Онтологический и основанный на LLM подход к гармонизации данных для федеративного обучения в здравоохранении

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1756555

LLM | гармонизация данных | онтологии | федеративное обучение | электронные медкарты

Введение

Семантическая гетерогенность в электронных медицинских картах (ЭМК) ограничивает масштабируемую и конфиденциальную аналитику в здравоохранении. Хотя федеративное обучение (ФО) позволяет проводить совместное моделирование без обмена сырыми данными, оно требует согласованных, соответствующих онтологии представлений. Мы представляем подход к гармонизации данных на основе онтологии и большой языковой модели (БЯМ) для поддержки безопасных, интероперабельных рабочих процессов ФО.

Методы

Мы предлагаем общий двухэтапный конвейер для преобразования или аннотирования клинического текста в заранее определённый формат онтологии. Во-первых, кандидаты на концепты извлекаются из целевого словаря с помощью поиска по семантической близости на основе эмбеддингов или кросс-ссылок онтологии. Во-вторых, БЯМ выступает в роли семантического валидатора, принимающего или отклоняющего кандидатов на основе явных критериев эквивалентности или включения. Подход

является онтологически-агностичным и конфигурируемым; в качестве примера реального использования продемонстрировано отображение в MONDO и HPO. Окончательные принятые отображения были оценены по сравнению с независимой оценкой экспертов-людей.

Результаты

В двух клинических наборах данных согласованность между экспертами и БЯМ достигала 92%, при этом общая производительность варьировалась от 78% до 91% в зависимости от стратегии генерации кандидатов. Извлечение само по себе было недостаточным для надёжного отображения, тогда как валидация на основе БЯМ существенно повысила точность, в то время как дополнительные стратегии извлечения улучшили полноту.

Обсуждение

Предложенный конвейер трансформирует онтологически-ориентированную гармонизацию из ручной задачи эксперта в повторно используемый и конфигурируемый рабочий процесс, подходящий для федеративных исследований в здравоохранении. Путём сочетания извлечения с высокой полнотой и семантической адъюдикации на основе БЯМ подход обеспечивает масштабируемое, конфиденциально-сохраняющее преобразование гетерогенного клинического текста в стандартизированные представления в различных областях.
