Can I start? Ok

Well. Let's start. My name is Dmitry Pogrebnoy and I want to present my work titled "Development of an automatic spelling correction tool for analyzing clinical text in Russian".

Thanks to impressive advances in machine learning, it has become possible to apply various predictive and decision-making models in medicine. In healthcare, such models are often based on electronic texts of patients' medical records. The quality of such models strongly depends on the quality of the original medical records, which are usually plain text. Such records often contain a lot of spelling errors, which significantly reduce the quality of the final models. A high-quality tool for automatic correction of spelling errors will be able to fix this problem and increase the quality of the models without additional costs.

Therefore, the purpose of this work is to develop a tool for automatic correction of spelling errors for the analysis of medical texts in Russian. The tool should accept the raw medical text and return the corrected text with a minimum number of errors.

The following tasks were set for this semester. Make an overview of the Russian medical texts correction. Analyze the existing tools for correcting spelling errors. And also develop the first draft of the architecture of the future tool.

Let's briefly look on the existing open source tools that support the Russian language. The tools were tested on Russian medical texts. Error and lexical precision were calculated for each tool, as well as their average value in the overall precision column. Performance tests were also made. Aspell-python and LangaugeTool-python tools stand out for precision, but they have low performance. It is also worth noting that Symspell and Jumspell tools have high performance, but low precision. The general low precision of the tools is due to the fact that these tools do not specialize in medical texts, but this tools for general purpose. Unfortunately, at the moment there are no open source tools that specialize in Russian medical texts.

To summarize, we can see that, all tools in general have rather low precision and slow performance. As a result, it is necessary to outperform these tools both in precision and performance.

Let's go next

Here is the principal architecture of the tool for correcting spelling errors. It contains five components. Preprocessor is responsible for parsing the incoming text and converting it into an internal representation. Error Detector is responsible for detecting incorrect words in the text. Error Model generates a list of potential candidates for correction, and the Language Model component selects the most suitable one from the list. Postprocessor build together the corrected text from the internal representation.

Let 's move on to the conclusion

As a result, over the past semester, an overview of the correction of Russian medical texts are made. The analysis of existing tools is carried out and the principle architecture of the tool is developed.

In the future, it is planned to complete the development of the architecture, implement an error correction tool and test it on real medical records.

Могу я начать? Хорошо

Хорошо. 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".

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

Поэтому целью данной работы является разработка инструмента для 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.

В этом семестре были поставлены следующие задачи. Сделать обзор коррекции русских медицинских текстов. Проанализировать существующие инструменты для исправления орфографических ошибок. А также разработать первый проект архитектуры будущего инструмента.

Кратко рассмотрим существующие инструменты с открытым исходным кодом, поддерживающие русский язык. Инструменты были протестированы на русских медицинских текстах. Для каждого инструмента были рассчитаны точность ошибок и лексическая точность, а также их среднее значение в колонке общей точности. Также были проведены тесты производительности. Инструменты Aspell-python и LangaugeTool-python выделяются по точности, но имеют низкую производительность. Также стоит отметить, что инструменты Symspell и Jumspell имеют высокую производительность, но низкую точность. Общая низкая точность инструментов объясняется тем, что эти инструменты не специализируются на медицинских текстах, а являются инструментами общего назначения. К сожалению, на данный момент не существует инструментов с открытым исходным кодом, специализирующихся на русских медицинских текстах.

Подводя итог, мы видим, что все инструменты в целом имеют достаточно низкую точность и медленную производительность. В результате необходимо превзойти эти инструменты как по точности, так и по производительности.

Идем дальше

Вот принципиальная архитектура инструмента для исправления орфографических ошибок. Она содержит пять компонентов. Препроцессор отвечает за разбор входящего текста и преобразование его во внутреннее представление. Детектор ошибок отвечает за обнаружение неправильных слов в тексте. Модель ошибок генерирует список потенциальных кандидатов на исправление, а компонент языковой модели выбирает из списка наиболее подходящий. Постпроцессор собирает исправленный текст из внутреннего представления.

Переходим к заключению

В результате за прошедший семестр сделан обзор коррекции русских медицинских текстов. Проведен анализ существующих инструментов и разработана принципиальная архитектура инструмента.

В будущем планируется завершить разработку архитектуры, реализовать инструмент исправления ошибок и протестировать его на реальных медицинских записях.

Переведено с помощью www.DeepL.com/Translator (бесплатная версия)