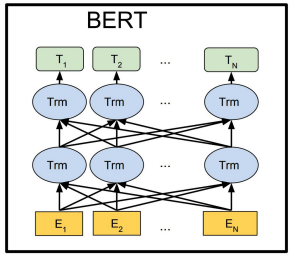
Investigation of approaches for evaluating semantic similarity of texts in the task of matching vacancies and resumes

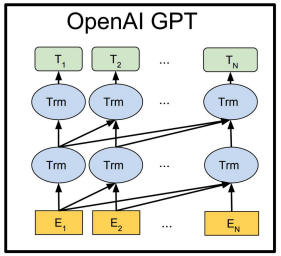
Main idea:

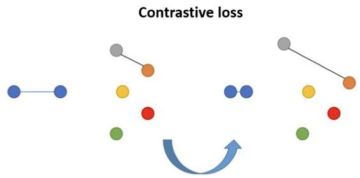
Obtain vector representations of texts and measure text proximity using **cosine similarity**

Method:

1. Ranking resumes by vacancy, quality metric: **MAP@K**
2. Use **Transformer-based models** for text vectorization: BERT & GPT
3. Additional layer training on positive pairs: contrastive loss
4. Fine-tuning models on positive pairs: contrastive loss







$$p@K = \frac{\sum_{k=1}^{K} r^{true}(\pi^{-1}(k))}{K} = \frac{\text{релевантных элементов}}{K}$$

	Random	BERT multilingual	BERT Russian	DeepPavlov RuBERT	RuGPT2 Large	RuGPT3 Large	text-embedding-ada-002
MAP@10	0.03	0.1	0.12	0.06	0.11	0.12	0.24
MAP@20	0.04	0.14	0.15	0.1	0.14	0.15	0.64