NATIONAL RESEARCH UNIVERSITY

Школа Иностранных Языков,
НИУ ВШЭ

Выпускная квалификационная работа на тему:

МЕТОДЫ МАШИННОГО ПЕРЕВОДА ДЛЯ СОЗДАНИЯ КОРПУСА УПРОЩЕННЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Выполнила:
Ижевская Александра Владимировна,
группа 1711-2
Научный руководитель:
Кандидат технических наук
Артемова Екатерина Леонидовна

План

- Введение
- Предметная область
- Применение методов машинного перевода для решения задачи упрощения предложений на русском языке
- Результаты
- Заключение

Введение

Актуальность работы:

- Автоматическое упрощение предложений позволяет представлять информацию в более доступном для понимания виде и имеет особую важность для людей, страдающих от когнитивных расстройств, детей и тех, кто учит иностранный язык

- Однако задача не была достаточно изучена в русском языке

Цель:

Изучение роли машинного перевода в обучении моделей для автоматического упрощения предложений и преодоления проблемы нехватки данных

Введение

Объект:
Методы машинного перевода

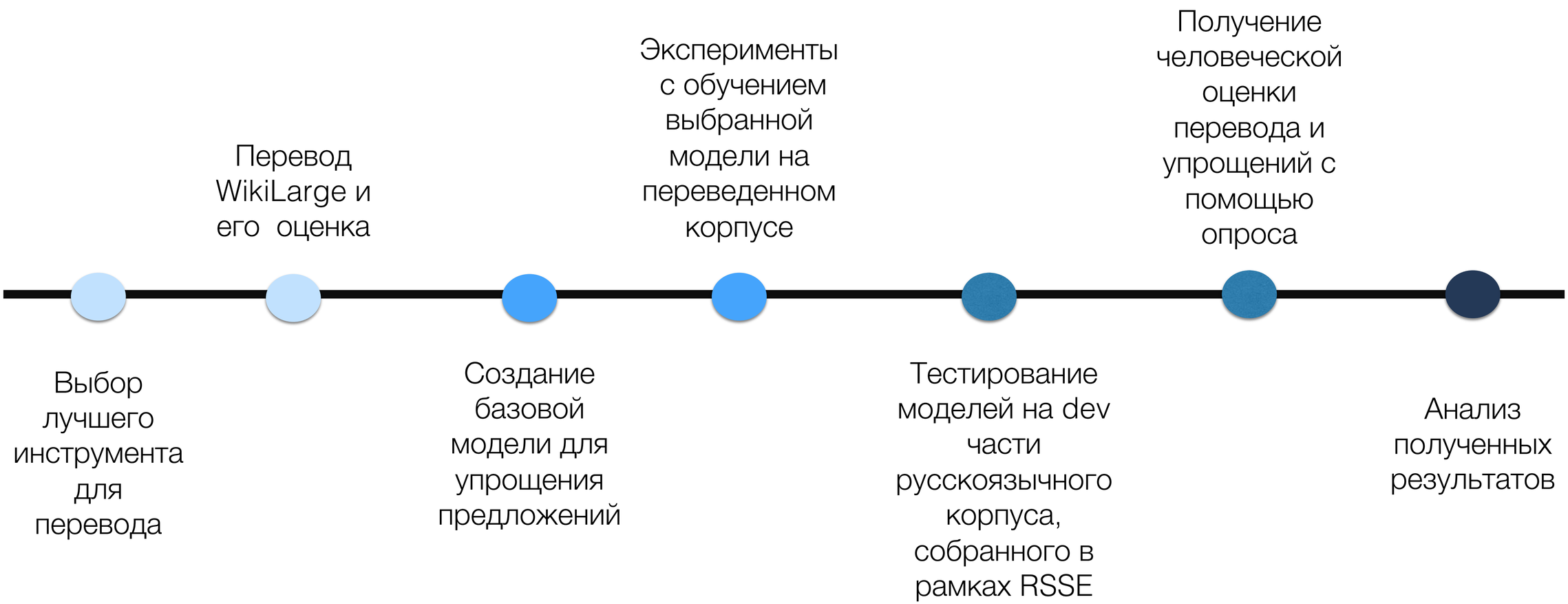
Предмет:
Применение методов машинного перевода для перевода параллельных корпусов простых предложений и обучения моделей для упрощения предложений на русском языке

Методы:
Сравнение, анализ, проведение и оценка экспериментов с помощью метрик и статистик, опрос

Задачи исследования

1. Перевести корпус WikiLarge
2. Оценить качество перевода
3. Использовать переведенные данные для обучения моделей для упрощения предложений
4. Автоматически оценить качество моделей и провести опрос, чтобы получить также и человеческую оценку упрощений
5. Обработать результаты и сделать выводы, основанные на данных.

Этапы работы



Предметная область

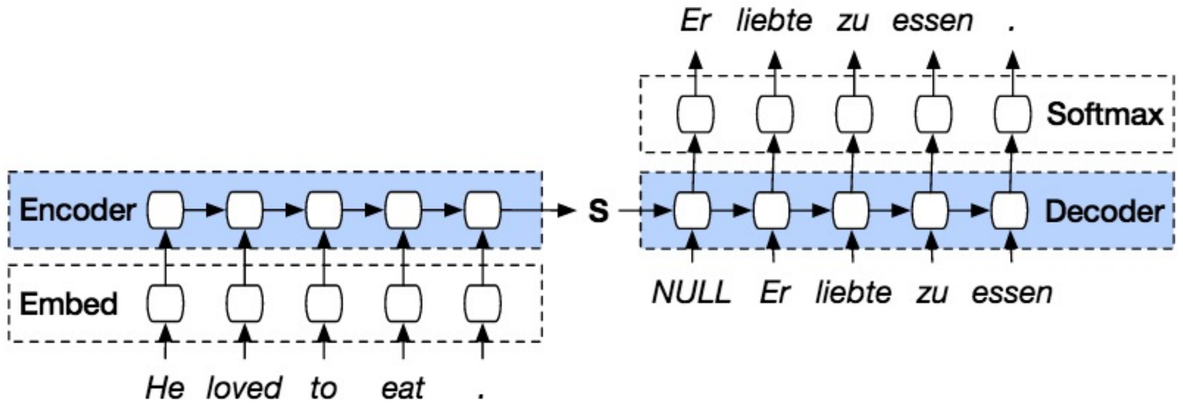
Ключевые термины:

- **Упрощение предложений**

They are culturally akin to the coastal peoples of Papua New Guinea.

Their culture is similar to the culture of the coastal peoples of Papua New Guinea.

- **Sequence-to-Sequence обучение**



- **Автоматические метрики упрощения предложений**

BLEU

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \leq r \end{cases}$$

$$BLEU = BP * \exp(\sum_{n=1}^{N}(w_n log p_n),$$

(часто $w_n = 1/N$)

SARI

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \bar{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(R \cap \bar{I})}$$

$$\#_g(O \cap \bar{I}) = \max(\#_g(O) - \#_g(I), O)$$
$$\#_g(R \cap \bar{I}) = \max(\#_g(R) - \#_g(I), O)$$

FKGL

$$FKLG = 0{,}39 * \left(\frac{total\ words}{total\ sentences}\right) + 11{,}8 * \left(\frac{total\ syllables}{total\ words}\right) - 15{,}59$$

Анализ теоретического материала

Хорошо изученные аспекты:

- Sequence-to-sequence подход к задаче
- Применение моделей машинного перевода
- Создание параллельных корпусов, подходящих для обучения

Для английского языка

Аспекты, которым не было уделено должное внимание:

- Применение моделей машинного перевода для решения задачи во многих других языках
- Создание параллельных корпусов, подходящих для обучения, во многих других языках (включая русский)

Применение методов машинного перевода для решения задачи упрощения предложений на русском языке

WikiLarge:

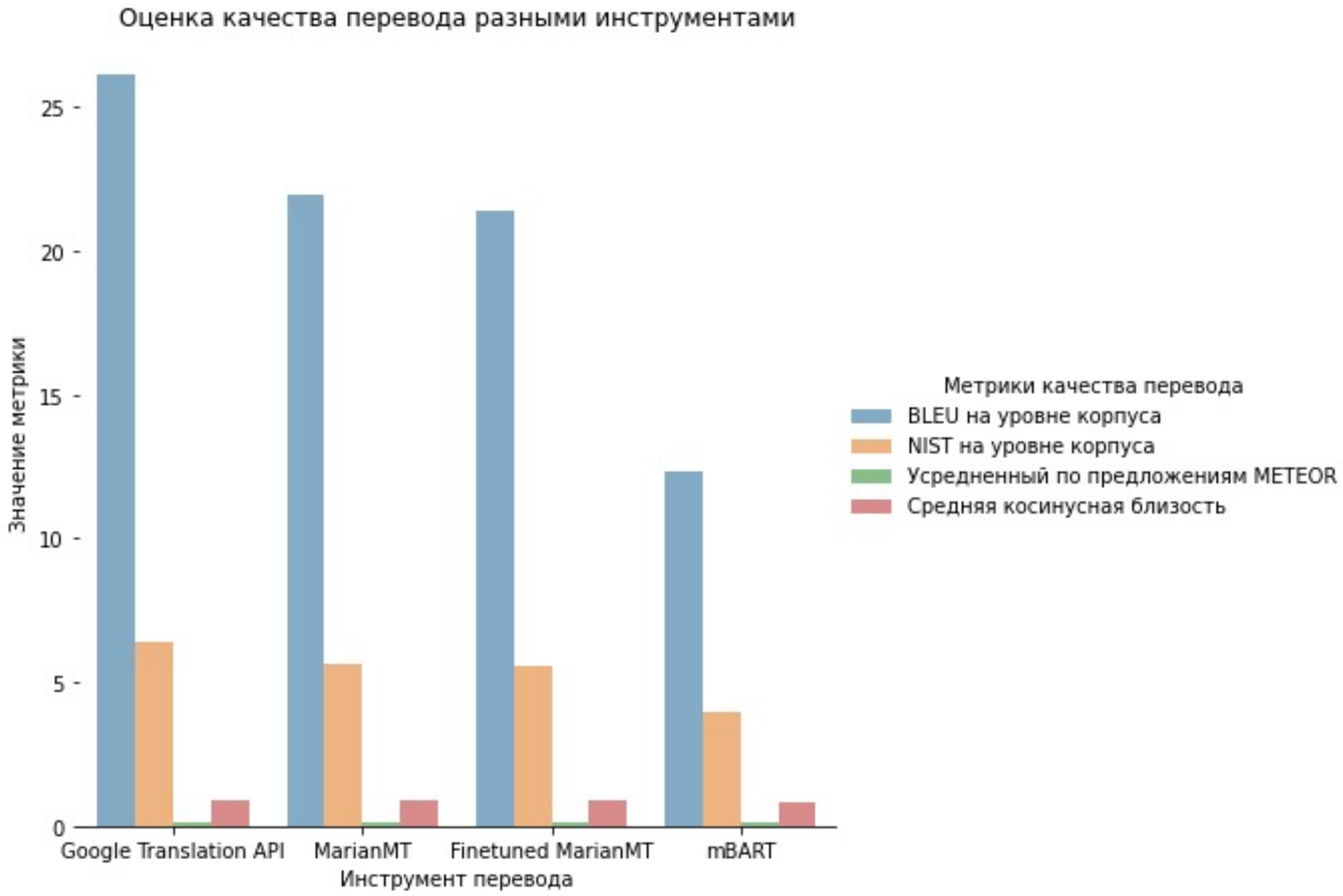
- Англоязычный корпус упрощений
- Содержит данные из Wikipedia и Simple Wikipedia
- Состоит из обучающей (296402 пар предложений), валидационной (2000) и тестовой части (359)

Исходное предложение	Упрощенный вариант
Mount Batur (Gunung Batur) is an active volcano located at the center of two concentric calderas north west of Mount Agung , Bali , Indonesia .	Mount Batur or Gunung Batur is a volcano on Bali .
Negros Oriental Filipino : Silangang Negros also called Oriental Negros , " Eastern Negros " is a province of the Philippines located in the Central Visayas region .	Negros Oriental , sometimes called Oriental Negros East Negros, is a province in the Philippines .

Этап 1. Перевод WikiLarge

Автоматические инструменты для перевода:

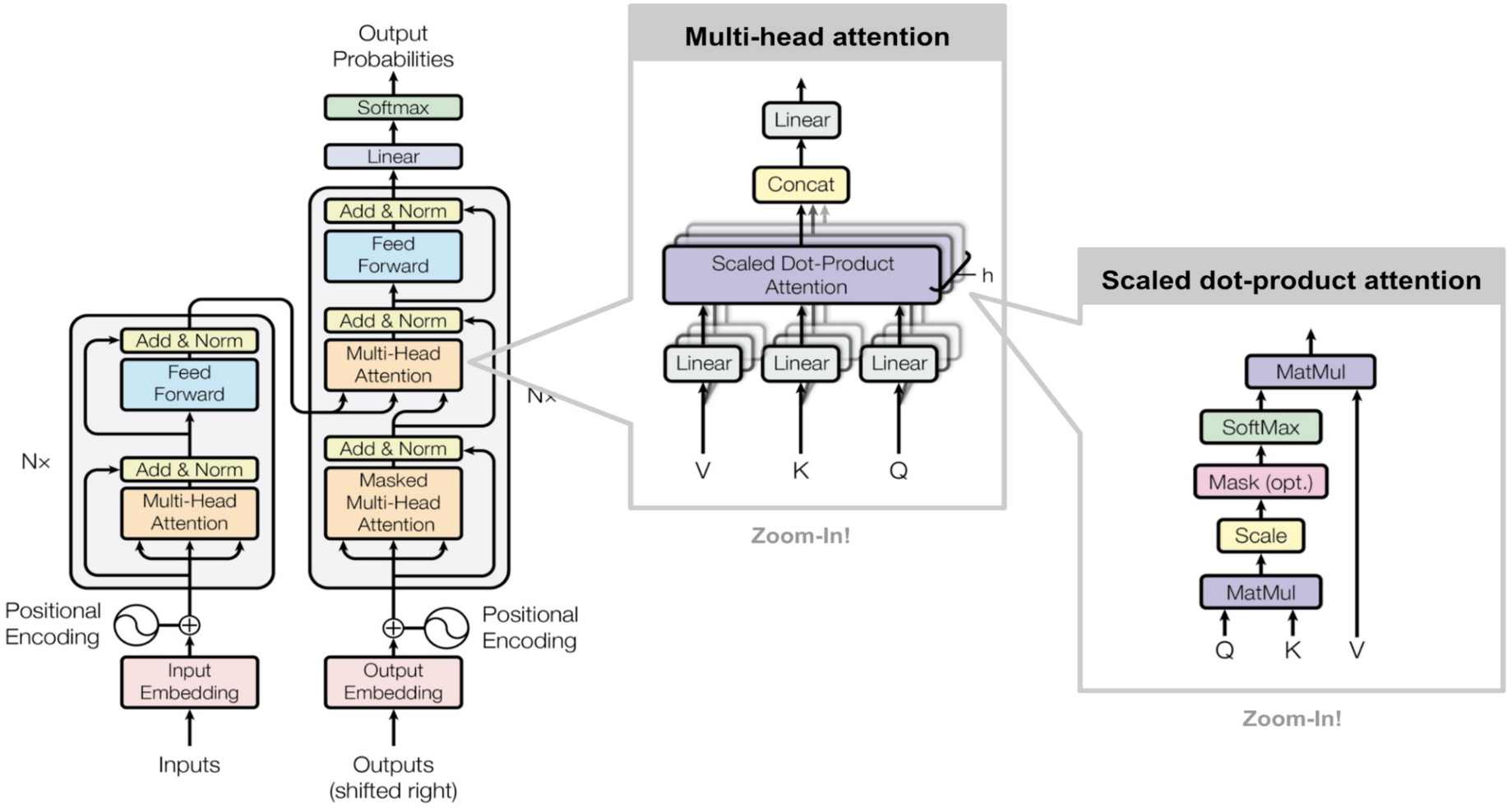
- Google Translation Api
- MarianMT, доступная в библиотеке Hugging Face
- FairSeq mBART



Этап 2. Применение переведенных данных для обучения моделей

mBART

Мультиязычная модель на основе архитектуры Трансформер



Этап 2. Применение переведенных данных для обучения моделей

Примененные улучшения:

- Очистка данных плохого качества и фильтрация предложений на основе длины
- Использование для обучения также корпуса парафраз ParaPhraserPlus
- Механизм «токенов контроля»

Этап 2. Применение переведенных данных для обучения моделей. Результаты обучения

Качество SARI

	Отфильтро-ванный WikiLarge	Объединение отфильтрованного WikiLarge и ParaPhraserPlus	Отфильтро-ванный WikiLarge + ParaPhraserPlus	ParaPhraserPlus	Объединение отфильтрованного WikiLarge и ParaPhraserPlus с токенами контроля
SARI	32.957	34.31	34.738	35.149	**38.823**
Compression ratio	0.635	0.68	0.632	0.624	0.932
Sentence splits	0.987	0.97	0.986	0.985	1.372
Levenshtein similarity	0.757	0.63	0.721	0.711	0.45
Exact copies	0.008	0.007	0.039	0.026	0
Additions proportion	0.019	0.217	0.045	0.053	0.6
Deletions proportion	0.392	0.546	0.418	0.441	0.595
Lexical complexity	10.724	10.732	10.734	10.732	10.489

Качество BLEU и FKGL

	Отфильтро-ванный WikiLarge	Объединение отфильтро-ванного WikiLarge и ParaPhraserPlus	Отфильтрованный WikiLarge + ParaPhraserPlus	ParaPhraserPlus	Объединение отфильтрованного WikiLarge и ParaPhraserPlus с токенами контроля
BLEU	36	**36.64**	33.98	30.31	8.7
FKLG corpus	13.89	13.84	13.65	13.83	**11.08**
ASL	10.69	11.28	10.58	10.04	12.23
AWS	2.89	2.87	2.88	2.9	2.65

Этап 2. Применение переведенных данных для обучения моделей. Примеры

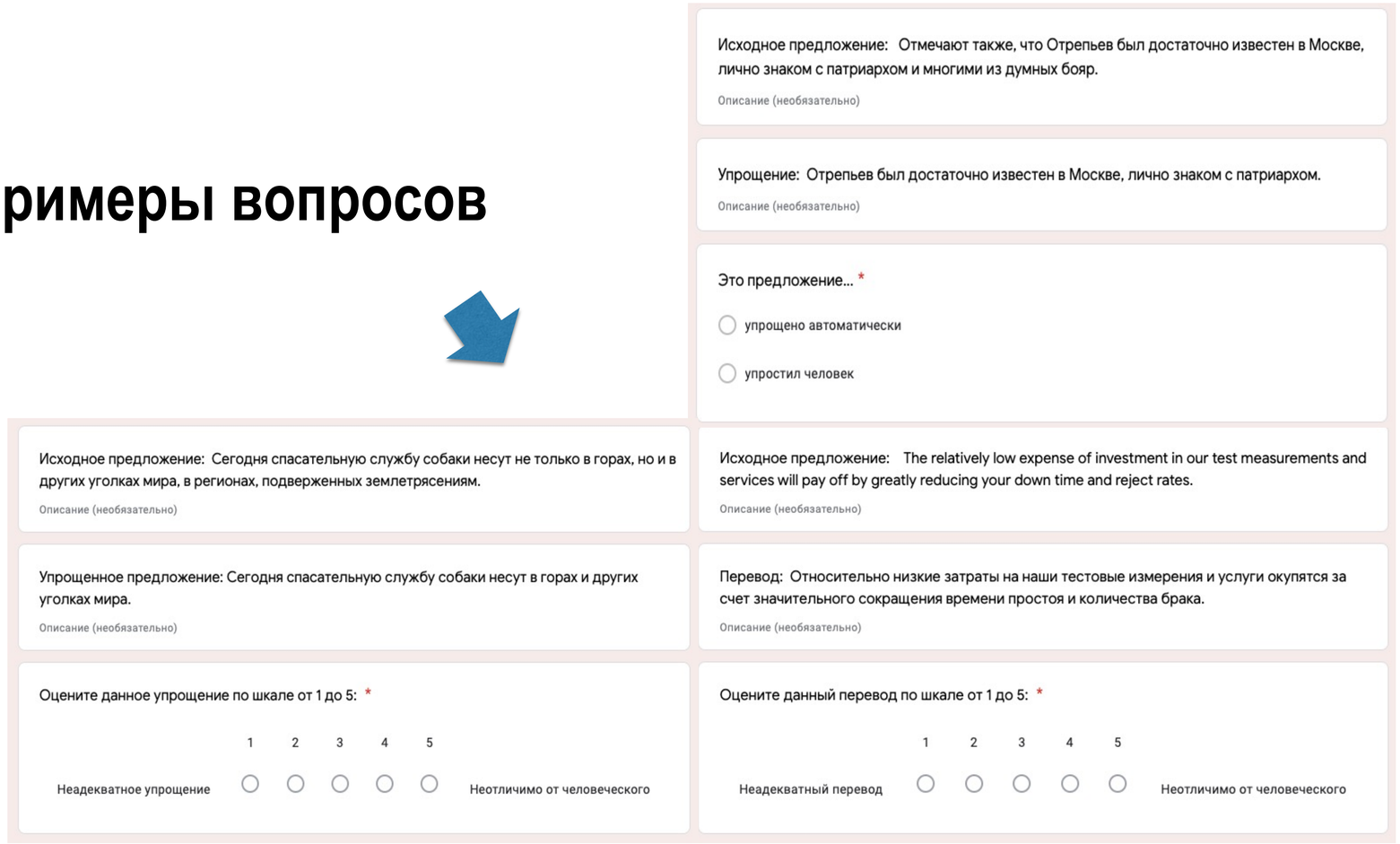
Исходное предложение:	**Упрощение, созданное человеком:**	**Упрощение модели:**
Юг и среднюю часть республики занимают горы и предгорья Большого Кавказа, на севере начинается Прикаспийская низменность.	Часть республики занимают горы, а часть низменность.	Юг и среднюю часть республики занимают горы и предгорья Большого Кавказа.
Чтобы иметь возможность показать картину возможно большей зрительской аудитории, создатели фильма значительно смягчили на экране любовный контекст книги.	В фильме меньше любовных сцен, чем в книге. Это сделано для того, чтобы показать его большому количеству зрителей.	Создатели фильма значительно смягчили на экране любовный контекст книги.
У части больных появляются тошнота и рвота, в отдельных случаях могут быть катаральные явления: першения в горле, сухой кашель, насморк.	У больных наблюдаются: тошнота, рвота, першение в горле, кашель, насморк.	У некоторых больных появляются тошнота и рвота.

Этап 3. Человеческая оценка автоматического перевода и упрощения предложений

Опрос

- В Google Forms
- Включает 3 блока
- Цель: получить человеческую оценку перевода и упрощений

Примеры вопросов



Этап 3. Человеческая оценка автоматического перевода и упрощения предложений

- Всего приняли участие 52 человека

Статистика оценки автоматического перевода

	Оценка автоматического перевода
Минимальное значение	1
Максимальное значение	5
Среднее значение	4.22
Медианное значение	4

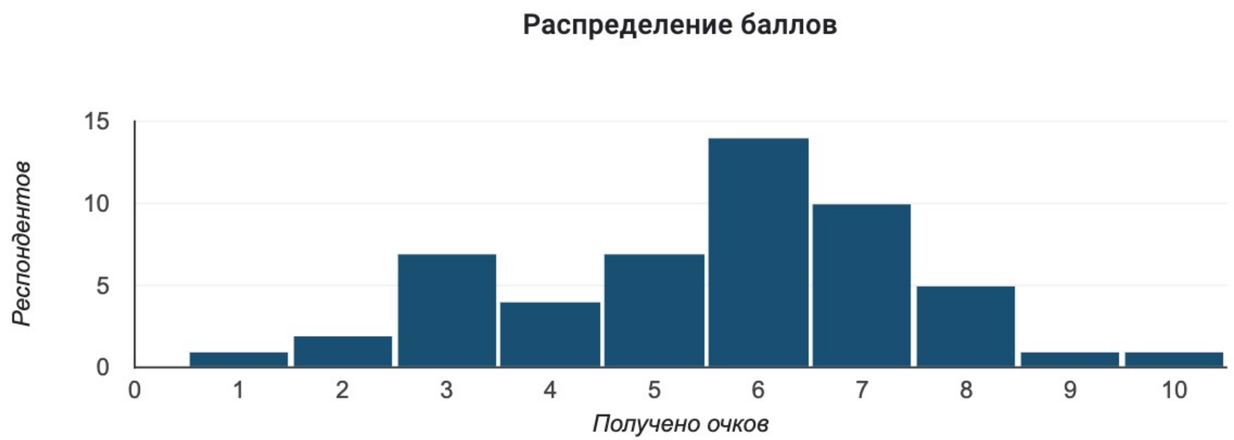
Статистика оценки автоматического упрощения

	Оценка автоматического упрощения
Минимальное значение	1
Максимальное значение	5
Среднее значение	3.5
Медианное значение	4

Этап 3. Человеческая оценка автоматического перевода и упрощения предложений

Распределение правильных ответов

Удовлетворительно	Медиана	Диапазон
Баллов: 5,58 из 10	Баллов: 6 из 10	Баллов: от 1 до 10



- Среднее значение среди людей, отметивших, что причастны к лингвистике или переводу, - 6.07
- Среднее среди остальных - 4.94

Результаты

- Корпус WikiLarge имеет серьезные недостатки, поэтому для улучшения результатов обучения на нем необходима фильтрация и исправление недочетов
- Лучшим инструментом для перевода с английского на русский язык является Google Translation API
- Переведенные данные могут быть использованы для обучения моделей для упрощения предложений на русском языке, которые достигают хорошего качества, особенно при использовании улучшений
- Упрощения модели уступают в креативности человеческим, но в целом она успешно справляется с задачей

Заключение

Выполненные задачи:

1. Параллельный корпус упрощений WikiLarge автоматически переведен на русский язык с помощью лучшего по качеству инструмента и отфильтрован
2. С помощью переведенных данных обучена русскоязычная модель для упрощения предложений, которая достигла высокого качества
3. В процессе обучения предложены и применены дополнительные улучшения
4. Получена как автоматическая, так и человеческая оценка упрощений, которая показала их высокое качество

Заключение

Теоретическая и практическая значимость исследования:

- Готовые модели могут быть применены на практике для решения задачи упрощения
- Переведенный корпус WikiLarge может быть использован для изучения задачи упрощения для русского языка в целом, а также для обучения и оценивания новых моделей в будущем.

Спасибо за внимание!