Извлечение гериатрических синдромов из выписных эпикризов: новый набор данных, схема аннотирования и первые результаты

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1591050

NLP | гериатрия | извлечение данных | машинное обучение | электронные медкарты

Введение

Гериатрические синдромы (GS), такие как падения, деменция, делирий и недостаточность питания, представляют собой сложные клинические состояния, поражающие пожилых людей, которые затрагивают несколько систем органов и оказывают значительное влияние на качество жизни и уход. GS охватывают различные категории заболеваний и недостаточно представлены в структурированных **электронных медицинских картах (EHR)**. **Обработка естественного языка (NLP)** предоставляет возможность извлекать ценную информацию, связанную с GS, из неструктурированного клинического текста, такого как выписные эпикризы из стационаров. Однако отсутствие высококачественных аннотированных наборов данных ограничивает эффективность моделей NLP в этой области. В данном исследовании представлен вручную аннотированный корпус, предназначенный для обнаружения GS, что позволяет обеспечить более точную идентификацию и классификацию этих синдромов.

Методы

Мы разработали комплексную и детальную схему аннотирования для маркировки 12 распространенных GS в выписных эпикризах, включающую такие ключевые атрибуты, как тип диагноза, отрицание и факт возникновения события. Корпус состоит из 2040 вручную аннотированных выписных эпикризов из больниц Национальной службы здравоохранения (NHS) Лотиана в Шотландии. Чтобы оценить эффективность NLP в извлечении GS, мы поэкспериментировали с несколькими предобученными моделями на основе трансформеров, включая **base BERT** (общего назначения), **BioBERT** (биомедицинского назначения), **BioClinicalBERT** (клинического назначения) и **BERT-cased** (регистрозависимая версия английской модели BERT). Модели были дообучены (fine-tuned) и протестированы для двух типов задач: **распознавание именованных сущностей (NER)** и маркировка на уровне документа. Мы также рассмотрели дополнительную задачу обнаружения контекстной информации для каждого упоминания GS (например, анамнез, подозрение, внутрибольничное состояние). При учете контекстной информации вводятся две новые задачи: **NER-C** и **DL-C**, представляющие собой NER и маркировку на уровне документа с учетом контекста соответственно.

Результаты

Наша оценка показала, что в задаче маркировки на уровне документа модель **BERT-cased** достигла самого высокого показателя **F1-меры** (0,897), а **BioClinicalBERT** показала лучший результат при учете отрицания (F1-мера: 0,888). Для задачи NER модели **BioClinicalBERT** и **BERT-cased** достигли F1-меры 0,883. **Хрупкость** (F1 = 1,0), **Падения** (F1 = 0,973) и **Делирий** (F1 = 0,946) являются сущностями GS с наилучшими показателями. В задаче NER-C модель **BERT-cased** достигла наилучшего показателя F1 (0,692), в то время как **BioBERT** показала худший результат (F1 = 0,658). В NER-C наилучшие результаты были достигнуты для контекстно-зависимых меток падений и хрупкости, особенно когда синдром подразумевался, а не заявлялся эксплицитно. Агрегация на уровне документа помогла уменьшить несоответствия, однако в экспериментах NER использовалось упрощенное представление оригинальных аннотаций, совместимое с форматом CoNLL, в котором прерывистые упоминания были преобразованы в кратчайшие охватывающие интервалы, а перекрывающиеся упоминания — объединены. Таким образом, представленные результаты NER следует интерпретировать

как базовую производительность на упрощенном представлении этих структур, при этом категории GS с низкой частотой встречаемости и редкие контекстные метки также отрицательно сказались на точности модели.

Обсуждение

Данное исследование демонстрирует эффективность NLP для извлечения гериатрических синдромов из неструктурированного клинического текста и представляет вручную аннотированный корпус с подробными руководствами для поддержки этой задачи. Результаты также показывают, что на производительность модели сильно влияет характеристика набора данных. Более частые и лексически четкие синдромы, такие как хрупкость, падения и делирий, достигли наиболее высоких результатов, тогда как более редкие категории и низкочастотные комбинации атрибутов, такие как «подозрение», «направление» и некоторые отрицательные или специфические для контекста метки, поддавались обучению труднее и давали более низкие и менее стабильные баллы. Аналогичным образом, мелкозернистое (fine-grained) аннотирование оказалось более сложным, чем крупнозернистое (coarse-grained), поскольку оно увеличивает разреженность меток и требует от модели различения тонких контекстуальных различий, таких как текущие упоминания в сравнении с упоминаниями из анамнеза, неявные упоминания и внутрибольничное начало заболевания. На извлечение на уровне сущностей дополнительно влияли прерывистые и перекрывающиеся упоминания, которые часто встречаются в клинических повествованиях и затрудняют определение границ, в то время как агрегация на уровне документа снижала влияние этих локальных ошибок и, следовательно, обеспечивала более высокие баллы. Эти результаты подчеркивают, что распределение данных, сложность аннотирования и структура упоминаний напрямую влияют на производительность модели и должны быть центральными аспектами при планировании будущих работ по извлечению гериатрических синдромов.