LOMONOSOV
all_to_summarize.bot

КОДИИМ

Команда: **“ЛОМОНОСОВ”**
(№ 4) Никита Кукузей, Максим Антонов, София Куршина

Проект: **"Обученная модель детекции парафразов для суммаризации русскоязычного текста"**

Концепция: Создание GAN модели для суммаризации русскоязычного текста на основе fine-tuned версий предобученных моделей суммаризации и парафразирования.

LOMONOSOV
all_to_summarize.bot

КОДИИМ

Актуальность проекта

Проблема: Информационный взрыв и рост вычислительной сложности обработки данных влекут:

1. интеллектуальную слепоту вследствие предвзятости и когнитивных искажений - психологических, социальных, вычислительных и др.
2. снижение рациональности и этичности в решениях, влияющих на жизнь личности, компаний, народов и человечества.

Потребность: Всем необходимо сокращение времени и повышение эффективности обработки любых обращений и входных данных.

Решение: автоматизация (1) обработки и (2)суммаризации входных данных, (3) распознавания потребностей пользователей, (4) быстрого поиска в базах знаний и (5) подсказывания верного ответа/решения.

LOMONOSOV
all_to_summarize.bot

КОДИИМ

Решение (1 этап)

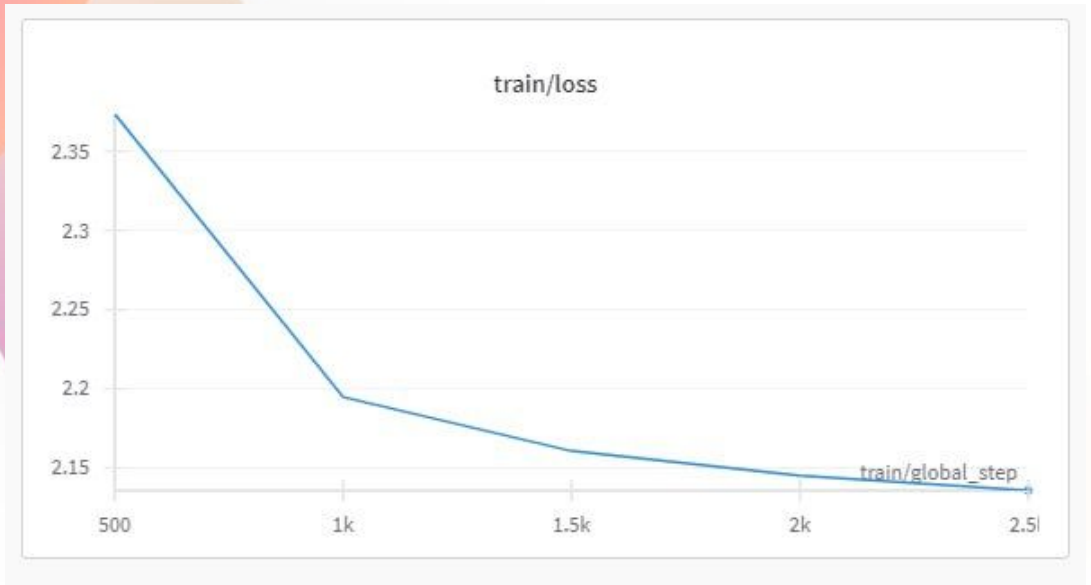
1. open source модель суммаризации русскоязычного текста, высокого качества, легко встраиваемая в API;
2. библиотека - transformers + pytorch
3. fine-tuning предобученной модели mt5 small от google на датасете "Газета";
4. Метрики качества:

- eval_loss:	**1.5822**	- eval_runtime:	**46.8578**
- eval_rouge1:	**12.4365**	- eval_samples_per_second:	**10.18**
- eval_rouge2:	**2.3859**	- eval_steps_per_second:	**0.64**
- eval_rougeL:	**12.4889**	- epoch:	**1.18**
- eval_rougeLsum:	**12.4796**	- eval_gen_len:	**15.7862**

LOMONOSOV
all_to_summarize.bot

КОДИИМ

Этап 1:
1. Обработка датасета газеты
2. Парсинг данных для загрузки в модель
3. Токенизация
4. Векторизация
5. Берем несколько предобученных моделей
   и дообучение на датасете
6. Телеграмм бот



Этап 2:
1. Берем датасет по парафразам
2. Обработка датасета
3. Токенизация и векторизация
4. В качестве метрики используем модель
5. Дообучаем модель по суммаризации

Этап 3:
1. experience & стресс-тестирование,
2. презентация и демо-просмотры.

ТГ-бот

демовидео

LOMONOSOV
all_to_summarize.bot

КОДИИМ

Заключение

При жестких временных и ресурсных ограничениях:

- реализована модель
- поддерживает русский язык;
- модель обучена на достаточно большом объеме данных;
- представлен ТГ-бот для демонстрации возможностей модели,

Планы и перспективы:

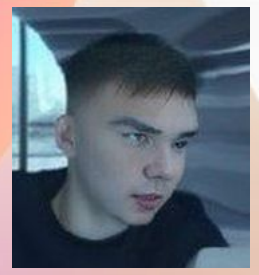
- мультимодальность ввода, обработки и вывода данных (текст, голос, графика, файлы и т.д.),
- мультиплатформенность интерфейса (бот, сайт и т.д.)
- полный цикл информационного “сжатия” смысла входных данных;
- интеграция с сервисами поиска знаний и предсказания потребностей пользователей
- реализация GAN-модели

LOMONOSOV
all_to_summarize.bot

КОДИИМ

“Ломоносов”
Команда
№ **4**


Никита Кукузей


Максим Антонов


София Куршина

All You Need Is Attention

Ист.:
*1) Attention Is All You Need, Ashish Vaswani et al., 2017, https://arxiv.org/abs/1706.03762
*2) Generative adversarial networks, https://habr.com/ru/articles/352794/
*3) https://github.com/google-research/multilingual-t5