

ОТЗЫВ

на автореферат диссертации Ватолина Алексея Сергеевича на тему «Обучение и оценивание мультязычных нейросетевых моделей семантического векторного представления научных текстов», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.8 «Информатика и информационные процессы».

Диссертационная работа А.С. Ватолина представляет собой целостное и методологически выверенное исследование, направленное на решение одной из проблем современной научной информатики – эффективной семантической векторизации научных публикаций, количество которых постоянно увеличивается. Актуальность работы не вызывает сомнений, поскольку предложенные автором решения напрямую отвечают на вызовы, связанные с необходимостью глубокого анализа текстовых данных в условиях информационного перенасыщения.

В основе исследования лежит разработка семейства специализированных двуязычных моделей SciRus, ключевой особенностью которых является достижение баланса между высоким качеством семантического представления и вычислительной эффективностью. Автор убедительно демонстрирует, что предложенный им двухэтапный подход к обучению, грамотно сочетающий общее языковое моделирование с последующим контрастивным дообучением на доменных данных, позволяет создавать компактные модели, не уступающие по своим характеристикам значительно более крупным универсальным аналогам.

Другим, не менее важным направлением работы является создание специализированных наборов данных, на которых можно провалидировать разработанные в рамках диссертационной работы языковые модели. Автором были разработаны новые русско-английские бенчмарки RuSciBench и RuSciFact, которые предоставляет научному сообществу первые стандартизированные инструменты для измерения качества векторных представлений и способности моделей к верификации фактов в русскоязычном научном домене. Также важным фактором является включение RuSciBench в международный бенчмарк для оценки качества ММТЕВ, что показывает признание мировым сообществом пользы разработанных инструментов. Этот шаг не только обеспечивает воспроизводимость результатов, но и закладывает основу для будущих исследований.

Практическая значимость диссертации подтверждается внедрением модели SciRus-tiny на портале eLibrary.ru, что свидетельствует о востребованности и промышленной готовности предложенных решений. Не менее важным практическим результатом является публикация моделей, бенчмарков и кодовой базы для оценки в открытый доступ.

В качестве замечаний, которые не влияют на общую положительную оценку, можно отметить следующее: в работе рассматривается задача регрессии для предсказания количества цитирований на основе семантических векторов (раздел 3.2). Однако известно,

что количество цитирований является кумулятивной величиной, существенно зависящей от времени, прошедшего с момента публикации (возраста статьи). В автореферате недостаточно подробно поясняется, как именно учитывался этот временной фактор при постановке задачи и оценке качества: использовалась ли нормализация целевой переменной по годам или модель обучалась предсказывать абсолютные значения исключительно по тексту, что могло затруднить корректное сравнение статей разных лет издания. Также одна из задач бенчмарка RuSciFact, а именно информационно-поисковая, вероятно достаточно проста для современных моделей векторизаторов, так как лучшие из представленных моделей показывают метрики качества близкие к 100%. Сформулированное в автореферате предположение о причинах подобных метрик «Это свидетельствует о том, что задача поиска релевантного контекста для современных моделей практически решена» не является корректным, так как задача решена только в данном конкретном случае. Во-вторых не хватает сравнения с классическими подходами (BM25), которые до сих пор не уступают нейросетевым решениям в ряде случаев.

В заключение, диссертация Ватолина Алексея Сергеевича является завершенной научно-квалификационной работой, в которой успешно решается важная научная задача. Работа обладает научной новизной, практической ценностью и соответствует всем требованиям, предъявляемым ВАК РФ к кандидатским диссертациям. Автор, Ватолин Алексей Сергеевич, в полной мере заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.8 «Информатика и информационные процессы».

Старший научный сотрудник НИВЦ МГУ,

канд. физ.-мат. наук,

Тихомиров Михаил Михайлович

