УДК 004.891.2

МУЛЬТИАГЕНТНАЯ АРХИТЕКТУРА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ МЕДИЦИНСКИХ ЗАПИСЕЙ

Золин И. М.[1]
Научный руководитель – канд. физ.-мат. наук, Чуканов В. С.[1]
[1]Университет ИТМО
zolin.work@yandex.ru

Работа выполнена в рамках темы НИР №15054 «Формирование подходов и методов комплексной оценки цифровых технологий для целей профилактики, диагностики и лечения социально-значимых заболеваний».

Введение

Рост объёмов данных реальной клинической практики (Real-World Data, RWD) создаёт значительный потенциал для проведения мультицентровых исследований, развития медицинской аналитики и внедрения систем поддержки принятия решений [1]. Однако повторное использование данных электронных медицинских карт (ЭМК) остаётся затруднённым вследствие низкой интероперабельности медицинских информационных систем, гетерогенности форматов хранения информации, широкого использования свободного текста и наличия персональных данных, подлежащих защите [2].

Несмотря на развитие международных стандартов медицинских данных, включая SNOMED CT, LOINC, HL7 FHIR и OMOP CDM [3], а также функционирование национальной нормативно-справочной информации (НСИ) Минздрава РФ [4], процессы извлечения и мэппинга данных по-прежнему остаются трудоёмкими и во многом выполняются вручную. Современные достижения в области больших языковых моделей и генеративного ИИ демонстрируют потенциал для автоматизации клинического NLP и извлечения знаний из ЭМК [5], однако требуют архитектурных решений, обеспечивающих контролируемость и валидацию результатов.

Целью работы являлась разработка и валидация мультиагентной системы для автоматизированного извлечения, обезличивания, структурирования и семантической гармонизации медицинских данных.

Основная часть

Система реализована в виде мультиагентного конвейера с центральным оркестратором, обеспечивающим модульную и масштабируемую обработку документов. Система поддерживает обработку текстов, таблиц, PDF и сканов медицинских документов.

Агент парсинга обеспечивает унификацию входных данных и подготовку их к последующим этапам обработки. Для распознавания сканированных документов реализован OCR/VLM-агент, использующий vision-language модели для извлечения текста и восстановления структуры документа. Полученные результаты демонстрируют высокую точность распознавания (Average Similarity до 94,5%), что подтверждает применимость современных VLM-подходов к задачам медицинского OCR.

Для обеспечения конфиденциальности данных разработан агент детекции и анонимизации персональных данных. Учитывая риски утечки чувствительной информации при использовании языковых моделей [6], применён комбинированный подход, объединяющий NER-библиотеки и развёрнутые во внутреннем контуре LLM. По результатам тестирования на корпусе NEREL [7] достигнут F1-score до 0,87, что превышает показатели отдельных библиотечных решений и подтверждает эффективность гибридной стратегии.

Агент структуризации выполняет извлечение ключевых медицинских сущностей (диагнозы, анамнез, лабораторные показатели, медикаменты) с формированием машиночитаемого JSON-представления в соответствии с формально заданными схемами (моделями данных). Реализован механизм автоматического сопоставления извлечённых сущностей с национальными и международными стандартами (НСИ Минздрава РФ, SNOMED CT, LOINC, МКБ-10, HL7 FHIR, OMOP CDM), что обеспечивает семантическую интероперабельность данных. Агент валидации осуществляет контроль полноты и корректности структурированных данных по сравнению с исходным текстом.

Для практического использования системы разработаны десктопное приложение на PyQt и веб-сервис на базе FastAPI с контейнеризацией Docker и возможностью развёртывания в облачной инфраструктуре.

Выводы

Разработана мультиагентная система, обеспечивающая автоматизированную обработку ЭМК с учётом интероперабельности и защиты данных. Использование генеративного ИИ совместно с формализованными моделями стандартов снижает трудозатраты и повышает воспроизводимость исследований на основе RWD.

В будущем планируется расширение функционала и комплексная валидация: интеграция и модификация AutoDP [8] для генерации и вариативности моделей данных и прогнозирования диагноза, улучшение OCR (включая рукописные документы), расширение модальностей (КТ, МРТ), внедрение предобработки и сравнительное тестирование на benchmark- и реальных данных.

Литература

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395-405. doi: 10.1038/nrg3208. PMID: 22549152.
2. Методика оценки качества данных электронных медицинских карт. Webiomed, 2023.
3. International review for a national interoperability framework: summary of commonly used health data interoperability standards / HealthData@IE. — Dublin: Health Information and Quality Authority, Nov. 2025. — 96 p. — URL: https://www.hiqa.ie/sites/default/files/2025-11/International-Review-For-a-National-Interoperability-Framework.pdf.
4. Федеральный реестр НСИ Минздрава России — https://nsi.rosminzdrav.ru
5. Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2
6. Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 946–959, Online. Association for Computational Linguistics
7. Loukachevitch N., Artemova E., Batura T., Braslavski P., Denisov I., Ivanov V., Manandhar S., Pugachev A., Tutubalina E. // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (Deep Learning for Natural Language Processing Methods and Applications). — Incoma Ltd., 2021. — P. 876–885. — DOI: 10.26615/978-954-452-072-4_100.
8. S. Cui, P. Mitra // NeurIPS 2024 Poster Session (Neural Information Processing Systems Conference).