Разработка системы для сбора данных из открытых источников

Лебедев Станислав Филиппович
руководитель от предприятия: В. Д. Танков
руководитель от факультета: А. А. Шпильман

Санкт-Петербургская школа физико-математических и компьютерных наук
НИУ ВШЭ – Санкт-Петербург

28 июня 2021 г.

Мотивация

Команда “Grazie“ компании JetBrain решает задачи обработки естественного языка:

- Автодополнение
- Суммаризация
- Перефразирование
- Исправление грамматических ошибок

Для этого требуется много данных, открытые источники – отличное решение для их сбора:

- В интеренете много разной информации. Можно набирать датасеты как узкоспециализированные, так и “общего“ плана
- Открытые источники помогают в комерциализации продукта

Задача исправление грамматических ошибок (GEC)

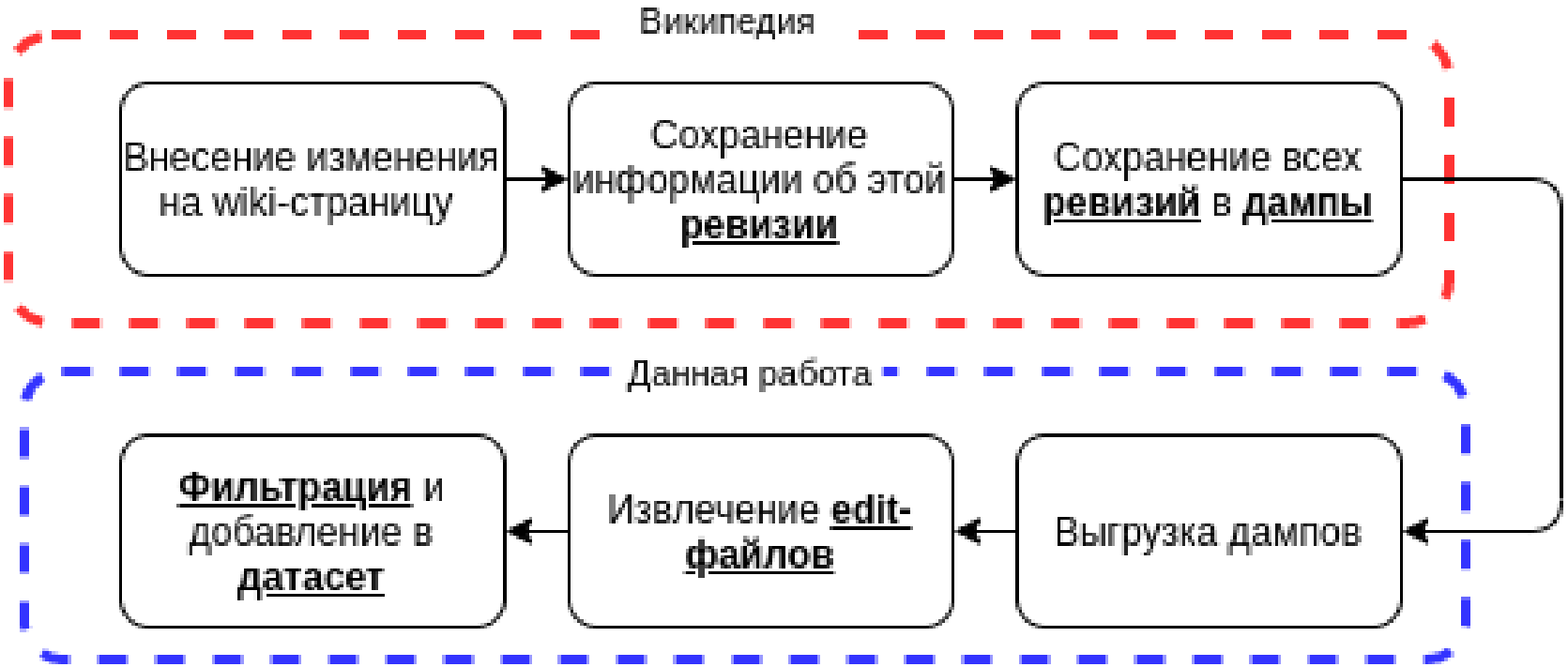
Задача исправления грамматических ошибок – детекция и исправления грамматической ошибки в тексте на естественном языке

Подход схож с описанным в статье: GECToR – Grammatical Error Correction: Tag, Not Rewrite Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, Oleksandr Skurzhanskyi Grammarly 15th Workshop on Innovative Use of NLP for Building Educational Applications (co-located with ACL 2020) [https://arxiv.org/pdf/2005.12592.pdf]

Особенности существующих решений

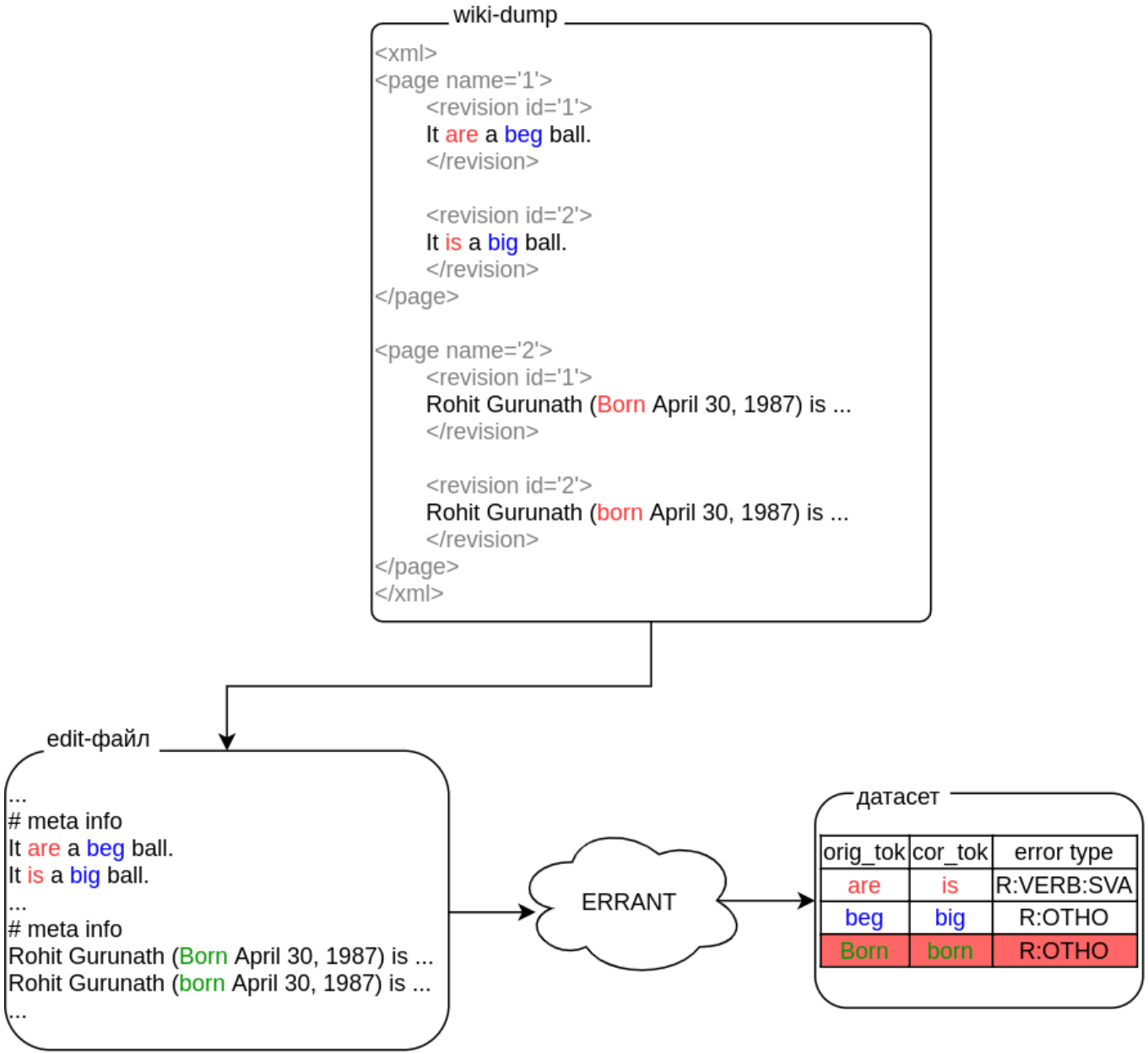
- Некоммерческие. Невозможно создавать коммерческий продукт.
- Сложно и дорого расширять, так как требуется работа живых людей.
- Смещение в область художественной литературы.
- Недостаточно объемны

Схема сбора данных



Ревизии – сохраненное состояние страницы
Дампы – зархивированные наборы ревизий
Edit-файлы – пары предложений с внесенными изменениями
ERRANT – фрейморк для классификации изменений

Подробнее


wiki-dump
<xml>
<page name='1'>
<revision id='1'>
It are a beg ball.
</revision>
<revision id='2'>
It is a big ball.
</revision>
</page>
<page name='2'>
<revision id='1'>
Rohit Gurunath (Born April 30, 1987) is ...
</revision>
<revision id='2'>
Rohit Gurunath (born April 30, 1987) is ...
</revision>
</page>
</xml>
edit-файл
...
meta info
It are a beg ball.
It is a big ball.
...
meta info
Rohit Gurunath (Born April 30, 1987) is ...
Rohit Gurunath (born April 30, 1987) is ...
...
ERRANT
датасет
orig_tok cor_tok error type
are is R:VERB:SVA
beg big R:OTHO
Born born R:OTHO

Цель и задачи

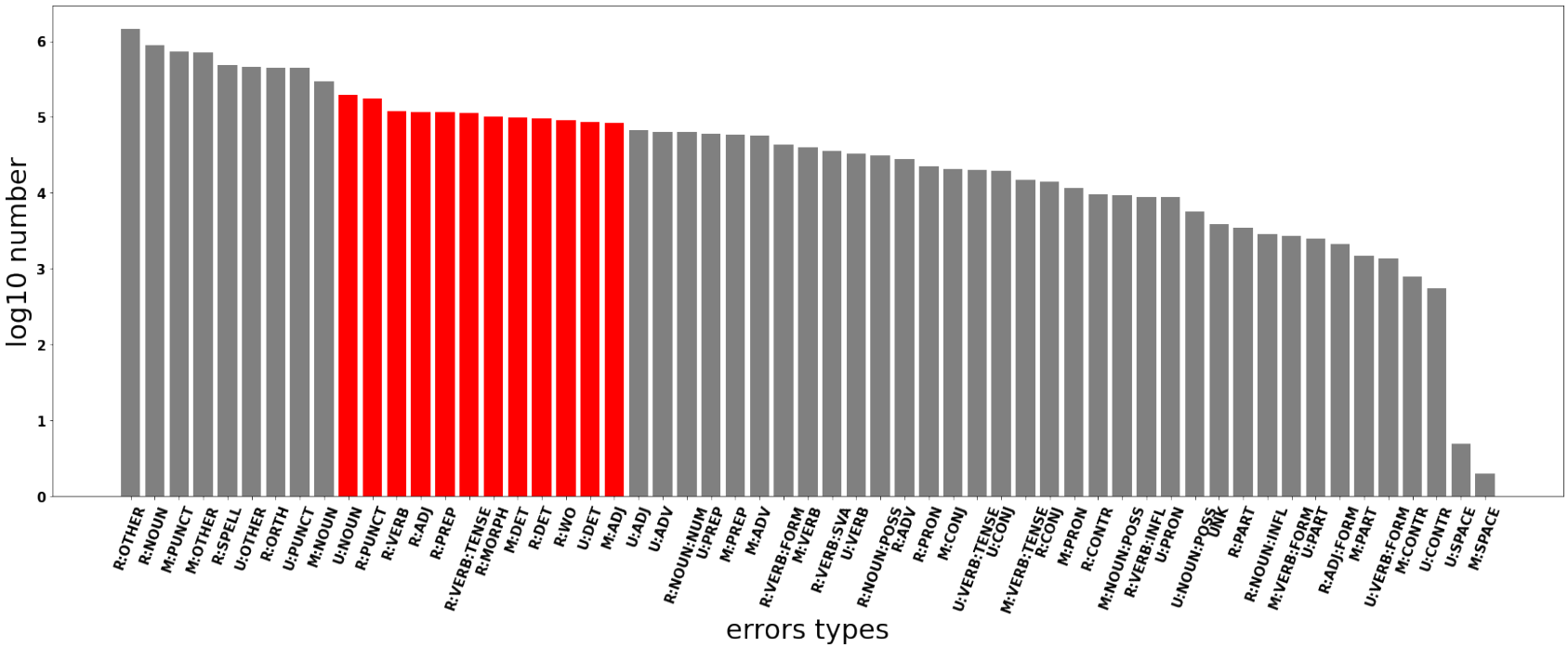
Цель: реализовать инструмент извлечения датасета из Википедии для задачи исправления грамматических ошибок.

Задачи:

- Изучить дампы и выдвинуть гипотезы о методах фильтрации
- Разработать программу для извлечения датасетов в соотвествии с гипотезами
- Собрать датасеты и проверить их качество

Фильтрация по индексу авторов

- Возьмем авторов с больше чем n ревизиями.
- Индекс – доля ревизий, попавших в класс “хороших“



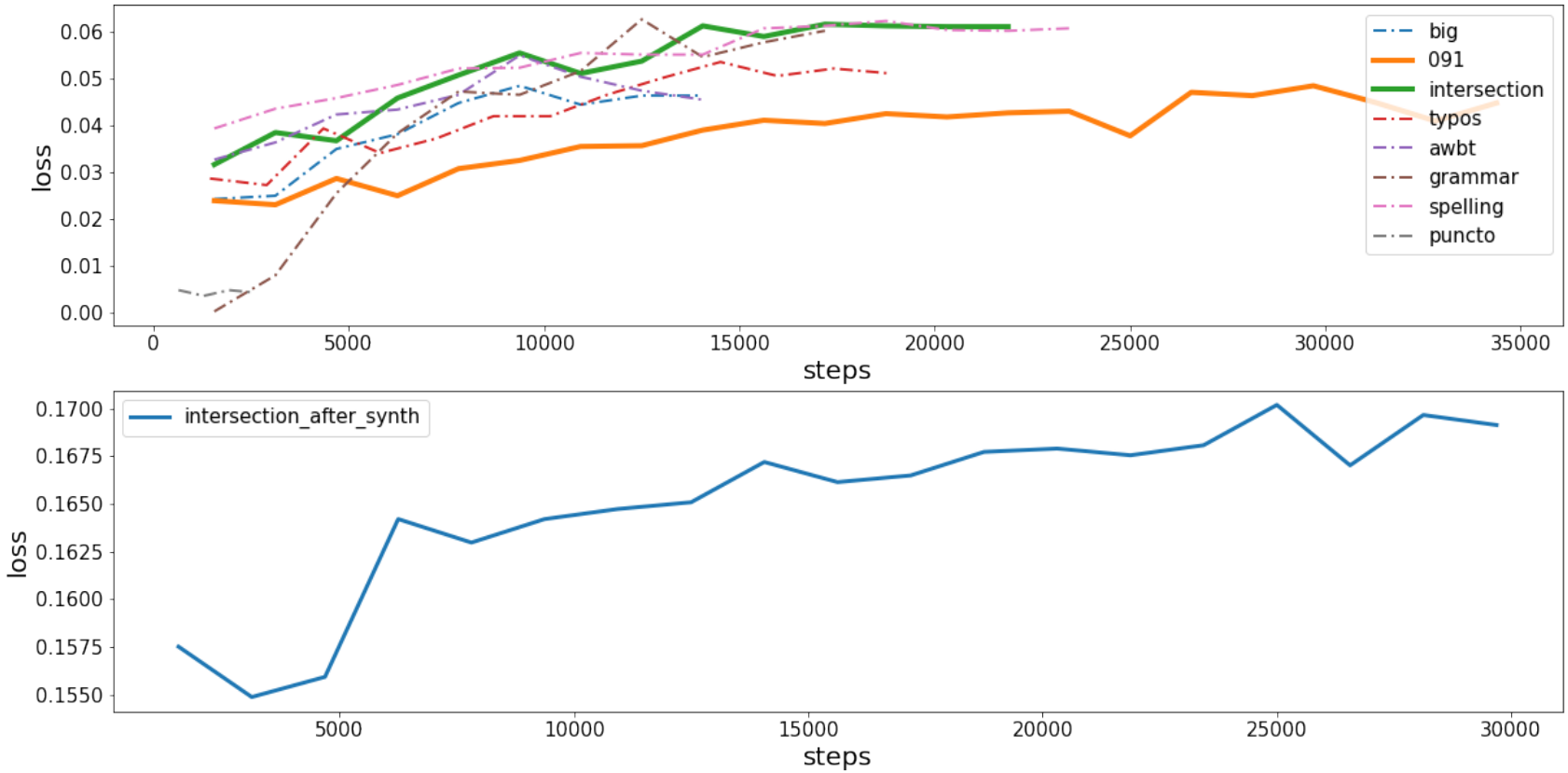
Фильтрация по комментариям

Фильтрация производилась регулярными выражениями.

- Отдельные выражения по упоминанию грамматики, написания и опечаток
- Комбинированный поиск
- Проект AutoWikiBrowser – проект на Википедии, собирающий и валидирующий регулярные выражения для поиска грамматических ошибок. В своих ревизиях оставляют специальный тэг

Валидация 1

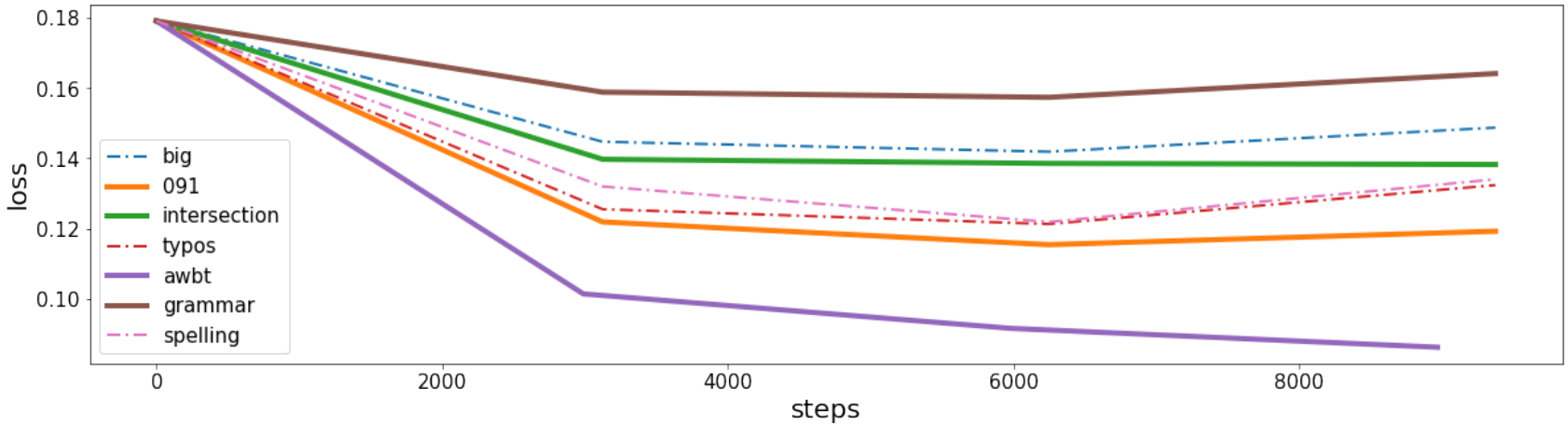
Датасет на 100 000 примеров



- Фильтрация по комментариям работает лучше, чем фильтрация по индексу
- Фильтрация по комментариям улучшает сентетику

Валидация 2

Датасет на 200 000 примеров



Из-за улучшения сентетики все датасеты показали себя хуже, однако можно сделать и положительные выводы:

- Фильтрация по комментариям работает лучше, чем фильтрация по индексу
- Фильтрация по комментариям иногда почти не портит сентетику

Из странное: фильтрация по комментариям проекта AutoWikiBrowser показывает себя очень плохо

Результаты

- Разработана система создания датасетов от списка ссылок на дампы до готовых к запуску датасетов
- Изучены разные подходы фильтрации датасетов
- Обнаружено что фильтрация по комментариям работает эффективнее

Apendix 1

- NUCLE
- Lang-8
- The First Certificate in English
- WI+LOCNESS

Доступно к скачиванию на
https://www.cl.cam.ac.uk/research/nl/bea2019st/