ИНЖЕНЕРНО-
МАТЕМАТИЧЕСКАЯ
ШКОЛА
НИУ ВШЭ И VK

Universal Sentence Encoder for Russian

Malashenko Boris

btmalashenko@edu.hse.ru

Abstract

The sentence encoder is a crucial component in various pipelines, ranging from traditional classification and clustering tasks to the increasingly popular Retrieval Augmented Generation (RAG). The quality of extracted embeddings often becomes a bottleneck for further processing, underscoring the importance of careful sentence encoder development and selection. Currently, the best solutions for the russian language are predominantly multilingual models. However, large benchmarks like MTEB indicate that monolingual models can achieve superior performance within their specific language. Consequently, we developed the **U**niversal **S**entence **E**ncoder for **R**ussian (**USER**) model, which outperforms all existing models of similar size on Encodechka and MTEB benchmarks by 0.012 and 0.021, respectively. For training, we gathered approximately 3 million text pairs from open sources and generated an additional 2 datasets, adding around half a million more text pairs. To achieve the best quality, we proposed a novel training method involving two models with data split by symmetry, followed by merging into a final model with weight tuning using LM-Cocktail. Our research represents a significant advancement for russian NLP, implementing a novel approach in monolingual model development.

Research goal

In the recent years, all the top models in the ranking of Russian sentence encoders benchmark (Encodechka) are multilingual. However, for other languages, modern monolingual models are more effective than multilingual ones, as demonstrated on the English-language MTEB benchmark. Therefore, we decided to develop a new russian monolingual model using the most up-to-date training approaches, aiming to bridge the performance gap between multilingual and monolingual models for Russian.

Related work

Currently, the top sentence encoders for the Russian language are multilingual models such as Multilingual E5 Text Embeddings, BGE M3-Embedding, and LaBSE. Monolingual models, for example rubert-tiny2 and LaBSE-en-ru, are positioned lower in the rankings. The rubert-tiny2 model is quite effective, but as the name suggests, its tiny size compromises its quality compared to larger models. LaBSE-en-ru is an adaptation of LaBSE through vocabulary reduction, enhancing its speed but not its quality. This indicates that there is limited research on sentence encoders for the russian language.

Our approach

In our work, we aimed to follow the BGE model training algorithm, but we made several improvements along the way. Due to limited resources, we used the pretrained model *deepvk/deberta-v1-base*. The subsequent step

involved *unlabeled* fine-tuning on the Russian part of the mMarco corpus. Next, we trained two separate models: one on *symmetric* data and the other on *asymmetric* data. We modified the instruction design by simplifying the multilingual approach to facilitate easier inference. For symmetric data, we added the instructions *«query: »/«query: »*, and for asymmetric data, we used *«query: »/ «passage: »*. Since we split the data, we could additionally apply the *AnglE* loss to the symmetric model, which enhances performance on symmetric tasks. Finally, we combined the two models, tuning the weights for the merger using *LM-Cocktail* to produce the final model, **USER**.

Data

The sentence encoder is trained on pairs of texts linked by various heuristics, primarily semantic relationships. We collected over 20 datasets from open sources, resulting in approximately 3 million positive pairs. However, effective training also requires hard negative pairs — texts that are lexically similar but semantically opposite. For the russian language, there are very few datasets with hard negative pairs, and they are generally small in size. To address this gap, we created two additional datasets: **RuHNP** (**R**ussian **H**ard **N**on-**P**araphrases) and **RuWANLI** (**R**ussian **W**orker and **AI** Collaboration for **NLI**), which together include over half a million positive pairs and an equal number of hard negative pairs.

Experiments

As a baseline, we chose the current top models from the Encodechka leaderboard table. In addition, we evaluate model on the russian subset of MTEB, which include 10 tasks. Unfortunately, we could not validate the bge-m3 on some MTEB tasks, specifically clustering, due to excessive computational resources. Besides these two benchmarks, we also evaluated the models on the MIRACL. All experiments were conducted using NVIDIA TESLA A100 40 GB GPU and fixed seed at 42. We use validation scripts from the official repositories for each of the tasks.

Results

Table 1 presents a comparison between the baselines and our model. Model sizes are shown, with larger models visually distinct from the others. Absolute leaders in the metrics are highlighted in bold, and the leader among models of our size is underlined.

In this way, our solution outperforms all other models of the same size on both Encodechka and MTEB. Given that the model is slightly underperforming in retrieval tasks relative to existing solutions, we aim to address this in our future research.

Table 1: Models validation results. *Millions of params, excluding embeddings layer.

Model	Size*	**Encodechka** *Mean S*	**MTEB** *Mean Ru*	**Miracl** $Recall_{100}$
bge-m3	303	**0.786**	-	**0.959**
multilingual-e5-large	303	0.78	0.665	0.927
paraphrase-multilingual-mpnet-base-v2	85	0.76	0.625	0.149
multilingual-e5-base	85	0.756	0.645	<u>0.915</u>
LaBSE-en-ru	85	0.74	0.599	0.327
sn-xlm-roberta-base-snli-mnli-anli-xnli	85	0.74	0.593	0.08
USER (this work)	85	<u>0.772</u>	<u>**0.666**</u>	0.763