Гранулярность по дизайну: подход к тонкозернистому распознаванию именованных сущностей для извлечения контекстно-зависимой информации о лекарствах

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1927272

NLP | извлечение информации | клинические исследования | фармаконадзор | электронные медкарты

Введение

Хотя структурированные электронные медицинские карты (ЭМК) обеспечивают базовую историю приема лекарств, они часто не способны зафиксировать важную в клиническом отношении контекстуальную информацию об использовании лекарственных средств, которая задокументирована только в неструктурированных клинических заметках. Значительная часть критически важной клинической информации — такой как не зарегистрированные побочные эффекты или использование безрецептурных биологически активных добавок — содержится исключительно в неструктурированных клинических заметках в формате свободного текста. Наша цель заключалась в разработке и оценке системы распознавания именованных сущностей (NER) с учетом контекста на детальном уровне, способной преобразовывать информацию о лекарственных средствах из неструктурированных клинических текстов на венгерском языке в структурированные данные, пригодные для клинических исследований.

Методы

Мы разработали систему распознавания именованных сущностей (NER) с учетом контекста на детальном уровне с использованием библиотеки обработки естественного языка spaCy. Клинические документы были получены из исследовательского хранилища данных UDBD-Health Университета Дебрецена. Для классификации упоминаний лекарственных средств по 6 контекстуальным категориям была разработана специализированная схема аннотирования: Текущие лекарства, Отмененные/ Отсутствие лекарств, Побочные эффекты и Аллергии (аббревиатура CAVE). Документы предварительно аннотировались с использованием правил на основе словарей, а затем вручную дорабатывались в программе Label Studio. Для оценки способности модели к обобщению и получения надежной оценки производительности без утечки данных мы использовали пятикратную перекрестную проверку с внутренним разделением на валидационную выборку, при этом выбор модели с помощью ранней остановки использовал подмножество валидации, взятое только из обучающих складок, тогда как каждая внешняя отложенная складка предназначалась исключительно для окончательной оценки. Чтобы справиться с выраженным дисбалансом между контекстными классами лекарственных средств, мы применяли перевзвешивание (oversampling) обучающих примеров редких классов на уровне документа. Производительность модели оценивалась с использованием точности (Precision), полноты (Recall) и F1-меры, которые представлены как среднее значение ± стандартное отклонение по пяти складкам, и сравнивались с правиловым базовым уровнем, построенным на основе той же процедуры сопоставления со словарем, которая использовалась для предварительной аннотации.

Результаты

Для извлечения информации о лекарственных средствах были обучены два типа моделей. Модель с учетом контекста на детальном уровне (CME) достигла общей F1-меры 0,944 ± 0,001 при пятикратной перекрестной проверке. На уровне сущностей категория CURRENT_MED (Текущие лекарства) показала наивысшую F1-меру (0,958 ± 0,001), за ней следовали CAVE_MED (0,900 ± 0,023), NO_MED (0,820 ± 0,005) и AE_MED (0,689 ± 0,026), последняя из которых демонстрировала наибольшую вариабельность от складки к складке. Применение перевзвешивания примеров редких классов на уровне документа (CME OS) привело к достижению общей F1-меры 0,946 ± 0,002 с умеренным улучшением F1-мер для редких классов. Модель грубого уровня (ME), обученная на задаче распознавания единственной сущности

MEDICATION (Лекарственное средство), достигла F1-меры 0,968 ± 0,001 при минимальной дисперсии, существенно превзойдя правиловой базовый уровень на основе сопоставления со словарем для той же задачи (F1 = 0,860 ± 0,003), который показал высокую точность, но сравнительно низкую полноту. Модели были развернуты в институциональном конвейере обработки данных, что позволило автоматизировать структурирование документов о лекарственных средствах из более чем 700 000 документов в год.

Обсуждение

В данном исследовании представлена система извлечения информации о лекарственных средствах с учетом контекста на детальном уровне для клинических документов на венгерском языке. Разработанные модели способствуют преобразованию неструктурированных текстов о лекарственных средствах в структурированные данные, обеспечивая основу для последующих фармакологических, эпидемиологических исследований и исследований реальных данных (real-world evidence). К ограничениям относятся специфичность институциональных данных, недостаточное представительство контекстов побочных эффектов, разделение перекрестной проверки на уровне документа, а не на уровне пациента, а также отсутствие независимой параллельной аннотации, что не позволило рассчитать коэффициент согласованности между аннотаторами. Будущие работы будут сосредоточены на расширении корпуса, межинституциональной валидации и включении дополнительных атрибутов, таких как дозировка и продолжительность приема.

Перевод выполнен: 31.10.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.