Давайте начнем. Меня зовут Дмитрий Погребной, я хочу представить свою работу под названием "Метод 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. Метод и инструмент для автоматического исправления орфографических ошибок смогут 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. Предложить новый подход. И после этого реализовать новый инструмент.

В данной работе рассматривается исправление следующих ошибок. Неправильные буквы в слове. 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 была сформулирована необходимая производительность будущего инструмента, которая составляет примерно 100 слов в секунду. Этой производительности достаточно, чтобы обрабатывать единичный анамнез почти мгновенно.

Давайте кратко рассмотрим существующие инструменты для исправления ошибок 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

Предлагается следующий метод исправления ошибок. После определения некорректного слова, с помощью расстояния левенштейна генерируется список возможных исправлений. Отличительной особенностью здесь является, то что для ускорения вычислений используется алгоритм SymDel, который позволяет на порядок ускорить операцию вычисления расстояний благодаря специальному индексу. Далее полученный список кандидатов на исправления ранжируется с помощью модели машинного обучения. В данном случае используется дообученая модель RuBERT.

Давайте пойдем дальше

На слайде представлена архитектура реализованного инструмента. 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.

На данный момент удалось достичь следующих метрик. Error precision – 0.55, lexical precision – 0.78. Производительность 326 слов в секунду. Однако это не конечный результат и еще есть большой простор для улучшений. Например можно усложнить алгоритм работы с

В результате проведен первичный анализ и обработка данных, выполнен анализ существующих инструментов, предложен новый подход для исправления орфографических ошибок в медицинских текстах и реализован прототип инструмента для коррекции.

В дальнейшем планируется протестировать прототип на большем наборе данных и провести полноценную апробацию. Также планируется попробовать дообучить модель на большем наборе данных. Помимо этого планируется провести эксперименты с другими моделями для ранжирования.