ОТЗЫВ

официального оппонента к.ф.-м.н. Девяткина Дмитрия Алексеевича на диссертационную работу Бизюковой Надежды Юрьевны «Формирование знаний о биологической активности низкомолекулярных органических соединений на основе автоматизированного анализа текстов», представленную на соискание ученой степени кандидата биологических наук по специальности 1.5.8. – Математическая биология, биоинформатика

Актуальность темы исследования

Актуальность диссертационного исследования Бизюковой Надежды Юрьевны определяется наличием существенного разрыва между объёмом доступной биомедицинской информации и возможностями её системного использования в научных исследованиях. В последние годы наблюдается устойчивый рост числа публикаций, патентов и иных источников данных, содержащих сведения о биологической активности химических соединений, что приводит к формированию высокоразмерных и слабо структурированных информационных массивов.

Существующие подходы к интеллектуальному анализу текстов, включая методы распознавания сущностей и извлечения отношений, как правило, ориентированы на решение частных задач (например, выделение отдельных типов объектов или бинарных связей) и не обеспечивают представления знаний, необходимого для анализа биологической активности соединений в полном объёме. В частности, отсутствуют универсальные решения, позволяющие одновременно учитывать разнообразие типов биомедицинских сущностей, вариативность их представления в текстах и необходимость их нормализации с использованием внешних онтологий и баз данных.

В этой связи разработка воспроизводимого и масштабируемого подхода к извлечению, нормализации и структурированию сведений о биологической активности низкомолекулярных органических соединений из текстов научных публикаций представляет собой актуальную научную задачу. Особую значимость приобретают методы, обеспечивающие переход от фрагментарных текстовых упоминаний к формализованным представлениям знаний, пригодным для последующего анализа, интеграции и использования в задачах биоинформатики и медицинской химии.

Структура диссертационной работы

Текст диссертации изложен на 150 страницах и включает в себя 17 рисунков и 12 таблиц. Работа состоит из введения, обзора литературы, описания материалов и методов, результатов работы и их обсуждения, заключения и выводов. Кроме того, текст содержит разделы «Список сокращений», «Финансирование работы», «Список литературы» (188 источников) и 2 приложения.

Характеристика диссертации

Диссертация имеет логичную структуру и включает введение, обзор литературы, раздел «Материалы и методы», результаты и их обсуждение, заключение и выводы. Содержание работы последовательно отражает этапы постановки и решения поставленной научной задачи.

Обзор литературы охватывает основные направления развития методов автоматизированного анализа биомедицинских текстов. Рассмотрены подходы к формированию текстовых коллекций, распознаванию наименований биомедицинских объектов и извлечению взаимосвязей между ними. Проведённый анализ позволяет выявить ограничения существующих решений и обосновать необходимость разработки подхода, реализованного в диссертационной работе.

Раздел «Материалы и методы» изложен подробно и содержит описание всех ключевых этапов исследования. Представлены подходы к формированию репрезентативных коллекций текстов, включая использование MeSH-терминов и методов машинного обучения, методы распознавания биомедицинских сущностей различных типов, а также алгоритмы извлечения и нормализации ассоциаций. Отдельное внимание уделено структуре создаваемой базы данных и принципам её наполнения.

Раздел, посвящённый результатам исследования, построен последовательно и отражает переход от разработки отдельных компонентов метода к их интеграции и практическому применению. Приведены результаты оценки качества разработанных алгоритмов, а также примеры применения предложенного подхода для анализа различных биомедицинских задач, включая исследования противовирусной активности и молекулярных механизмов заболеваний.

Текст диссертации изложен в целом последовательно, с соблюдением научного стиля. Материал структурирован, основные положения работы выделены и логически связаны между собой. Иллюстрации и таблицы способствуют пониманию представленных результатов. Отмечаются отдельные неточности формулировок, не оказывающие существенного влияния на общее восприятие работы.

Научная новизна

Научная новизна диссертационной работы заключается в разработке целостного подхода к автоматизированному извлечению сведений о биологической активности низкомолекулярных органических соединений, ориентированного на последовательное решение взаимосвязанных задач обработки биомедицинских текстов. В отличие от существующих исследований, в которых, как правило, рассматриваются отдельные этапы анализа (например, только распознавание сущностей или извлечение связей), в данной работе предложено решение, объединяющее формирование текстовых коллекций, идентификацию биомедицинских объектов, установление взаимосвязей между ними и их последующую нормализацию в рамках единого методического подхода.

Отдельный вклад представляет предложенный подход к извлечению ассоциаций, учитывающий не только наличие совместного упоминания объектов, но и их взаимное расположение в тексте, что позволяет повысить точность выявления связей. Дополнительным

элементом новизны является реализация этапа нормализации извлекаемых данных с использованием внешних онтологий и баз знаний, что обеспечивает сопоставимость результатов и возможность их интеграции с существующими информационными ресурсами.

Таким образом, научная новизна работы определяется разработкой воспроизводимого и методически обоснованного подхода к извлечению и структурированию знаний из биомедицинских текстов, обеспечивающего согласованное решение ключевых задач анализа текстовой информации в данной предметной области.

Теоретическая и практическая значимость

Теоретическая значимость диссертационной работы связана с развитием подходов к системному представлению знаний, извлекаемых из биомедицинских текстов. В работе показано, каким образом разнородные сведения, содержащиеся в научных публикациях, могут быть приведены к единой структуре, обеспечивающей их сопоставимость и дальнейший анализ. Существенным является также уточнение роли отдельных этапов обработки текстов в формировании итогового качества извлекаемых знаний.

Практическая значимость работы определяется возможностью использования разработанных решений при анализе больших массивов научной литературы. Полученные результаты могут применяться для поддержки исследований, связанных с изучением биологической активности соединений, включая сопоставление данных из различных источников, выявление недостаточно изученных направлений и формирование обоснованных научных гипотез.

Дополнительное значение имеет созданная база знаний, позволяющая аккумулировать и структурировать информацию о химических соединениях и их биологических эффектах, что делает возможным её использование как в исследовательской, так и в прикладной практике, в том числе при решении задач медицинской химии и фармакологии.

Обоснованность научных выводов и положений

Обоснованность полученных результатов определяется комплексным характером проведённого исследования, включающего как разработку методов, так и их проверку на тематических коллекциях биомедицинских текстов различного объёма и направленности. Выводы работы основаны на результатах последовательного анализа всех этапов обработки данных – от формирования выборок публикаций до извлечения и сопоставления ассоциаций между биомедицинскими объектами.

Достоверность научных положений подтверждается согласованностью полученных результатов с данными существующих фактографических ресурсов, а также анализом полноты извлекаемой информации. В работе показано, что предложенный подход позволяет воспроизводить значительную часть известных сведений и дополнять их новыми связями, выявленными в текстах научных публикаций. Существенным фактором является также использование нормализации данных с опорой на внешние базы и онтологии, что повышает сопоставимость результатов.

Дополнительным подтверждением обоснованности результатов является проведение вычислительных экспериментов на различных тематических выборках, включая задачи,

связанные с анализом противовирусной активности соединений и исследованием молекулярных механизмов заболеваний. Это позволяет говорить о воспроизводимости полученных результатов и устойчивости предложенного подхода при изменении предметной области.

Основные положения диссертационной работы прошли апробацию на российских и международных научных мероприятиях и отражены в публикациях в рецензируемых изданиях, что свидетельствует о признании полученных результатов научным сообществом.

Вопросы и замечания

1. В диссертации заявлен интегральный характер подхода, однако вопрос распространения ошибок между последовательными этапами обработки раскрыт недостаточно подробно. Поскольку результат зависит от качества отбора публикаций, распознавания сущностей, извлечения ассоциаций и нормализации, представляло бы интерес увидеть более явный анализ того, как ошибки ранних этапов влияют на конечное качество базы знаний. Такой поэтапный анализ в явном виде в работе не представлен.
2. Существуют работы, в которых показано высокое качество (по F1-мере) извлечения отношений между сущностями (например, белками и соединениями) с применением нейронных сетей «Трансформеров», однако в настоящей работе они не используются. Необходимо пояснить причины отказа от использования сетей-Трансформеров при извлечении ассоциаций.
3. При описании извлечения ассоциаций указано, что для части связей фактически используется совместная встречаемость объектов в тексте ввиду отсутствия выраженных семантических маркеров. В диссертации недостаточно подробно обсуждён риск появления ложноположительных ассоциаций в таких случаях, особенно для текстов с высокой плотностью терминов. Было бы полезно более чётко развести ассоциации, извлекаемые по семантическим шаблонам, и ассоциации, основанные лишь на совместном упоминании. Полезно, также, проанализировать применимость использования синтаксических связей между сущностями для снижения доли ложноположительных результатов.
4. Предложенные в диссертации методы апробировались путем анализа аннотаций научных публикаций из открытых ресурсов. Было бы полезно оценить качество работы методов на полных текстах, в которых потенциально при описании биологической активности соединений могут использоваться кореференции.
5. В тексте диссертации на стр. 40 отсутствует описание параметров оценок качества классификации текстов и их фрагментов (TP,FN, FP). Не указано также, каким образом усреднялись эти параметры: на уровне словоупотреблений текстов, или на уровне сущностей.

Перечисленные замечания и вопросы носят в основном уточняющий характер, не снижают общей положительной оценки работы и не умаляют значимости полученных результатов.

Заключение

В целом диссертационная работа Бизюковой Надежды Юрьевны «Формирование знаний о биологической активности низкомолекулярных органических соединений на основе автоматизированного анализа текстов» представляет собой методически проработанное и логически завершённое исследование, в котором решена актуальная научная задача, связанная с разработкой подходов к извлечению и структурированию знаний из биомедицинских текстов. По уровню научной новизны, теоретической и практической значимости диссертация соответствует требованиям пункта 9 «Положения о порядке присуждения ученых степеней», утвержденного Постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842 (в действующей редакции), предъявляемым к диссертациям на соискание ученой степени кандидата наук.

Автор диссертации, Бизюкова Надежда Юрьевна, заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8 – «Математическая биология, биоинформатика».

Официальный оппонент:

Кандидат физико-математических наук,

Девяткин Дмитрий Алексеевич

«3». апреля 2026

Руководитель 73 отдела Федерального государственного учреждения "Федеральный исследовательский центр "Информатика и управление" Российской академии наук".

Кандидатская диссертация защищена по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Адрес основного места работы: ул. Вавилова, д.44, кор.2, г. Москва, 119333, Россия. Адрес электронной почты: devyatkin@isa.ru.

Телефон: +79295045184.

Собственноручную подпись Девяткина Дмитрия Алексеевича удостоверяю

Подпись Д.А. Девяткина заверяю
Ученый секретарь ФИЦ ИУ РАН
д.т.н. В.Н. Захаров
3 04 2026 г.

«3». 04 2026