Несогласие между моделями как сигнал без использования эталона для приоритизации проверки человеком при транскрибации медицинской речи

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1829902

NLP | автоматическое распознавание речи | диагностика | машинное обучение | медицинская документация

Введение

Ambient AI scribes (фоновые ИИ-ассистенты для ведения записей) генерируют транскрипты в больших масштабах, однако рутинный контроль качества ограничен отсутствием верифицированных человеком эталонных транскриптов в большинстве условий развертывания. Мы оценивали, может ли расхождение между гетерогенными системами **автоматического распознавания речи (ASR — Automatic Speech Recognition)** служить информативным сигналом для локализации неопределенности транскрипции, используя публичный англоязычный корпус медицинской речи вместо записей клинических приемов.

Методы

Восемь коммерческих и открытых систем **ASR** были применены к 50 аудиоклипам медицинского образовательного назначения (8 ч 14 мин). Выходы нескольких моделей были выровнены, и использовалась процедура консенсуса по принципу «исключения одного» (leave-one-out), чтобы оценить согласованность каждой модели, минимизируя при этом цикличность.

Результаты

Расхождения между моделями были редкими и локализованными: 72,1% позиций демонстрировали высокую степень согласованности (7–8 систем были согласны), тогда как только 2,5% составляли позиции высокого риска с минимальной согласованностью (0–3 системы). Регионы с низкой согласованностью систематически были обогащены смыслоразличительными лексическими различиями, определенными как лексические несоответствия после исключения пунктуации, сокращений, числовых данных и вариаций слов-паразитов. Единичный слой валидации с человеческой коррекцией (**HC — Human-Corrected**) показал, что ошибки транскрипции монотонно возрастали по мере снижения согласованности. При иллюстративном апостериорном пороге, маркировка позиций, в которых согласились шесть или менее систем, позволила отобрать 28,6% токенов, при этом выявив 93,7% ошибок, верифицированных одним аннотатором (HC) в данном прокси-корпусе.

Обсуждение

Эти результаты позволяют предположить, что расхождение между моделями может помочь сосредоточить проверку человеком на небольшом количестве регионов транскрипта, вероятно, подверженных ошибкам. Однако согласованность между всеми системами не гарантирует правильности, поскольку общие ошибки могут остаться необнаруженными при таком подходе. Перед операционным развертыванием требуется валидация на реальных данных клинических приемов.
