Отзыв научного руководителя

на диссертационную работу Беляевой Оксаны Владимировны на тему:
“Автоматическое восстановление структуры текстовых документов”,

представленную на соискание ученой степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Беляева О.В. в 2018 году получила степень магистра с отличием по направлению “Программная инженерия” в Федеральном государственном бюджетном образовательном учреждении высшего образования “Московский государственный технический университет имени Н.Э. Баумана”. С 2019 года Беляева О.В. активно работает в научной группе, исследующей методы распознавания изображений сканированных документов.

С 2022 будучи аспирантом ИСП РАН, Беляева О.В. начала руководить исследованиями в области автоматической интеллектуальной обработки электронных текстовых документов неструктурированных форматов и распознаванию изображений с использованием методов машинного обучения. В процессе работы над диссертацией, под ее руководством было выполнено несколько дипломных проектов студентов факультетов ВМК МГУ, ФУПМ МФТИ, МИЭМ ВШЭ.

Актуальность темы диссертационного исследования Беляевой О.В. обусловлена необходимостью создания эффективных методов автоматической обработки неструктурированных документов, позволяющих обрабатывать интенсивные потоки данных и приводящих поступающие разнородные документы к единому формату, пригодному для дальнейшего интеллектуального анализа.

Научный вклад диссертационной работы заключается в создании эффективного метода восстановления иерархической структуры в произвольных неструктурированных документах. Беляева О.В. предложила решение, которое превосходит другие существующие методы, что было

подтверждено победой в независимом международном конкурсе FinTOC 2022 (Марсель, Франция). Кроме того, для увеличения применимости предложенного метода восстановления структуры, Беляева О.В. создала метод обработки некорректных PDF-документов, который существенно снижает время обработки и повышает точность извлечения текстовой информации.

Результаты научных исследований были реализованы в системе Dedoc. Разработанная система была на конкурсной основе поддержана грантом ФСИ для проектов с открытым кодом, и сейчас продолжает активно развиваться. Особенностью системы является ее расширяемость и поддержка автоматической структуризации множества форматов документов. Dedoc является одним из ключевых сервисов платформы Talisman и внедрен в ряде организаций, как в составе Платформы, так и в других решениях, что подтверждено актами о внедрении.

Диссертационная работа Беляевой О.В. представляет собой завершенное научное исследование, которое вносит значимый вклад в развитие области автоматической интеллектуальной обработки текстовых документов и содержит результаты, полезные для дальнейшего развития данной научной области.

Считаю, что диссертационная работа полностью соответствует требованиям ВАК к кандидатским диссертациям по специальности 2.3.5 “Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”, и ее автор, Беляева Оксана Владимировна, заслуживает присуждения ученой степени кандидата технических наук.

Кандидат физико-математических наук, заведующий отделом Информационных систем ИСП РАН — Турдаков Денис Юрьевич

6 февраля 2025 года