ITMO UNIVERSITY

Saint Petersburg, Russia

Разработка инструмента автоматической коррекции орфографии для анализа клинического текста на русском языке

Дмитрий Погребной

Руководитель: Ковальчук С.В.

Научная и учебно-методическая конференция
Университета ИТМО
Санкт-Петербург
2022

Актуальность

- Существует множество различных моделей машинного обучения основанных на информации из медицинских карт пациентов.
- Качество таких моделей сильно зависит от качества исходных текстов.
- Электронные карты пациентов обычно представлены в виде простого текста и содержат орфографические ошибки
- Орфографические ошибки значительно снижают качество итоговых моделей и поэтому требуют исправления.

Цели и задачи

Цель: Разработать метод и реализовать инструмент автоматической коррекции орфографии для анализа клинических текстов на русском языке.

Задачи:

- Аналитический обзор
- Первичный анализ и предобработка данных
- Анализ существующих инструментов
- Реализация нового инструмента

Ошибки в текстах

Type of mistake	Incorrect text	Correct text
Missing caracters	туб□ркулез	туберкулез
Extra characters	туберкпулез	туберкулез
Shuffled characters	тубрекулез	туберкулез
Missing word separator	острый|туберкулез	острый_туберкулез
Extra word separator	туб_еркулез	туберкулез

Первичный анализ данных

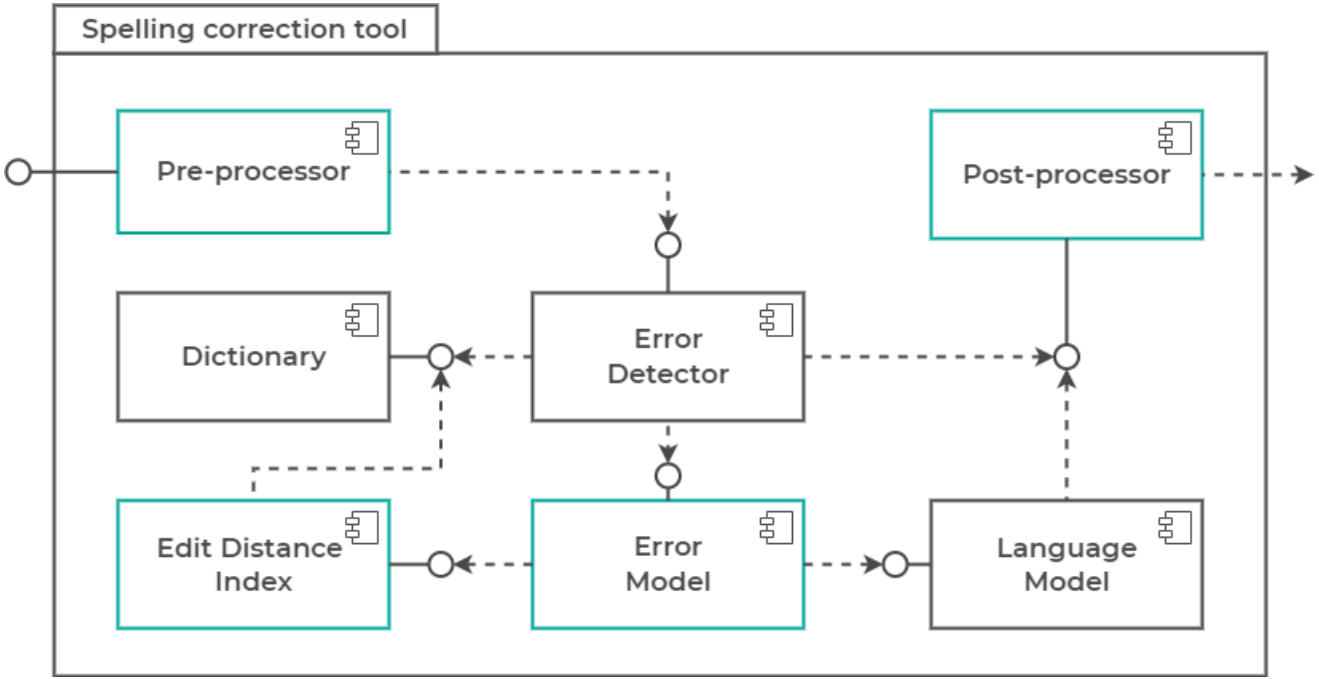
- Корпус из 2356 анамнезов пациентов медицинского центра им. Алмазова
- Анамнезы токенизированы, отфильтрованы и лемматизированы.
- 91 токен - 99 перцентиль количества отфильтрованных токенов в анамнезе
- Необходимая производительность ≈100 слов в секунду

Существующие инструменты

Инструмент	Error precision	Lexical precision	Overall precision	Среднее количество слов в секунду
Aspell-python	0.65	0.775	0.7125	353
PyHunspell	0.59	0.49	0.54	11.5
PyEnchant	0.6	0.455	0.5275	26.4
LanguageTool-python	0.64	0.845	0.7425	19.1
PySpellChecker	0.335	0.765	0.55	4.3
SymspellPy	0.42	0.78	0.6	15892.1
Jumspell	0.395	0.925	0.66	2043.2
Spellchecker prototype	0.41	0.83	0.62	0.07

ITMO UNIVERSITY

Архитектура инструмента



Заключение и дальнейшие планы

- Проведён первичный анализ и обработка данных
- Выполнен анализ существующих инструментов
- Частично реализован инструмент для коррекции

Дальнейшие планы:

- Закончить разработку нового инструмента
- Провести аппробацию полученного инструмента

Thank you for your attention!

www.ifmo.ru

ITMO more than a UNIVERSITY

Тестирование инструментов

- **Error precision** – отношение количества корректно исправленных слов к общему количеству некорректных слов
- 200 медицинских слов с орфографическими ошибками для расчета error precision
- **Lexical precision** – отношение количества неизмененных корректных слов к общему количеству корректных слов
- 200 корректных медицинских текстов для расчета lexical precision
- **Производительность** замерялась на ноутбуке под Ubuntu 20.04 с 16 GB RAM и Intel Core i7-9750H CPU @ 2.60GHz * 12