Сбор корпуса параллельных текстов для задачи исправления грамматических ошибок на основе истории пользовательских правок

Ермилов А. Н.

Научный руководитель: к.ф.-м.н. И. Е. Куралёнок

Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург

Исправление грамматических ошибок

Задача: нахождение и исправление пунктуационных, грамматических, синтаксических и других ошибок в текстах

Мотивация:

- Ошибки ведут к недопониманию и сказываются на впечатлении
- Требуются хорошие инструменты для исправления текстов
- Разработка таких инструментов упирается в наличие данных

Обзор работ. Экспертная разметка

- Корпус — набор пар из оригинальных и исправленных предложений
- Исправления производятся экспертами вручную

Название корпуса	# предл.	Источник	Год
FCE	34k	ESL[1]	2011
Lang-8	1M	lang-8.com[2]	2012
NUCLE	57k	ESL	2013
AESW	1.2M	научные статьи	2016
JFLEG	6k	ESL	2017
BEA	43k	writeandimprove.com[2]	2019

[1] экзаменационные работы English as a Second Language

[2] онлайн-платформа для помощи в изучении языка

Обзор работ. Извлечение правок текстов

Корпусы предложений на основе данных Википедии:

- WikEd Error Corpus (2014) и WikiAtomicEdits (2018)
- Много исправлений/дополнений смысла, мало исправлений ошибок

Корпусы предложений на основе данных GitHub:

- GitHub Typo Corpus (2019)
- Много исправлений орфографии, отсутствуют правки структуры

Выводы

Проблемы:

- Экспертная разметка — гарантирует уровень качества датасета, но требует много времени и денег
- Извлечение правок из истории редактирования — быстро и дёшево, но нужны хорошие источники текстов с исправлениями

Возможный источник — научные статьи:

- Использовались для сбора корпуса AESW
- История редактирования — любая онлайн-платформа для редактирования latex-документов, например Papeeria[3]

[3] https://papeeria.com/

Цели и задачи

Цель: сбор корпуса параллельных предложения для задачи исправления грамматических ошибок на основе истории редактирования научных статей

Задачи:

- Обработать данные публичных проектов Papeeria
- Собрать датасет на основе обработанных данных
- Оценить качество полученного датасета

Обработка данных Papeeria

Патч = низкоуровневое описание правки
История редактирования = документ[4] + патчи

Проблемы правок на уровне патчей:

- Создаются каждые 2-3 секунды
- Не хранят контекст

Правка на уровне патчей	Реальная правка
It is known, that... ⇓ It is wid known, that... ⇓ It is widely known that...	It is known, that... ⇓ It is widely known, that...

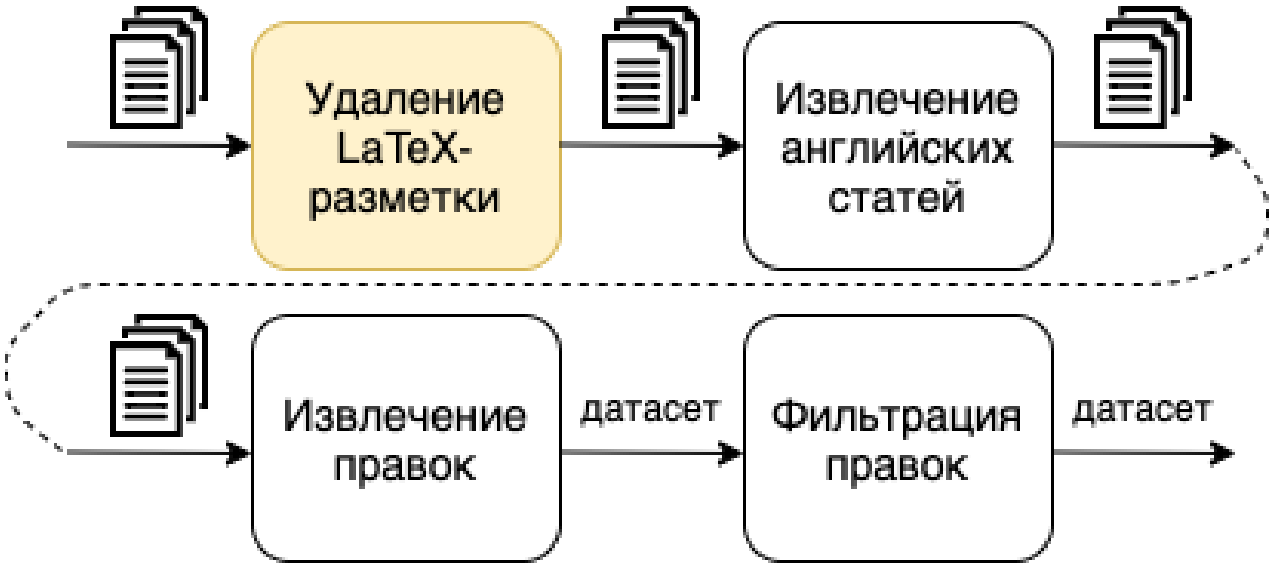
[4] хранится только последняя версия документа

Обработка данных Papeeria

- Объединение патчей в группы по времени создания и расстоянию
- Получение промежуточных версий документов для извлечения исправлений с контекстом

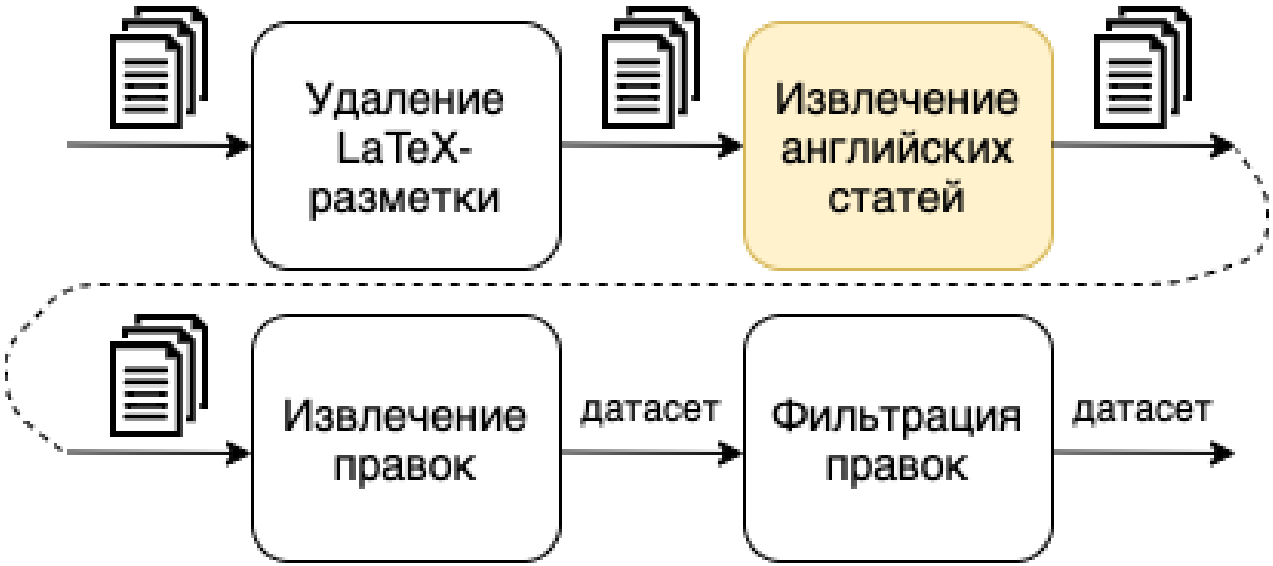
Сбор датасета

- Замена части элементов разметки тегами
  Пример: This process is modeled in FIGURE
- Удаление оставшейся разметки



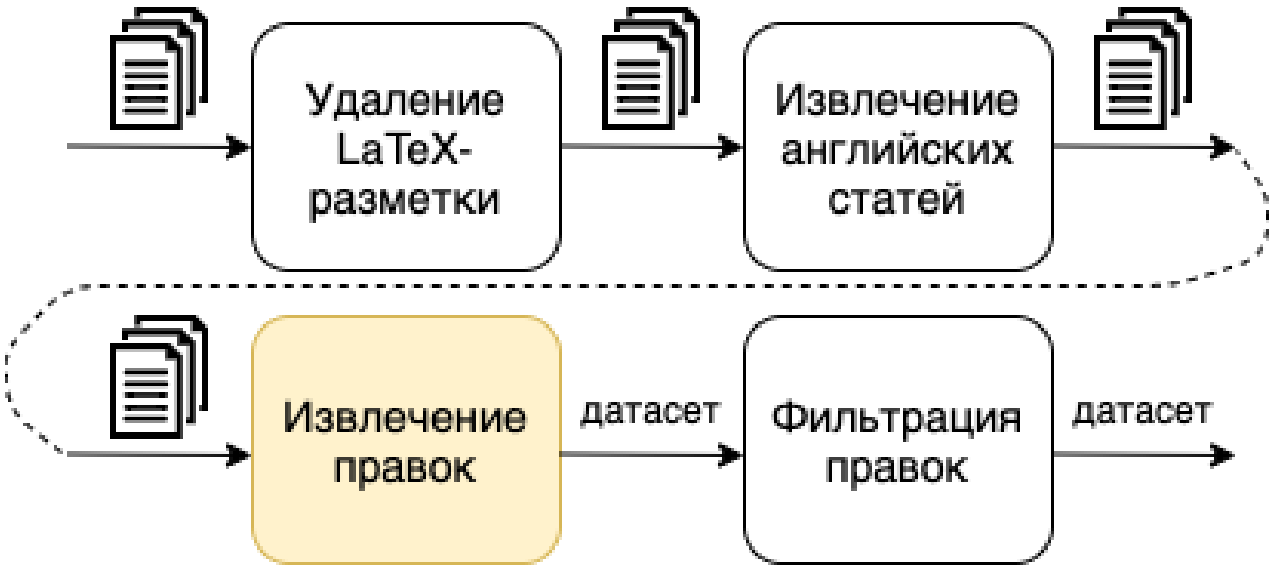
Сбор датасета

- Фильтрация документов по языку с помощью библиотеки `langdetect`
- Извлечение документов со средней длиной предложений, выше порогового значения



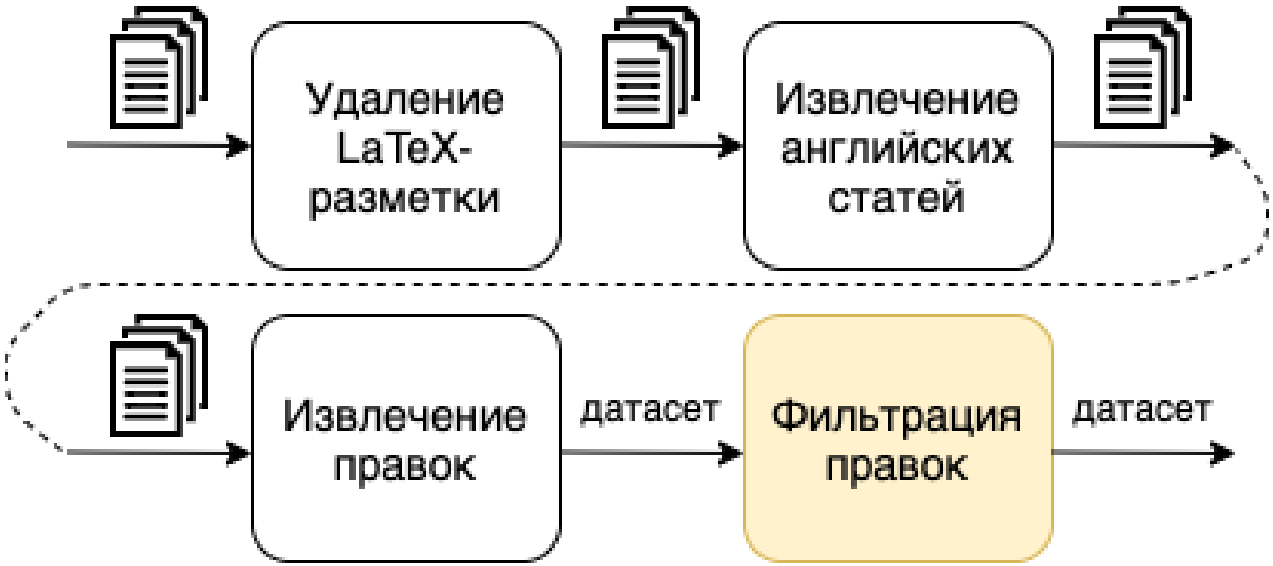
Сбор датасета

- Разбиение соседних версий на предложения
- Поиск похожего предложения в k-окрестности исходного предложения
- Определение похожести с помощью BLEU



Сбор датасета

- Удаление примеров с большим редакционным расстоянием
- Удаление примеров с исправлениями в числах и тегах



Сбор датасета

- Реализован пайплайн с возможностью автоматического сбора корпуса параллельных предложений
- По результатам обработки ≈ 7.200 документов и ≈ 450.000 патчей собран корпус из ≈ 22.000 предложений

Оценка качества

Хочется понять, насколько извлечённые исправления естественны и разнообразны, как сильно они улучшают качество текстов

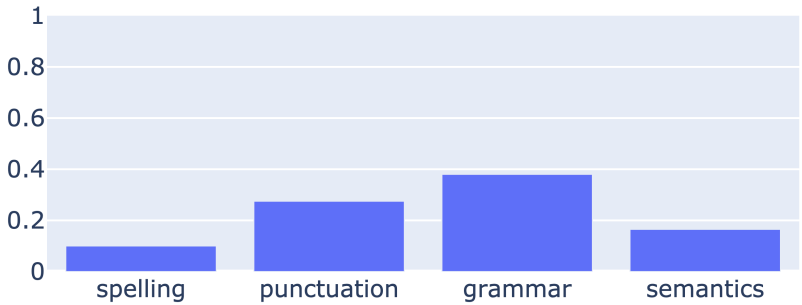
Для этого нужно:

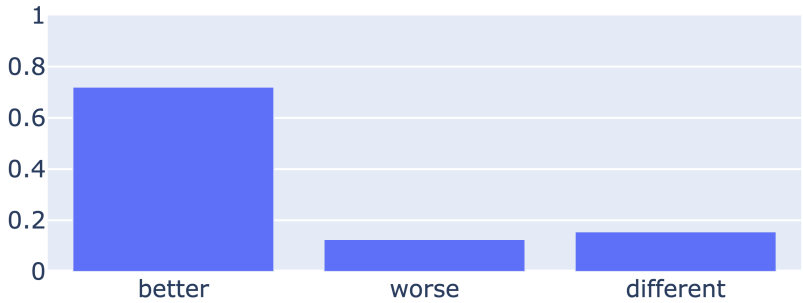
1. оценить улучшение качества текстов с точки зрения носителей языка
2. оценить улучшение качества текстов с точки зрения языковых моделей
3. сравнить поведение языковых моделей на различных корпусах

Оценка качества. Разметка в MTurk

Разметка 200 случайных примеров из собранного корпуса в Amazon Mechanical Turk на основе двух заданий:

1. Какие лингвистические особенности текста (орфография, пунктуация, грамматика, семантика) затрагиваются правкой?
2. Как приведённая правка влияет на качество текста?





Оценка качества. Языковые модели

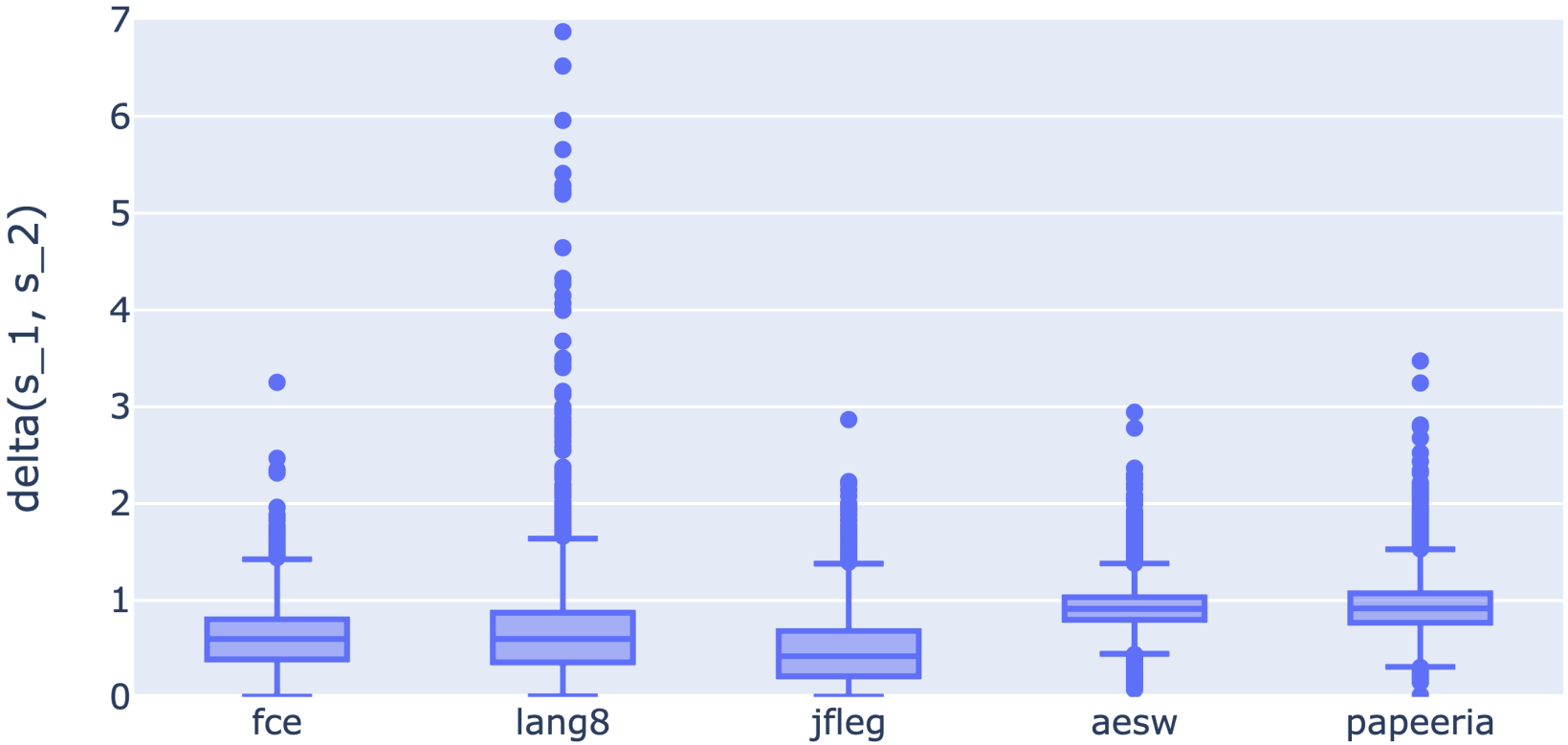
Перплексия — мера того, насколько текст t вероятен с точки зрения языковой модели

$$\mathrm{PP}(t) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log \mathrm{Prob}(t_i)\right)$$

Уменьшение перплексии $\Longleftrightarrow$ повышение качества

$$\delta(s_1, s_2) = \mathrm{PP}(s_2) \,/\, \mathrm{PP}(s_1)$$

Оценка качества. Языковые модели



Изменение перплексии языковой модели GPT-2

FCE, Lang-8, JFLEG — ESL-корпуса
AESW — корпус правок научных статей
Papeeria — собранный корпус

Результаты

- Реализована библиотека[5] для извлечения правок из истории пользовательского редактирования latex-документов
- Собран корпус из ≈ 22.000 пар предложений
- По результатам оценки качества корпуса[6]:
  - большая часть ошибок естественная и затрагивает грамматику и пунктуацию
  - изменение перплексии GPT-2 на текстах схоже с корпусом AESW
- Полученные результаты подтверждают качество собранного датасета и способность описанного метода извлекать разумные правки

[5] https://github.com/AntonYermilov/papeeria-patch-processor
[6] https://github.com/AntonYermilov/gec-dataset-analyzer