессор Федерального государственного

автономного образовательного

учреждения высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

Телефон: +7 (495) 772-95-90 доб. 27318

e-mail: skuznetsov@hse.ru

С.О. Кузнецов

06.11.2025

Подпись заверяю

