RuBERT-Score

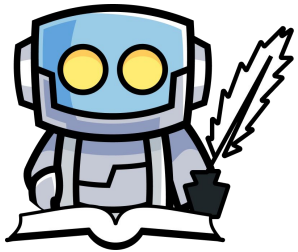
сибирские
нейросети

Спикер: Батурова Дари

О себе

Батурова Дари

- Выпускник Института интеллектуальной робототехники НГУ
- Разработчик-исследователь, “Сибирские нейросети”



О компании:

Решения на базе искусственного интеллекта для бизнеса

- Мультимодальный ИИ
- Большие языковые модели
- Распознавание речи

План доклада

1. Зачем нужна точная оценка качества генерации текстов?
2. Что такое BERTScore?
3. Почему нужен BERT-Score для русского языка?
4. Описание использованных данных
5. Модели для генерации и оценивания
6. Результаты исследования

Проблема

Эталонный текст:
"Кошка сидела на окне, смотрела на улицу и грелась на солнце."

Сгенерированный текст:
"На подоконнике сидела кошка, греясь в лучах солнца и наблюдая за тем, что происходит снаружи."

Как же автоматически понять, насколько хорошо модель сгенерировала текст?

Обзор существующих метрик

Традиционные метрики: BLEU, ROUGE, METEOR, TER и др.

Статистические методы оценки качества основаны на сравнении распределений двух текстов – сгенерированных автоматически и написанных людьми. Примеры метрик: MAUVE, GLTR.

Нейросетевые метрики: BERTScore, BARTScore.

LLM-as-a-Judge: TigerScore, INSTRUCTSCORE, Prometheus, Prometheus-2 и др.

Что такое BERTScore?

BERTScore — это метрика для оценки качества сгенерированных текстов, которая сравнивает векторные представления слов (эмбеддинги).

Примеры использования:

1. Машинный перевод
2. Автоматическое реферирование текстов
3. Перефразирование

\+ **Учет контекста**: анализирует смысл, а не просто совпадение слов.

\+ **Гибкость**: подходит для различных языков и задач.

\- **Высокие вычислительные затраты**: требует больше ресурсов для подсчета по сравнению с традиционными метриками (но при этом сильно меньше, чем LLM-as-a-Judge).

\- **Чувствительность к выбранной модели**: результаты зависят от качества используемой языковой модели.

Оригинальная статья

BERTScore: Evaluating Text Generation with BERT (2019) - Zhang et al.

BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

Tianyi Zhang[*†‡◇], **Varsha Kishore**[*‡], **Felix Wu**[*‡], **Kilian Q. Weinberger**[†‡◇], and **Yoav Artzi**[‡§]

[‡]Department of Computer Science and [§]Cornell Tech, Cornell University
{vk352, fw245, kilian}@cornell.edu {yoav}@cs.cornell.edu

[◇]ASAPP Inc.
tzhang@asapp.com

ABSTRACT

We propose BERTSCORE, an automatic evaluation metric for text generation. Analogously to common metrics, BERTSCORE computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches, we compute token similarity using contextual embeddings. We evaluate using the outputs of 363 machine translation and image captioning systems. BERTSCORE correlates better with human judgments and provides stronger model selection performance than existing metrics. Finally, we use an adversarial paraphrase detection task to show that BERTSCORE is more robust to challenging examples when compared to existing metrics.

Почему для русского нужен свой BERTScore?

➜ Исследование BERTScore в оригинальной статье было посвящено английскому, китайскому и турецкому языкам.

➜ Для русского языка такое исследование не проводилось .

➜ Русский язык имеет свои грамматические и синтаксические особенности.

Поэтому необходимо провести отдельное исследование, чтобы выявить, какие модели лучше всего подходят для оценки текстов на русском языке.

Примеры BERTScore

Задача: упрощение текста

Исходный текст: Александр обучался также военным наукам; английскому, французскому и немецкому языкам, рисованию; фехтованию и другим дисциплинам.

Эталонный текст: Александр учился воевать, говорить на иностранных языках, рисовать, фехтовать и делать многое другое.

Сгенерированный текст: Александр изучал военное дело, а также английский, французский и немецкий языки, рисование, фехтование и другие предметы.

Оценки BERTScore:

XLM-RoBERTa-base: 0.94

mBART-large-cc25: 0.62

RuBERT-Score: цель

Найти языковые модели для русского языка, векторные представления которых лучше всего коррелируют с человеческой оценкой

RuBERT-Score: задачи

- **Подготовка датасетов**, которые различаются между собой по типам задач, доменам и жанрам;
- **Генерация ответов** моделей для данных задач;
- **Получение экспертной оценки** для сгенерированных ответов;
- **Расчет корреляций** между значениями BERT-Score (с векторными представлениями от разных моделей) и экспертной оценкой;
- **Выработка рекомендаций** по использованию моделей для русского языка.

Набор датасетов

Датасет	Описание	Средняя длина вход. текста	Средняя длина выход. текста	Кол-во уникальных слов во вход. текстах	Кол-во уникальных слов в выход. текстах
dialogsum-ru	суммаризация диалогов	757	117	9 907	5 400
reviews-russian	суммаризация отзывов пользователей на отели и гостиницы	1 390	448	5 096	2 115
ru-simple-sent-eval	упрощение текстов	137	91	15 300	**20 694**
science-summarization-dataset	суммаризация научных статей	**20 155**	**843**	**112 380**	13 469
telegram-financial-sentiment-summarization	тексты постов из Телеграмма и их краткие содержания	313	169	26 818	20 426
yandex-jobs	тексты с описанием вакансий Яндекса	925	38	8 364	504

Модели для генерации ответов

Мы использовали две модели для русского языка:

Gigachat Lite

YandexGPT Lite

Для каждой из задач были вручную написаны промпты для моделей. Для каждой модели был использован один и тот же промпт, без изменений.

Примеры промптов

Датасет	Подводка для модели
dialogsum-ru	Сгенерируй краткий пересказ этого диалога в 1-2 предложениях:
reviews-russian	Сгенерируй краткий пересказ этого отзыва. Выделяй только важные факты. Отзыв:
ru-simple-sent-eval	Перепиши предложение так, чтобы его стало проще понимать:
science-sum	Ниже приведена научная статья. Выдели главные факты и напиши краткое содержание этой статьи.
telegram-fin	Напиши краткий пересказ этой новости. Выделяй только важные факты. Текст новости:
yandex-jobs	Напиши название должности для этой вакансии:

Получение экспертных оценок

1. Мы применили **подход LLM-as-a-judge**, чтобы получить оценки с помощью языковой модели без привлечения экспертов.
2. Проблема: модели склонны переоценивать свои ответы. Мы использовали **“перекрёстную” проверку**: YandexGPT оценивала ответы GigaChat, и наоборот.
3. Для каждой модели были использованы **три различных промпта** для оценки одного текста.
4. Промпты включали описание задачи, эталонный и сгенерированный ответы, и инструкцию для **оценки по 5-балльной шкале** (1 — полностью не соответствует, 5 — полностью соответствует).

Пример промпта для оценивания

Системный промпт: "Ты система, которая оценивает качество сгенерированного текста."

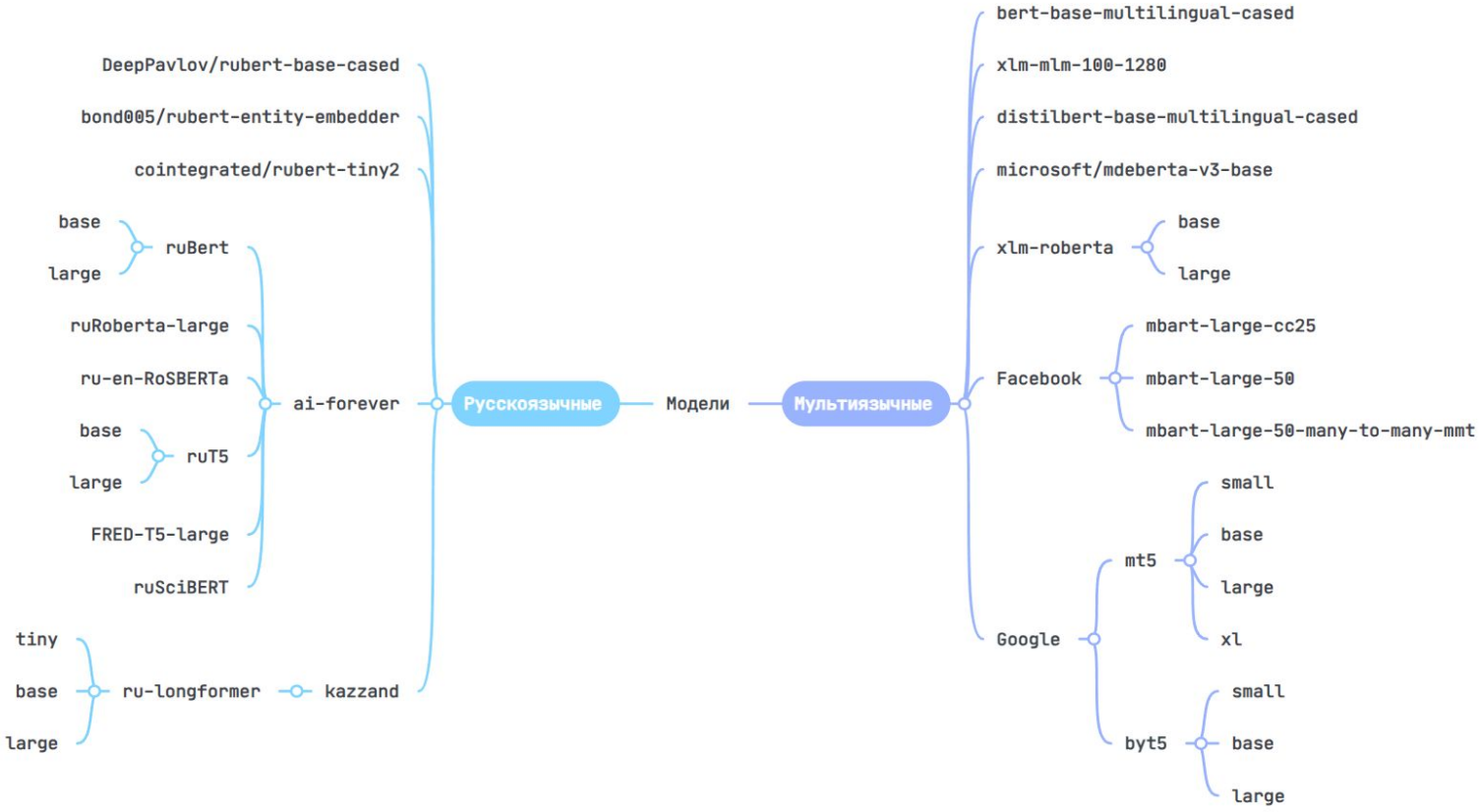
Задачи генерации: 'dialogsum_ru': 'Генерация краткого пересказа диалога'

'reviews_russian': 'Генерация краткого пересказа отзыва об отеле'

...

Промпт: “Оцени, насколько сгенерированный текст совпадает с эталонным, по шкале от 1 до 5.
Укажите одно число, где 1 - это полное несоответствие, а 5 - полное совпадение.
Задача генерации: {task}
Эталонный текст: {ref_text}
Сгенерированный текст: {pred_text}”

Корреляции


Модели
Русскоязычные
DeepPavlov/rubert-base-cased
bond005/rubert-entity-embedder
cointegrated/rubert-tiny2
ai-forever
ruBert
base
large
ruRoberta-large
ru-en-RoSBERTa
ruT5
base
large
FRED-T5-large
ruSciBERT
kazzand
ru-longformer
tiny
base
large
Мультиязычные
bert-base-multilingual-cased
xlm-mlm-100-1280
distilbert-base-multilingual-cased
microsoft/mdeberta-v3-base
xlm-roberta
base
large
Facebook
mbart-large-cc25
mbart-large-50
mbart-large-50-many-to-many-mmt
Google
mt5
small
base
large
xl
byt5
small
base
large

Корреляции метрик для текстов, сгенерированных Gigachat

Dataset	Best embedding	BERT Score	Best embedding (ru)	BERT Score (ru)	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
tg	microsoft/mdeberta-v3-base (7)	**0.827**	ai-forever/ruBert-base (22)	0.825	0.423	0.645	0.497	0.607
ru_simple_sent_eval	google/byt5-large (31)	0.615	ai-forever/ru-en-RoSBERTa (20)	**0.673**	0.096	0.281	0.149	0.25
science	facebook/mbart-large-50 (10)	**0.749**	ai-forever/ru-en-RoSBERTa (20)	**0.749**	0.282	0.599	0.481	0.560
dialogsum_ru	facebook/mbart-large-cc25 (11)	**0.447**	ai-forever/ru-en-RoSBERTa (7)	0.415	0.158	0.236	0.114	0.247
reviews_russian	facebook/mbart-large-50-many-to-many-mmt (6)	**0.678**	ai-forever/ru-en-RoSBERTa (20)	0.655	0.178	0.346	0.206	0.346
yandex	google/byt5-base (6)	0.433	ai-forever/ru-en-RoSBERTa (23)	**0.454**	0.078	0.192	0.165	0.192
AVG (слой)	google/byt5-large (29)	0.594	ai-forever/ru-en-RoSBERTa (20)	**0.630**	0.191	0.379	0.257	0.359
AVG (модель)	google/byt5-large	0.561	ai-forever/ruBert-base	**0.564**	0.191	0.379	0.257	0.359

*После названия модели в скобках указан номер слоя, векторные представления которого показали самую высокую корреляцию

Корреляции метрик для текстов, сгенерированных YandexGPT

Dataset	Best embedding	BERT Score	Best embedding (ru)	BERT Score (ru)	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
tg	microsoft/mdeberta-v3-base (8)	0.299	ai-forever/ruBert-base (7)	**0.345**	0.157	0.238	0.197	0.239
ru_simple_sent_eval	xlm-roberta-large (16)	**0.224**	ai-forever/ru-en-RoSBERTa (21)	0.170	0.068	0.079	0.051	0.067
science	distilbert-base-multilingual-cased (5)	**0.217**	ai-forever/ru-en-RoSBERTa (20)	0.199	0.017	0.116	0.134	0.122
dialogsum_ru	google/mt5-xl (23)	0.291	ai-forever/ru-en-RoSBERTa (22)	**0.333**	0.077	0.149	0.157	0.155
reviews_russian	google/mt5-xl (23)	**0.418**	ai-forever/ruSciBERT (10)	0.397	0.138	0.234	0.176	0.213
yandex	google/byt5-base (10)	**0.509**	ai-forever/ruBert-base (0)	0.488	0.169	0.303	0.214	0.301
AVG (слой)	google/byt5-base (10)	**0.284**	ai-forever/ru-en-RoSBERTa (20)	0.275	0.096	0.171	0.138	0.166
AVG (модель)	facebook/mbart-large-50-many-to-many-mmt	0.249	ai-forever/ruBert-base	**0.260**	0.096	0.171	0.139	0.166

*После названия модели в скобках указан номер слоя, векторные представления которого показали самую высокую корреляцию

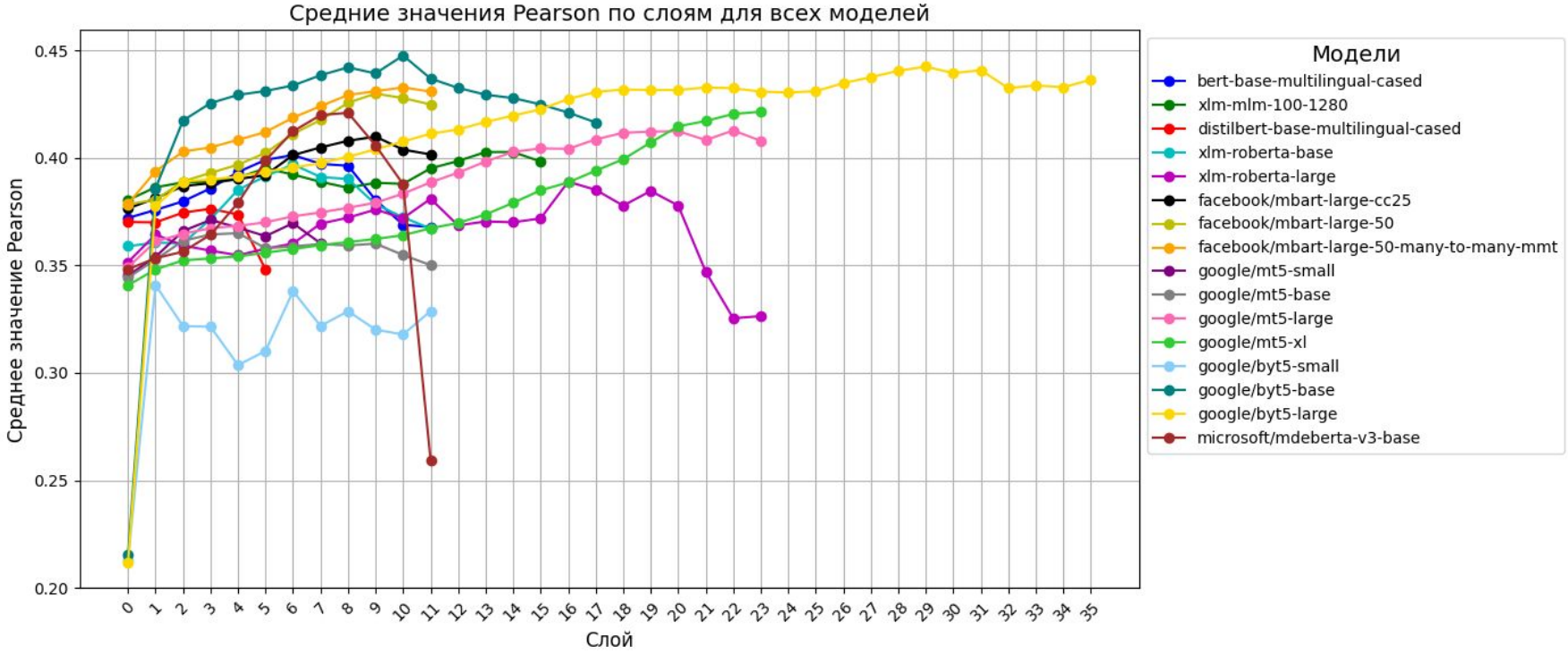
Корреляции для мультиязычных моделей

Модель	Кол-во параметров	Слой	Pearson
google/byt5-base	528M	10	0.447
google/byt5-large	1.23B	29	0.442
facebook/mbart-large-50-many-to-many-mmt	611M	10	0.433
facebook/mbart-large-50	611M	10	0.430
google/mt5-xl	3.7M	23	0.421
microsoft/mdeberta-v3-base	280M	8	0.421
google/mt5-large	1.2B	22	0.413
facebook/mbart-large-cc25	610M	9	0.410
xlm-mlm-100-1280	570M	15	0.403
bert-base-multilingual-cased	179M	6	0.401
xlm-roberta-base	279M	6	0.397
xlm-roberta-large	561M	16	0.389
distilbert-base-multilingual-cased	135M	3	0.376
google/mt5-small	300M	3	0.371
google/mt5-base	580M	4	0.365
google/byt5-small	300M	1	0.341

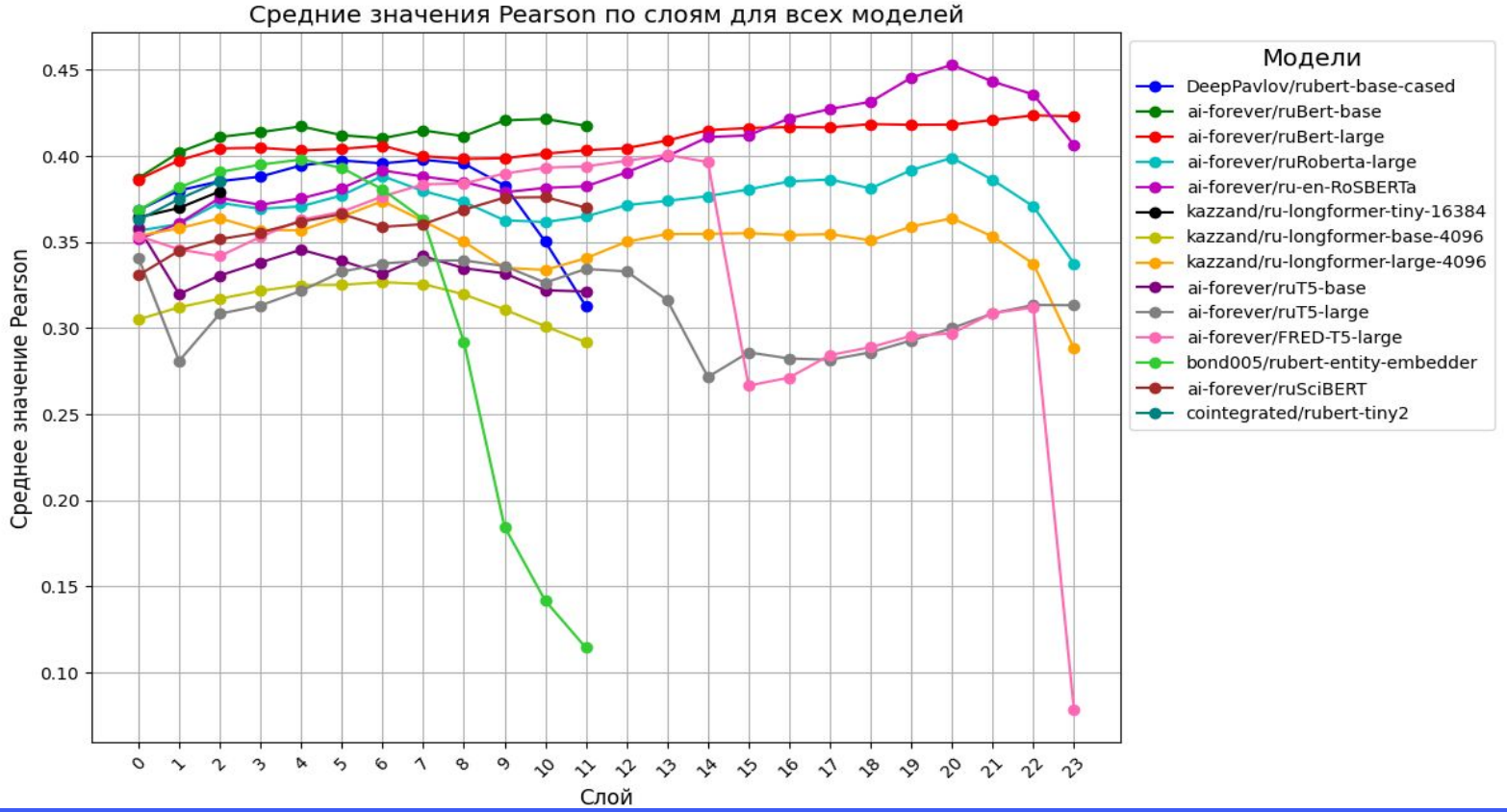
Корреляции для русскоязычных моделей

Модель	Кол-во параметров	Слой	Pearson
ai-forever/ru-en-RoSBERTa	404M	20	0.453
ai-forever/ruBert-large	427M	22	0.424
ai-forever/ruBert-base	178M	10	0.421
ai-forever/FRED-T5-large	820M	13	0.401
ai-forever/ruRoberta-large	355M	20	0.399
bond005/rubert-entity-embedder	180M	4	0.398
DeepPavlov/rubert-base-cased	180M	7	0.398
cointegrated/rubert-tiny2	29M	2	0.385
kazzand/ru-longformer-tiny-16384	34.5M	2	0.379
ai-forever/ruSciBERT	123M	10	0.376
kazzand/ru-longformer-large-4096	434M	6	0.373
ai-forever/ruT5-base	222M	0	0.358
ai-forever/ruT5-large	737M	0	0.341
kazzand/ru-longformer-base-4096	148M	6	0.327

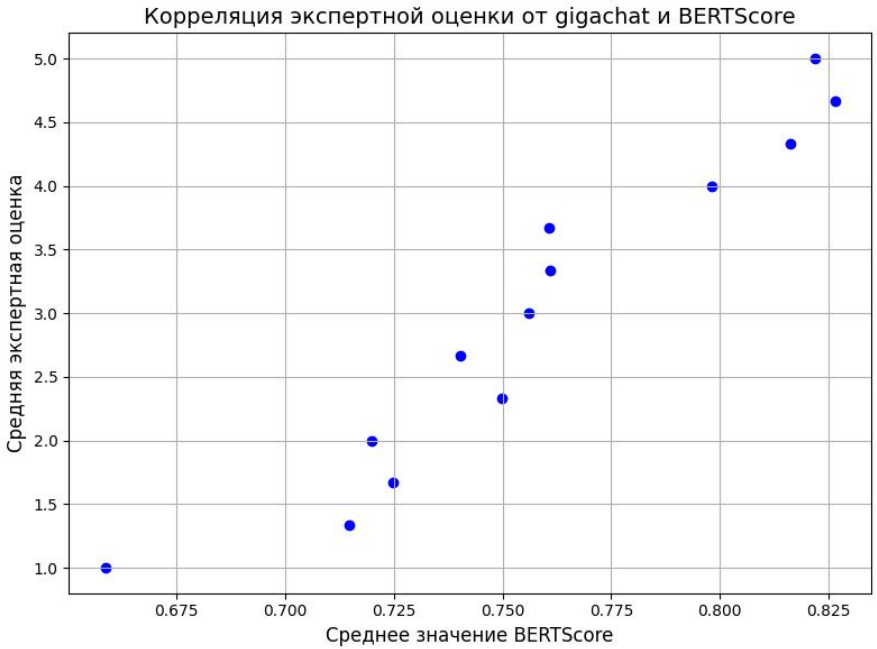
Визуализация корреляций по слоям мультиязычных моделей

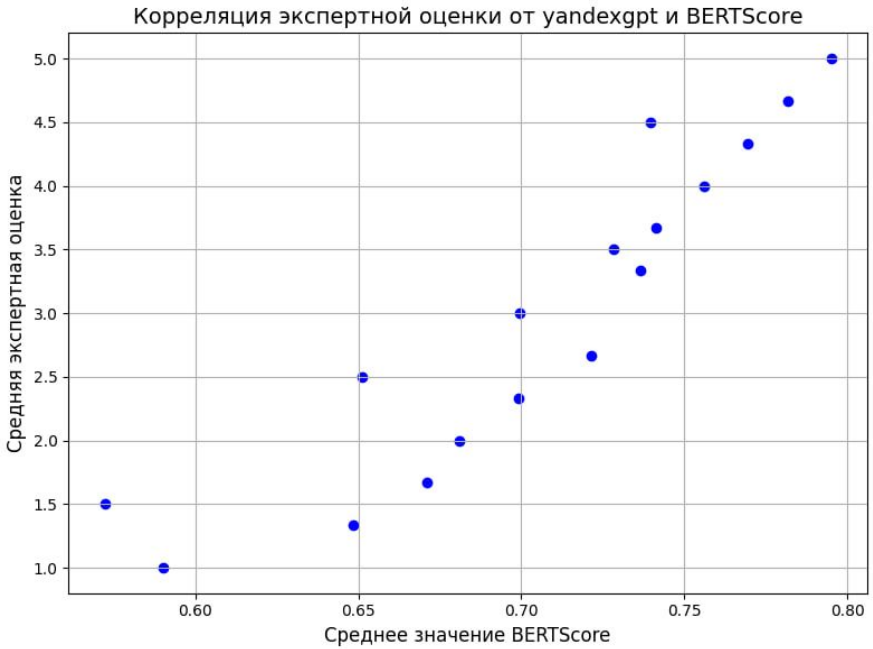


Визуализация корреляций по слоям русскоязычных моделей



Визуализация зависимости экспертной оценки и BERTScore





Выводы

Топ лучших конфигураций: 1. 20 слой модели “ai-forever/ru-en-RoSBERTa”

2. 10 слой модели “google/byt5-base”

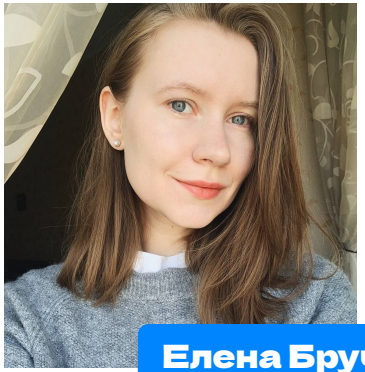
3. 29 слой модели “google/byt5-large”

- Правильный выбор модели и ее слоя критически важен для достижения высокой корреляции с экспертными оценками.
- Подход использования LLM для автоматической оценки текста показал свою эффективность, но он требует больше ресурсов по сравнению с BERTScore.
- Оценка длинных текстов остается проблемой, так как у каждой модели есть свое ограничение на длину входных данных.

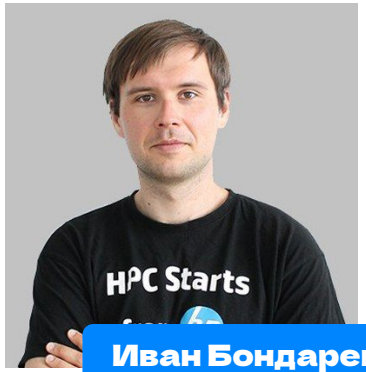
Планы на будущее

- Расширить датасеты для увеличения разнообразия текстов;
- Дополнить выборку моделей;
- Подобрать модели для специфических доменов;
- Провести сравнение с человеческими оценками для повышения точности.

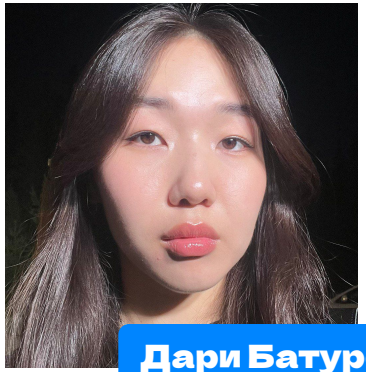
Наша команда



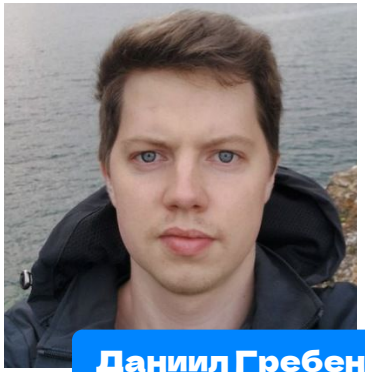
Елена Бручес



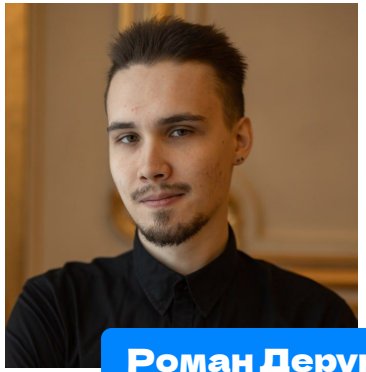
Иван Бондаренко



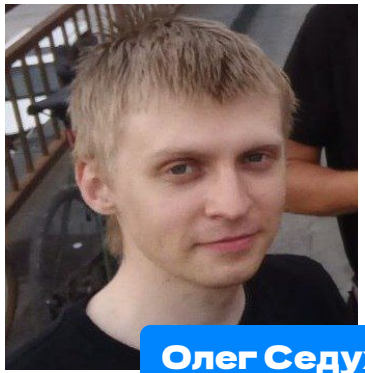
Дари Батурова



Даниил Гребенкин



Роман Дерунец



Олег Седухин

Наши контакты

ООО «СИБИРСКИЕ НЕЙРОСЕТИ»

+7 (913) 477-19-57

info@sibnn.ai

sibnn.ai

Спикер:
Батурова Дари

@dialoger_tech

SibNN/ru_bert_score

@dori_b