Университет ИТМО
Факультет Цифровых Трансформаций

МЕТОД АВТОМАТИЧЕСКОЙ КОРРЕКЦИИ ОРФОГРАФИИ ДЛЯ АНАЛИЗА КЛИНИЧЕСКОГО ТЕКСТА НА РУССКОМ ЯЗЫКЕ

Погребной Дмитрий Андреевич

УНИВЕРСИТЕТ ИТМО
КМУ
XI КОНГРЕСС МОЛОДЫХ УЧЕНЫХ

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
  - Предложение нового подхода
  - Реализация нового инструмента

Ошибки в текстах

Type of mistake	Incorrect text	Correct text
Wrong characters	тубиркулез	туберкулез
Missing characters	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

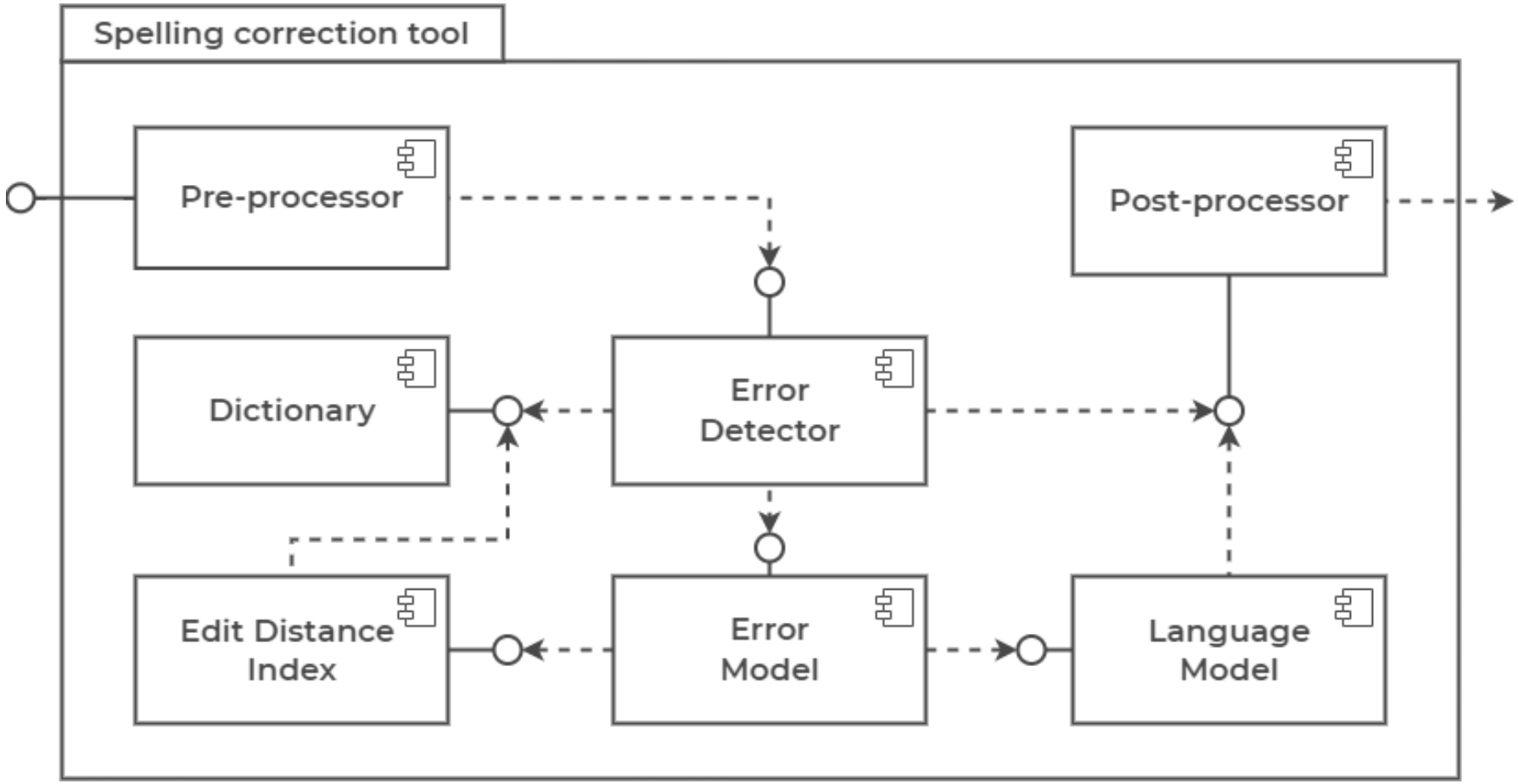
Существующие инструменты

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

Предлагаемый метод

- Модель ошибок генерирует кандидатов для исправления.
- Модель ошибок – расстояние Дамерау-Левенштейна и алгоритм SymDel для ускорения вычислений.
- Языковая модель ранжирует кандидатов для исправления.
- Языковая модель – модель RuBERT.

Архитектура инструмента



Достигнутые метрики

- Error precision – 0,55
- Lexical precision – 0,78
- Производительность ≈326 слов в секунду
- Есть простор для улучшений!

Заключение

- Проведён первичный анализ и обработка данных.
- Выполнен анализ существующих инструментов.
- Предложен новый метод для исправления орфографический ошибок на русском языке.
- Реализован прототип инструмента для коррекции.

Дальнейшие планы

- Протестировать полученный прототип на большем количестве данных.
- Дообучить языковую модель на большем объеме данных.
- В качестве языковой модели использовать другие модели.

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