УДК 004.8

МЕТОД АВТОМАТИЧЕСКОЙ КОРРЕКЦИИ ОРФОГРАФИИ ДЛЯ АНАЛИЗА КЛИНИЧЕСКОГО ТЕКСТА НА РУССКОМ ЯЗЫКЕ

Погребной Д.А. (Университет ИТМО)

Научный руководитель – к.т.н., доцент Ковальчук С.В.

(Университет ИТМО)

Доклад посвящен проблеме исправления орфографических ошибок в русских медицинских текстах для более качественной предобработки исходных данных при обучении моделей машинного обучения в здравоохранении. В работе будет представлен обзор и их сравнительная оценка существующих инструментов для исправления ошибок. И предложен альтернативного подход для коррекции орфографических ошибок.

Введение. Разработка моделей машинного обучения в здравоохранении - это сложный процесс, который включает в себя множество различных задач, помимо простого написания кода. Одним из аспектов построения моделей машинного обучения является подготовка и предобработка исходных данных. Качество финальных моделей сильно зависит от качества данных используемых для обучения. В здравоохранении модели машинного обучения часто обучаются на электронных текстах медицинских карт пациентов, которые представляют собой обычный неструктурированный текст. Такие записи содержат множество орфографических ошибок, которые негативно влияют на качество моделей и снижают их метрики. Однако существуют методы и инструменты для автоматической коррекции орфографических ошибок, которые способны устранить эту проблему и повысить качество моделей с минимальными затратами и без усложнения самих моделей.

Основными этапами автоматического исправления орфографической ошибки является генерация кандидатов на замену некорректному слову и выбор из списка кандидатов наиболее подходящего слова для исправления. Для генерации кандидатов часто используют расстояние редактирования или его модификации, а выбор наиболее подходящего слова осуществляется с помощью различных метрик частотности слов или с использованием языковой модели на базе машинного обучения. Подход с использованием частотности слов игнорирует контекст исправляемого слова, но является более производительным по сравнению с использованием языковых моделей, которые этот контекст учитывают.

Основная часть.

В данной работе проведено сравнение открытых инструментов для коррекции орфографических ошибок в медицинских текстах на русском языке. А также представлен новый подход исправления орфографических ошибок, который оптимизирует вычисление расстояний редактирования для генерации списка слов кандидатов и использует языковую модель учитывающую контекст для выбора наиболее подходящего слова.

Выводы. Предложенный подход позволяет получить более высокую производительность по сравнению с подходами, использующими стандартное вычисление расстояний редактирования и языковые модели на базе машинного обучения. А также достигает более высоких метрик точности по сравнению с подходами на базе частотности слов.

Погребной Д.А. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись