Давайте начнем. Меня зовут Дмитрий Погребной, и я хочу представить свою работу под названием "Разработка инструмента автоматической коррекции орфографии для анализа клинического текста на русском языке".

Начнем с актуальности. Благодаря впечатляющим достижениям в области машинного обучения стало возможным применение различных моделей прогнозирования и принятия решений в медицине. В здравоохранении такие модели часто основываются на электронных текстах медицинских карт пациентов. Качество таких моделей сильно зависит от качества оригинальных медицинских записей, которые обычно представляют собой обычный текст. Такие записи часто содержат множество орфографических ошибок, которые значительно снижают качество конечных моделей. Инструмент для автоматического исправления орфографических ошибок сможет устранить эту проблему и повысить качество моделей без дополнительных затрат.

Поэтому целью данной работы является разработка метода и инструмента автоматического исправления орфографических ошибок для анализа медицинских текстов на русском языке. Инструмент должен принимать необработанный медицинский текст и возвращать исправленный текст с минимальным количеством ошибок.

Для достижения цели были поставлены следующие задачи. Необходимо было выполнить аналитический обзор. Провести первичный анализ и предобработку данных. Выполнить анализ существующих инструментов. И после всего этого реализовать новый инструмент.

В данной работе рассматривается исправление следующих ошибок. Пропущенные буквы, лишние буквы и переставленные местами буквы. А также пропущенные или лишние пробелы.

В качестве данных в данной работе используются боле двух тысяч анамнезов пациентов из медицинского центра имени Алмазова. Все анамнезы токенизированы, отфильтрованы и лемматизированы. В результате почти все анамнезы в корпусе содержат менее 91 токена. Исходя из этого можно сформулировать требуемую производительность будущего инструмента, которая составляет примерно 100 слов в секунду. Этой производительности достаточно, чтобы обрабатывать единичный анамнез почти мгновенно.

Давайте кратко рассмотрим существующие инструменты с открытым исходным кодом, поддерживающие русский язык. Все метрики вычислялись на основании русских медицинских текстов. Error и lexical precision были рассчитаны для каждого инструмента, а также их среднее значение в столбце Overall precision. Также были проведены замеры производительности. В результате инструменты Aspell-python и LangaugeTool-python отличаются своей точностью, но имеют низкую производительность. Также стоит отметить, что инструменты Symspell и Jumspell обладают впечатляющей производительностью, но низкой точностью. Общая низкая точность инструментов связана с тем, что эти инструменты общего назначения и не специализируются на медицинских текстах. Отмечу, что к сожалению, на данный момент не существует инструментов с открытым исходным кодом, специализирующихся на русских медицинских текстах. В данной работе планируется превзойти эти инструменты как по точности, так и по производительности.

Давайте пойдем дальше

На данный момент ведется активная разработка инструмента. А на слайде представлена текущая архитектура. Зеленые компоненты уже реализованы, а черные еще ожидают своей очереди. Инструмент содержит семь компонентов. Preprocessor отвечает разбор входящего текста и преобразование его во внутреннее представление. Error Detector отвечает за обнаружение неправильных слов в тексте. Error Model генерирует список потенциальных кандидатов для исправления. Edit Distance Index это индекс который позволяет значительно ускорить вычисление расстояния редактирования для генерации потенциальных кандидатов. Language Model ранжирует и выбирает наиболее подходящего кандидата из списка кандидатов. Postprocessor собирает вместе исправленный текст из внутреннего представления.

И в результате проведен первичный анализ и обработка данных, выполнен анализ существующих инструментов и начата реализация инструмента для коррекции.

В дальнейшем планируется завершить разработку инструмента и провести его аппробацию.