Разработка универсального семантического энкодера текстов для русского языка

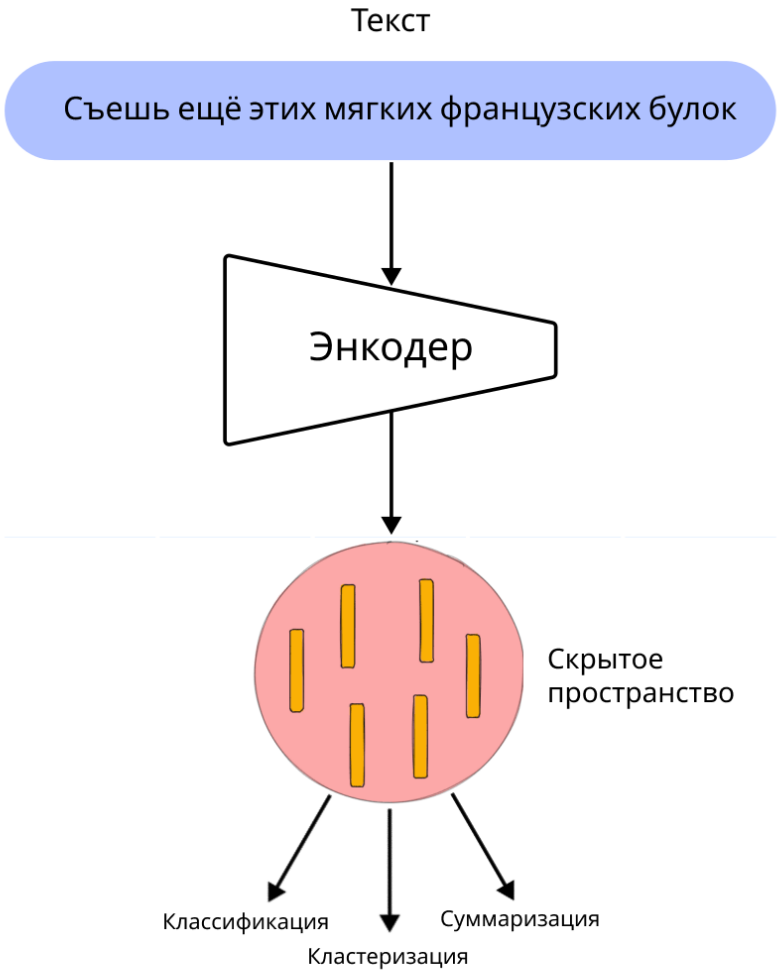
Выполнил: Малашенко Б. Т.
Научный руководитель: Мухин М. С.
Соруководитель: Спирин Е. С.

Что такое эмбеддинги и зачем они нужны

Энкодер переводит текст в скрытое пространство векторов фиксированной размерности,
называемых **эмбеддингами**.

Они наследуют семантические отношения исходных текстов. Похожие по смыслу предложения будут
ближе друг к другу в скрытом пространстве, чем противоположные по смыслу.

Такая структура скрытого пространства позволяет решать задачи:
классификации, кластеризации, суммаризации.



Мотивация

Все лидирующие места в наиболее актуальном русскоязычном бенчмарке «Рейтинг русскоязычных энкодеров предложений» (далее «Энкодечка») [1] занимают мультиязычные модели.

Топ 2 модель на «Энкодечке» находится лишь на 25 месте в англоязычном бенчмарке MTEB [2] среди моделей такого же или меньшего размера.
Все места выше занимают моноязычные модели.

Энкодечка

№	Модель	Encodechka *(Mean S)*	Размер модели *Млн. парам.**
1	bge-m3	0.786	303
2	multilingual-e5-large	0.780	303
3	paraphrase-multilingual-mpnet-base-v2	0.762	85

Massive Text Embedding Benchmark (MTEB)

№**	Модель	MTEB *(Average)*	Размер модели *Млн. парам.**
1	mxbai-embed-large-v1	64.68	303
…	…	…	…
25	multilingual-e5-large	61.5	303

**без учёта слоя эмбеддингов*
***позиция в лидерборде среди моделей размера < 500 Млн парам.*

Цель и задачи

Цель — разработать энкодер текстов для русского языка, который покажет более высокое качество, чем существующие модели.

Задачи:

1. провести обзор предметной области;
2. разработать алгоритм обучения универсального семантического энкодера, содержащий новейшие подходы, подготовить данные для обучения;
3. обучить модель по разработанной методологии;
4. апробировать полученную модель, сравнить с существующими решениями.

Подготовка данных

Эвристика — семантическая связь между парой текстов.

Энкодер текстов обучается без учителя и требует данные в формате пар, связанных различными эвристиками. Эвристики можно разделить на асимметричные и симметричные.

Из открытых источников удалось собрать примерно 2.5 миллиона пар текстов.

Дополнительно 2 датасета собрано нами.

Распределение найденных датасетов по эвристикам

RuHNP

Russian **H**ard **N**on **P**araphrases (**RuHNP**) — датасет перефразированных текстов, особенность которого в количестве сложных отрицательных пар.

Для каждого из 100 тысяч текстов было сгенерировано 5 положительных и 5 отрицательных пар.

Характеристики датасета:

- более миллиона пар;
- нейтральный домен — Википедия;
- медианная длина текста — 115 символов.

Пример из RuHNP

query: После войны продолжал службу в Советской армии на штабной работе.

pos: На штабе в Советской армии продолжал служить после войны.

neg: После войны переквалифицировался и ушел из Советской армии на завод.

RuWANLI

Russian **W**orker and **AI** collaboration for **NLI** (**RuWANLI**) — NLI датасет, который собирался по алгоритму WANLI [3]. В дополнение к синтетической генерации, использовалась фильтрация с помощью аннотаторов.

Характеристики датасета:

- 100000 примеров;
- благодаря наличию троек данных, можно эффективно использовать для обучения энкодера текстов.

Пример из RuWANLI

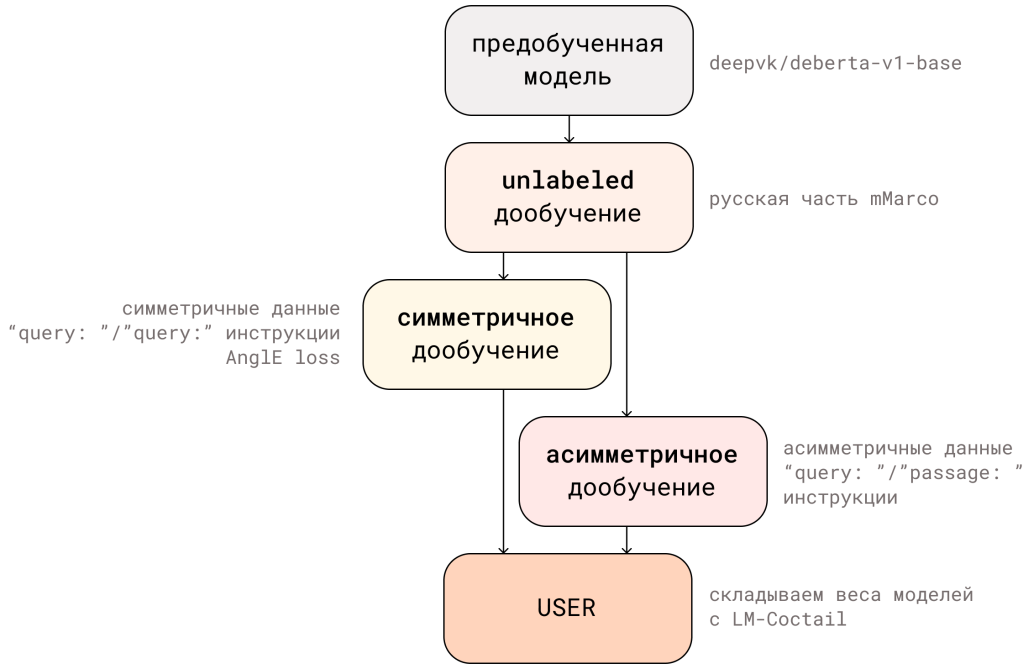
text:	Группа друзей отдыхает на пляже пьют коктейли под пальмами.
entail:	Друзья наслаждаются отдыхом у моря, попивая напитки в тени деревьев.
contr:	Группа друзей находится в городском кафе и кушает бургеры.

Детали обучения

При построении пайплайна я следовали алгоритму обучения BGE-en [4], но в процессе работы пришел к нескольким улучшениям. Шаги алгоритма:

1. предобученная модель deepvk/deberta-v1-base [5];
2. первый этап дообучения на mMarco;
3. добавление инструкций, обучение двух моделей на различных типах данных;
4. объединение в итоговую модель с помощью LM-Coctail [6].

Алгоритм обучения модели USER



Анализ шагов алгоритма

Анализ промежуточных решений показывает, что все шаги улучшают финальное качество.

Шаг 4 предложен нами, без него итоговое качество отличается практически на 0.03 в абсолютном значении.

Оценка вариаций алгоритма

№	Модель	Энкодечка *(Mean S)*
1	USER	0.772
2	USER без инструкций	0.769
3	USER без AnglE loss	0.759
4	USER без разделения	0.743
5	USER без первого этапа дообучения	0.739
6	USER без предобучения	0.711

Сравнение моделей

Энкодечка – наиболее актуальный русскоязычный бенчмарк.

Mean S — среднее значение на всех задачах в Энкодечке, кроме NER.

Mean CLF — среднее значение на задачах классификации: TI, SA, IA.

MTEB — крупнейший бенчмарк, ориентированный на англо и мультиязычные модели. Однако, он содержит 10 подзадач на русском, которыми я дополнил валидацию.

«Энкодечка» лидерборд

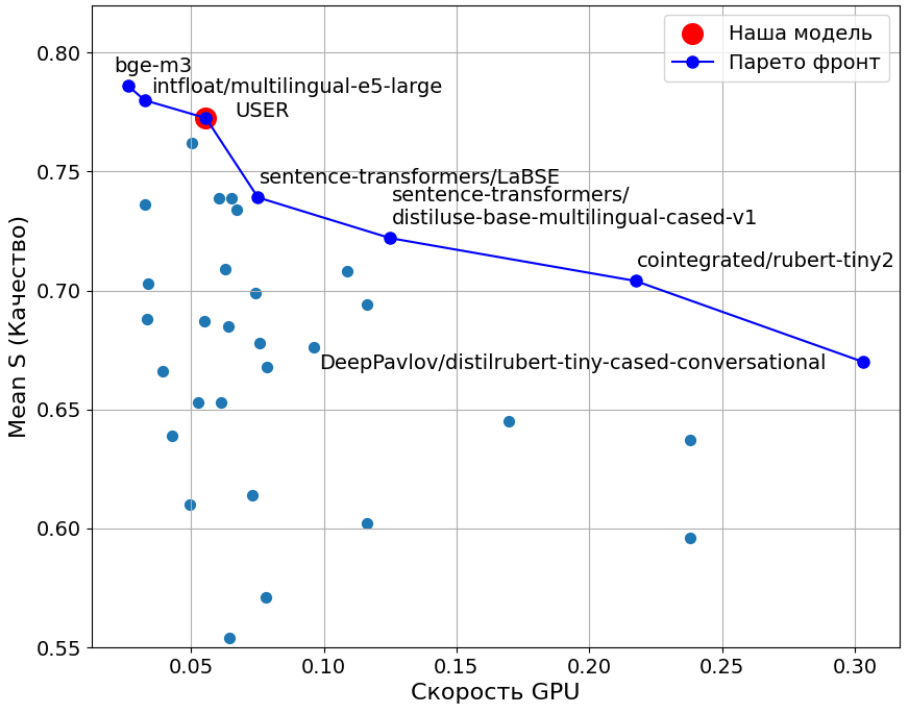
№	Модель	Размер модели *Млн. парам.**	Encodechka *mean S*	Encodechka *mean CLF*	MTEB *ru subset*
1	bge-m3 *(dense)*	303	**0.786**	0.859	**0.705**
2	mult-e5-large	303	0.78	0.863	0.695
3	USER	85	0.772	**0.871**	0.703
4	mpnet-base-v2	85	0.762	0.84	0.675
5	mult-e5-base	85	0.756	0.849	0.676

**без учёта слоя эмбеддингов*

Парето оптимальность

Посмотрим на все модели из Энкодечки, их качество и оценку скорости генерации. Окажется, что полученная нами модель является оптимальной по Парето.

Фронт Парето всех моделей из «Энкодечки»



Дальнейшая работа

MIRACL, MLDR – задачи выделения релевантных текстов, содержат длинные тексты. Ими я дополняю валидацию для более объективной оценки, в частности на Retrieval Augmented Generation (RAG).

Recall@100 — оценка качества обнаружения релевантных документов.

Задачи выделения релевантных текстов

№	Модель	Размер модели *Млн. парам.**	MIRACL *Recall@100*	MLDR *Recall@100*
1	bge-m3 *(dense)*	303	**0.959**	0.6
2	mult-e5-large	303	0.927	**0.66**
3	USER	85	0.763	0.605
4	mpnet-base-v2	85	0.365	0.43
5	mult-e5-base	85	*0.915*	*0.65*

**без учёта слоя эмбеддингов*

Результаты проекта

Рассмотрим результаты:

1. Проведён обзор предметной области, изучены подходы к обучению энкодеров текстов.

2. Разработан алгоритм обучения энкодера текстов с новейшими подходами, предложены улучшения. Подготовлены данные: собрано 2.5 млн положительных пар текстов из 23 датасетов, дополнительно сгенерировано 2 датасета суммарно с 540 тыс. положительных и таким же числом отрицательных пар.

3. Обучена модель в соответствии с предложенным алгоритмом.

4. Проведена апробация полученной модели. На целевой метрике удалось достичь роста качества среди моделей того же размера на 0.012 абсолютных единиц. Однако, дополнительные метрики показали, что модель недостаточно эффективна в задачах извлечения релевантных текстов.

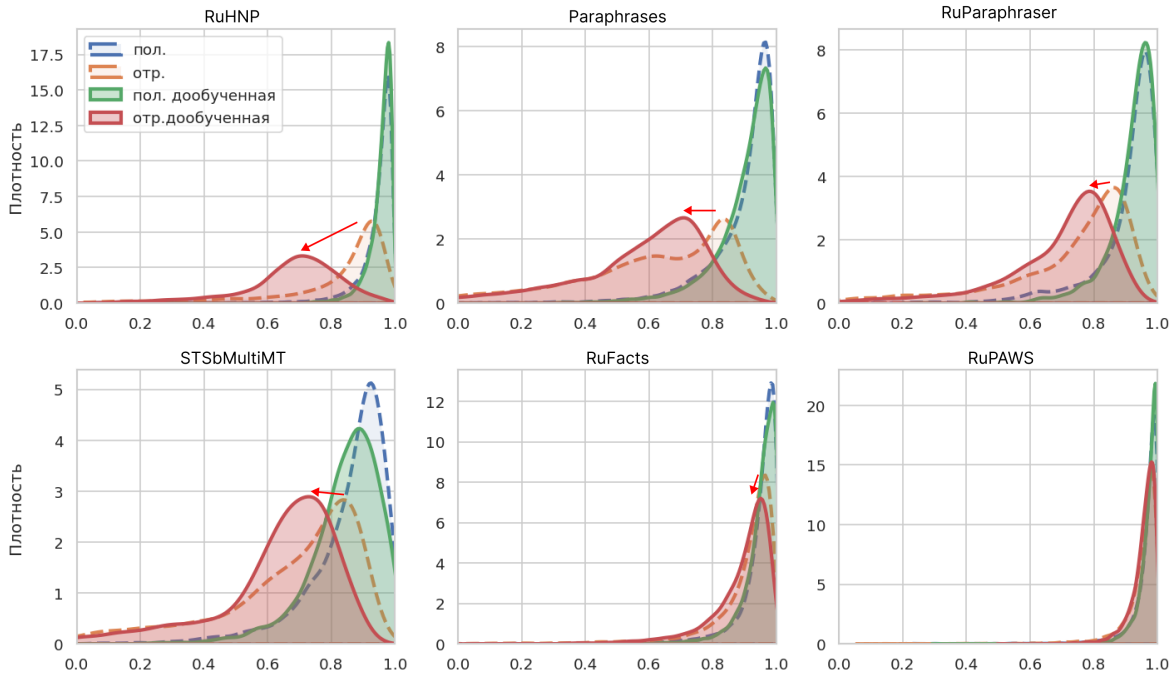

ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •

Оценка качества RuHNP

Для оценки полученного датасета мы провели эксперимент, в котором построили графики распределения косинусных близостей до и после обучения группы моделей на RuHNP.

Датасет консистентен, так как на его тестовой части распределение для отрицательных пар сместилось левее. Аналогично для тестовых частей других датасетов, что говорит о генерализуемости.

Сравнение распределений косинусных близостей



Список источников

1. Dale David. Рейтинг русскоязычных энкодеров предложений. — 2022. — June. — [Online; posted 5-June-2022]. Access mode: https:// habr.com/ru/ articles/669674/.
2. Muennighoff N. et al. MTEB: Massive text embedding benchmark //arXiv preprint arXiv:2210.07316. – 2022.
3. Liu Alisa, Swayamdipta Swabha, Smith Noah A, and Choi Yejin. Wanli: Worker and ai collaboration for natural language inference dataset creation // arXiv preprint arXiv:2201.05955. — 2022.
4. Chen Jianlv, Xiao Shitao, Zhang Peitian, Luo Kun, Lian Defu, and Liu Zheng. Bge m3-embedding: Multi-lingual, multi-functionality, multi- granularity text embeddings through self-knowledge distillation // arXiv preprint arXiv:2402.03216. — 2024.
5. deepvk. DeBERTa-base: Pretrained Bidirectional Encoder for Russian Language. — 2023. — Access mode: https://huggingface.co/deepvk/ deberta-v1-base. Model type: DeBERTa. Languages: Mostly Russian. License: Apache 2.0.
6. Xiao S. et al. Lm-cocktail: Resilient tuning of language models via model merging //arXiv preprint arXiv:2311.13534. – 2023.
7. Li X., Li J. Angle-optimized text embeddings //arXiv preprint arXiv:2309.12871. – 2023.