Huawei Academic Exchange Day October 4, 2024

ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
ИНФОРМАТИКА И УПРАВЛЕНИЕ
РОССИЙСКОЙ АКАДЕМИИ НАУК

Knowledge Factory: the instrumentalization of Informational Retrieval for Researchers

Konstantin Vorontsov

Dr.Sc., professor of RAS;
Chief Researcher of Intelligent Systems dept., FRC "Computer Science and Control" of RAS;
Head of lab. Machine Learning and Semantic Analysis, Institute of AI, MSU

The motivation of the «Knowledge Factory» project

An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized.
We need a sort of mental clearing house for the mind:
a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared

– Herbert Wells, 1940

Today AI technologies allow us to solve these challenging problems

From Information Retrieval to Knowledge Factory

What is missing from conventional search engines:

- How to search for new knowledge?
- What to do next with what you find?

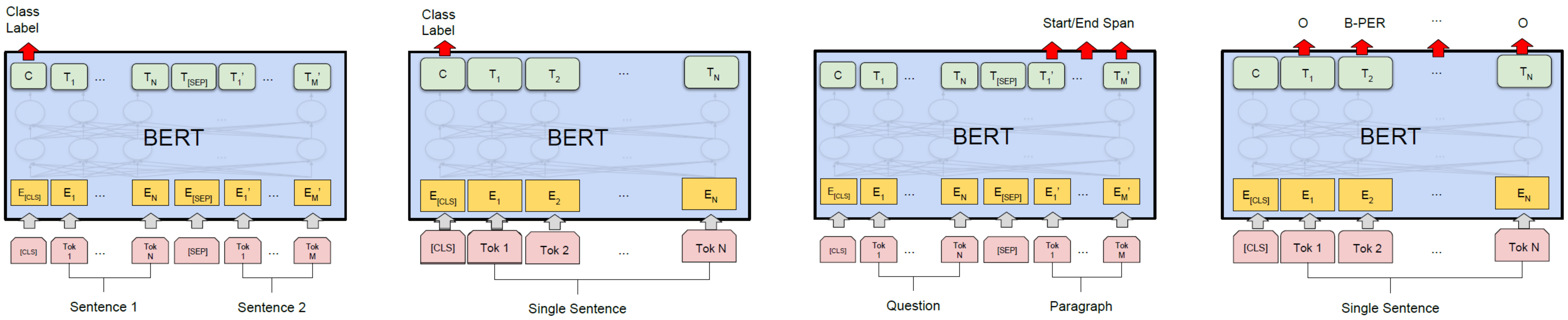
Knowledge Factory is a toolkit for automating further types of operations with large amount of texts (papers, books, manuals, instructions, etc.):

- I seek documents – to save them and accumulate in a collection
- I collect them – to read again and to understand them better
- I understand them – to extract and systematize knowledge from them
- I systematize knowledge – to apply it and to transfer it to other people

Today AI technologies allow us to solve these challenging problems

Transformers: deep neural Large Language Models

- LLM learns to vectorize and predict words from the context
- LLM learns from terabytes of texts, «it has seen everything in languages»
- LLM is multilingual: learn on dozens of languages
- LLM is multitask and multipurpose: for each new NLP/NLU task, pre-trained model & few-shot learning on small data may be sufficient



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

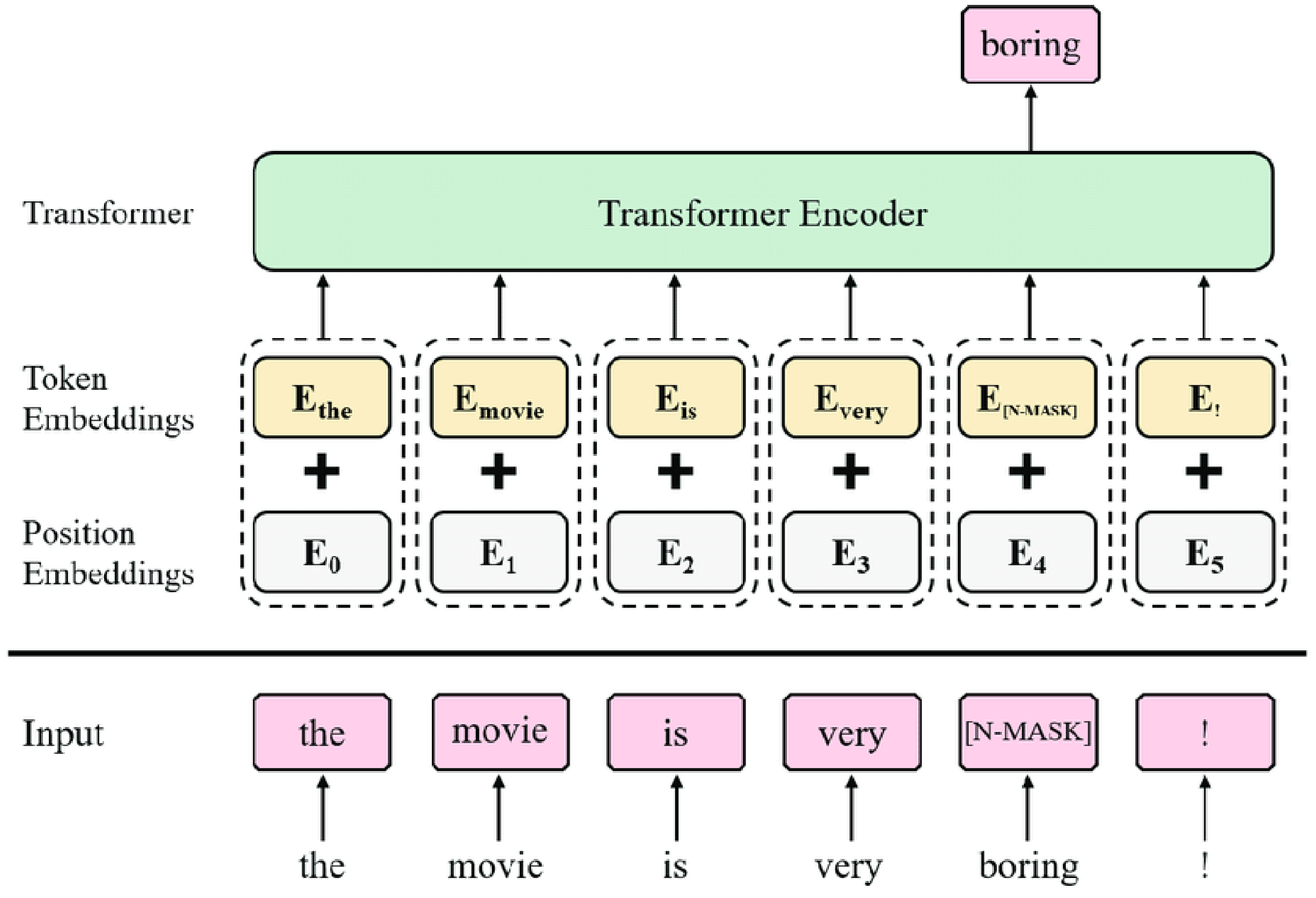
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

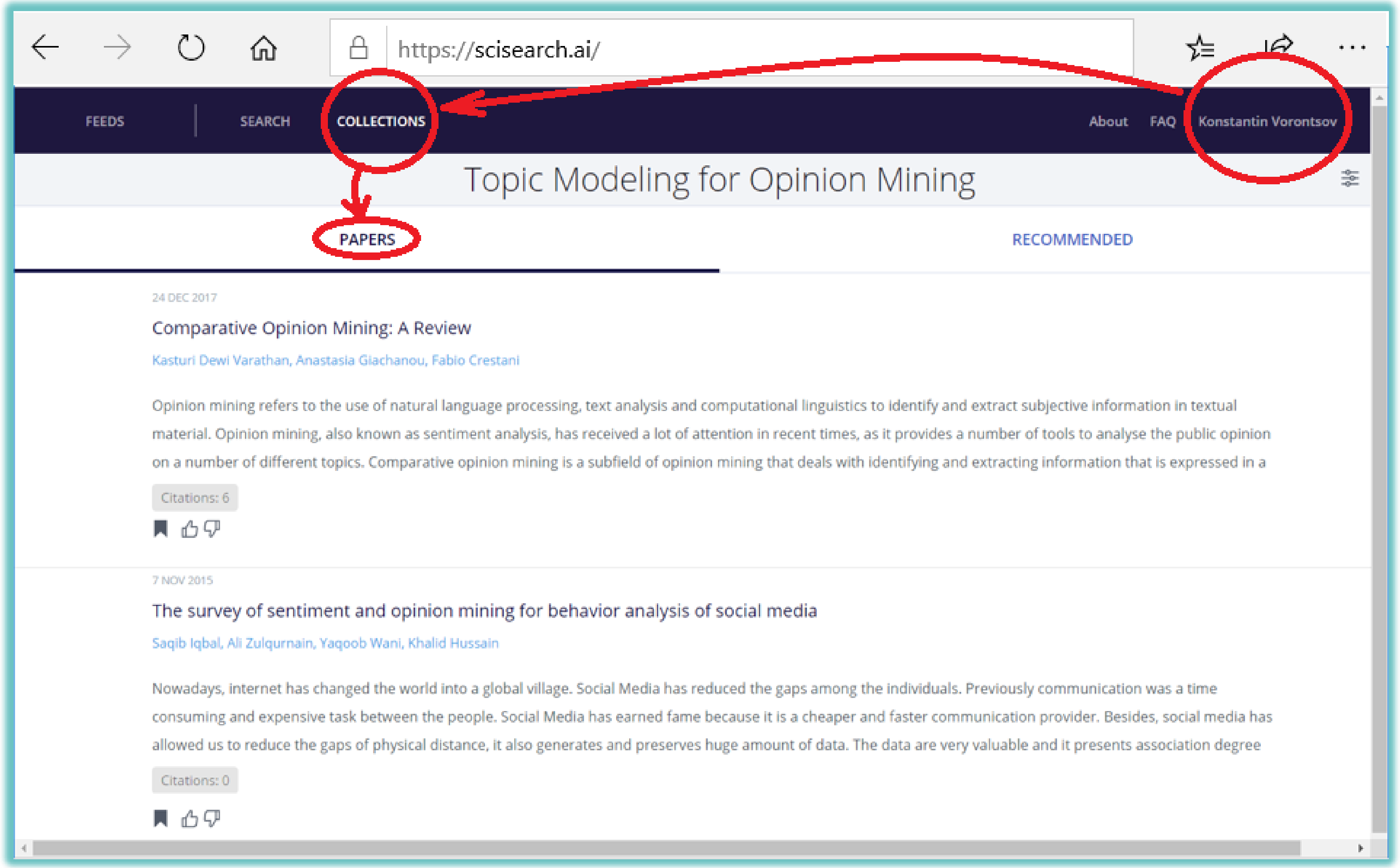
Large Language Models of scientific text

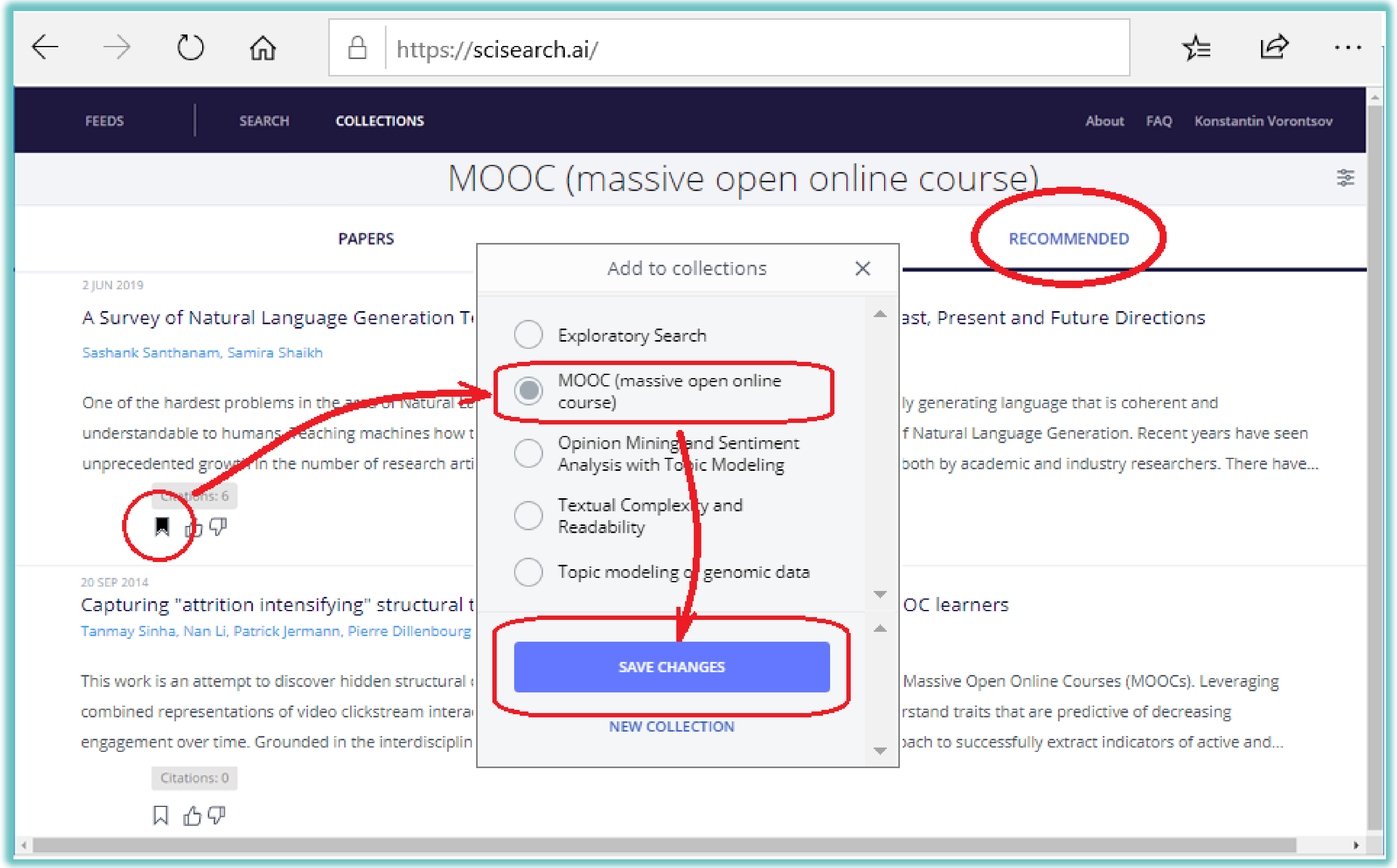
- **SciBERT (2019)** *Beltagy et al.*
  SciBERT: A pretrained language model for scientific text
- **SPECTER (2020)** *Cohan et al.*
  SPECTER: Document-level representation learning using citation-informed transformers
- **LaBSE (2020)** *Feng et al.*
  Language agnostic BERT sentence embedding
- **MPNet (2020)** *Song et al.*
  MPNet: Masked and permuted pre-training for language understanding
- **SPECTER-2 (2022)** *Singh et al.*
  SciRepEval: A multi-format benchmark for scientific document representations
- **SciNCL (2022)** *Ostendorff et al.*
  Neighborhood contrastive learning for scientific document representations with citation embeddings
- **mE5 (2024)** *Wang et al.*
  Multilingual E5 text embeddings: A technical report. 2024.



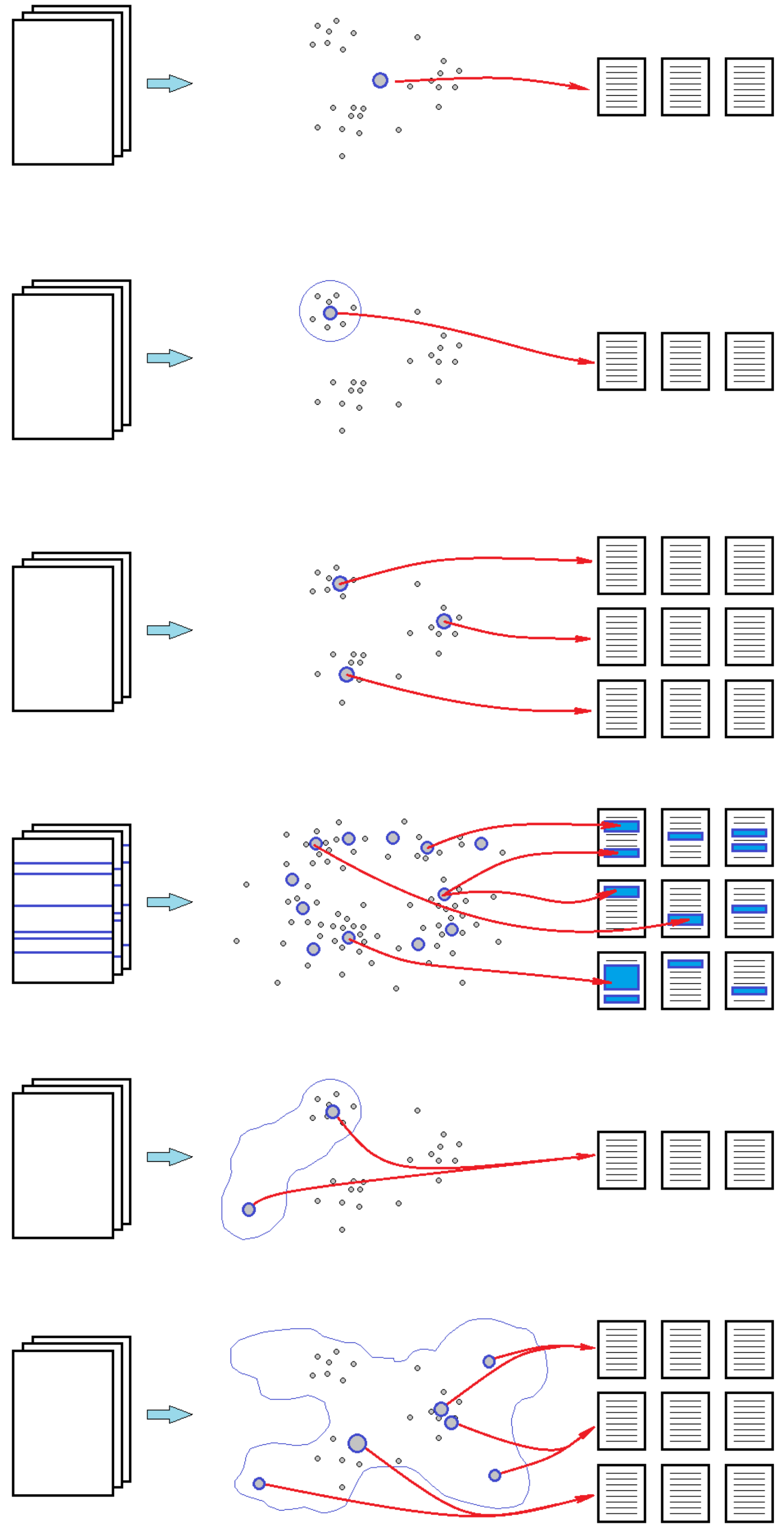
Search and recommendation (GUI prototype)

User's collection plays the role of a search query and search results at the same time





Vector-based document-by-collection search strategies



1. Search by average vector of the collection
   (the simplest, but not the most successful strategy)
2. Search by document from the collection or several semantically similar documents
3. Splitting the collection into clusters and searching by central documents of clusters
4. Splitting documents of the collection into segments and searching by segments of documents
5. Search by documents of related topics for a document or part of documents of the collection
6. Search by topics related to the entire collection

Motivations for our study

The model should be applicable in Russian-language services for searching, recommending, classifying, and analyzing scientific publications (our Knowledge Factory, eLibrary.ru, other scientific electronic libraries)

Model requirements:

- minimization of model size (23M parameters)
- the quality comparable to the best (SOTA) models
- the ability to calculate embeddings without GPUs
- multilingual setting: first English and Russian, then Chinese, Arabic, etc.
- the ability to fine-tune the model on citation data
- quality assessment based on known and new (ours) benchmarks

Datasets

Data for pre-training:

- **S2ORC — Semantic Scholar Open Research Corpus**
  205M publications, 121M authors
  30M (12B tokens) for learning LLM,
  title+abstract, 85% in English, 2% in Russian

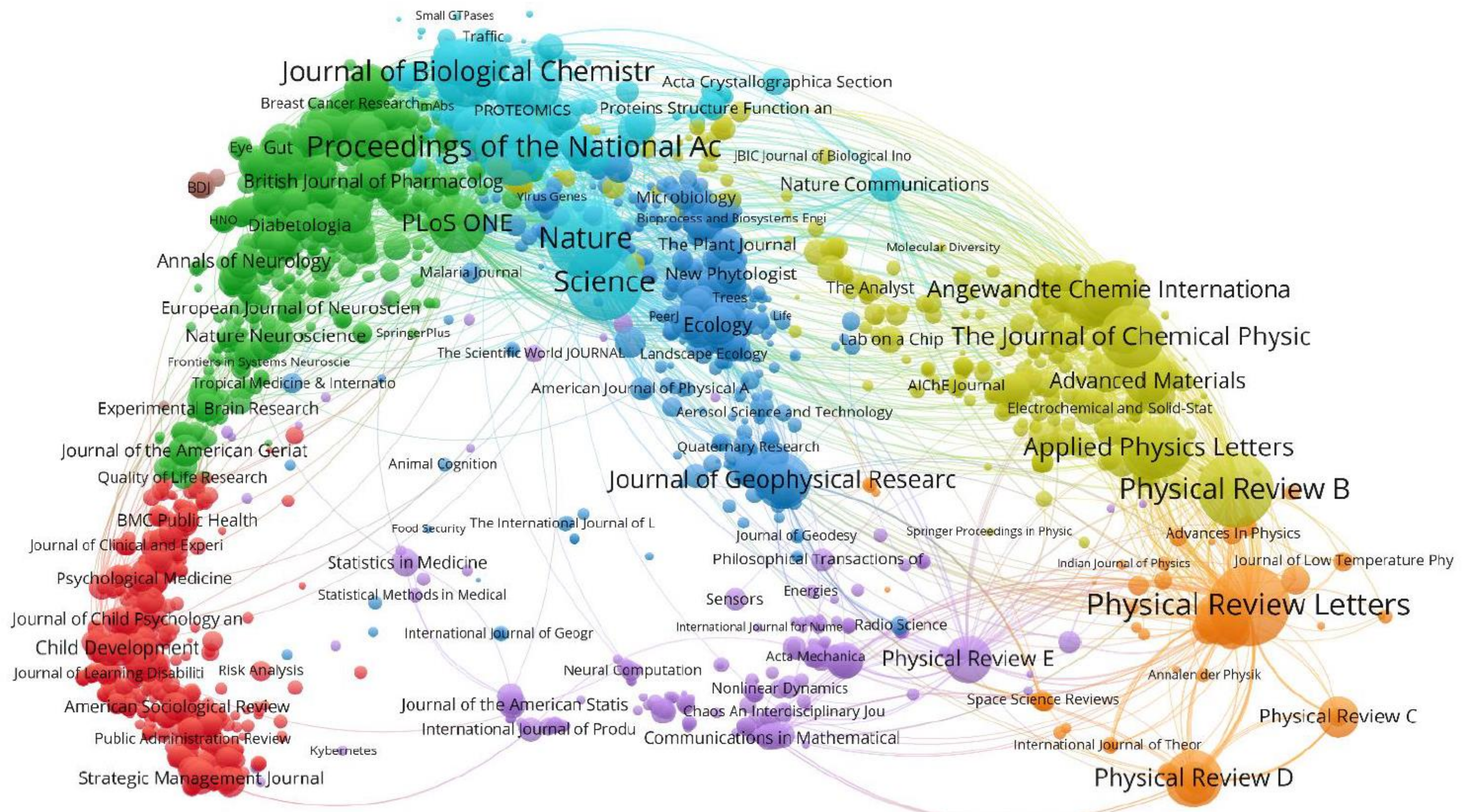
SEMANTIC SCHOLAR

- **eLibrary**, title+abstract:
  8.6M (2B tokens) in Russian
  8.8M (1.2B tokens) in English

eLIBRARY.RU

Data for contrastive trainig:

- **S2AG — Semantic Scholar Academic Graph**
  sources: Crossref, PubMed, Unpaywall и др.
  2.5B citation links



Benchmarks

SciDocs: 6 tasks

- document classification by MeSH categories / topics
- direct citations and co-citations prediction
- user activity prediction, paper recommendations

SciRepEval: 24 tasks, вкл. SciDocs (кроме рекомендаций):

- classification, regression, proximity, and ad-hoc search
- author disambiguation, paper-reviewer matching

(ours) RuSciBench: 8 tasks

- classification by OECD and GRNTI categories (ru / en / ru+en)
- search for an abstract by its translation (ru→en / en→ru)

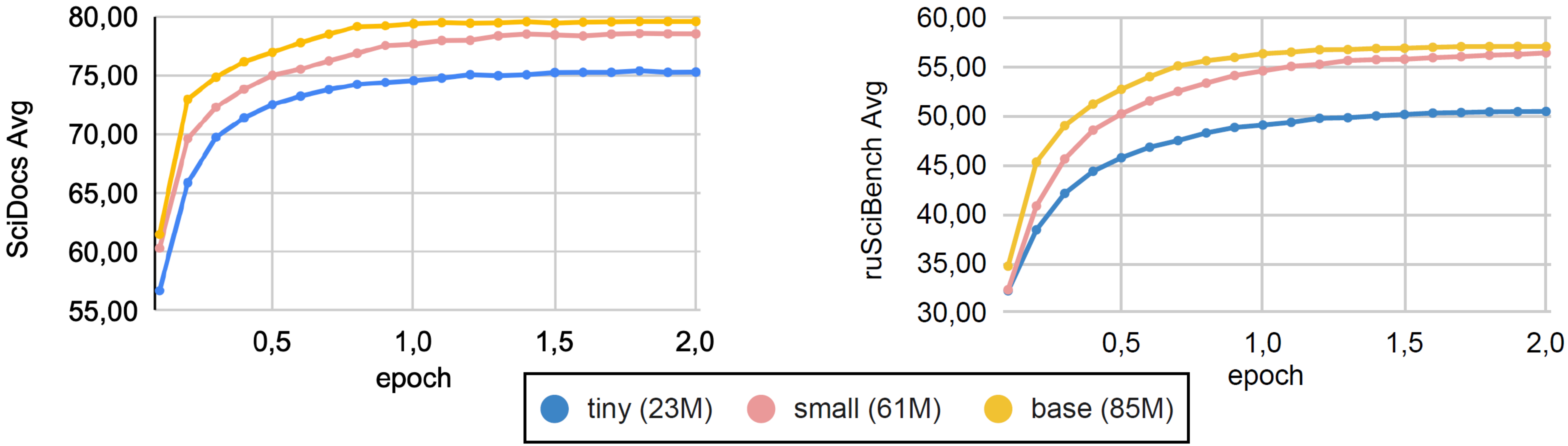
N.Gerasimenko, A.Vatolin. RuSciBench benchmark. 2023. https://github.com/mlsa-iai-msu-lab/ru_sci_bench/tree/main

Stage 1: MLM Pre-training for SciRus-tiny

Base architecture: RoBERTa (Y.Liu et al., 2019) initialized randomly:

tiny (sz=23M, dim=312), small (sz=61M, dim=768), base (sz=85M, dim=1024)

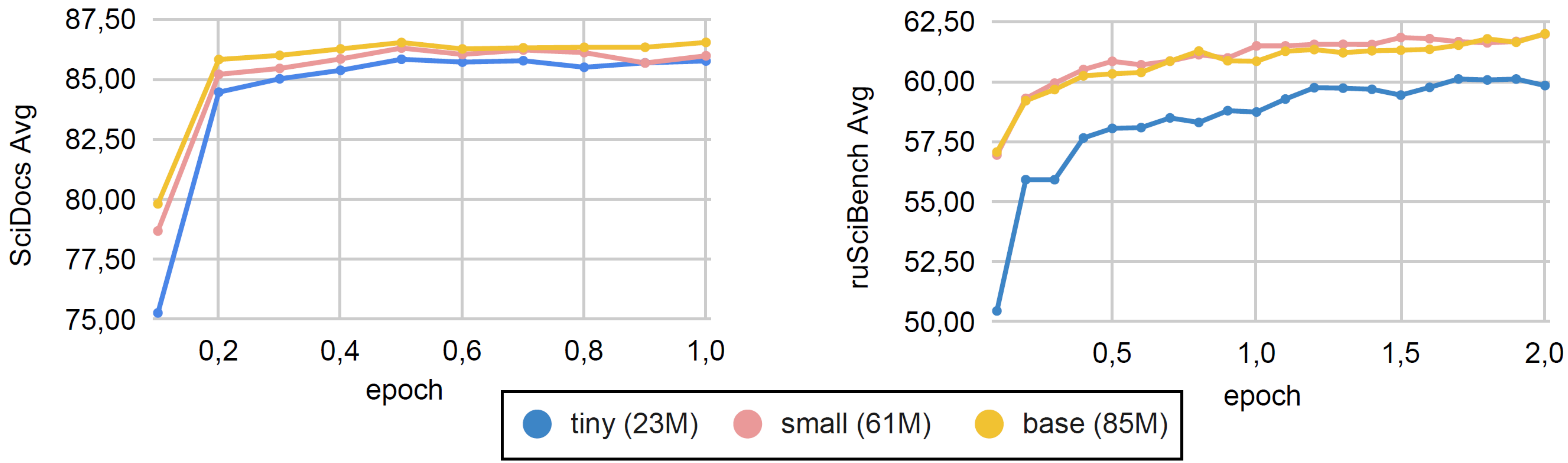
- masked language modeling (MLM)
- two epochs
- Avg — F1-measure, averaged over all benchmark tasks



Stage 2: Contrastive training on title-abstract pairs

Make embeddings closer to each other for all {title, abstract, ru, en) pairs

- 30.6M pairs from S2AG dataset
- 17.8M pairs from eLibrary dataset



L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

Stage 3: Contrastive training on cite/co-cite pairs

Make paper embeddings closer to each other for all (A,B) paper pairs:

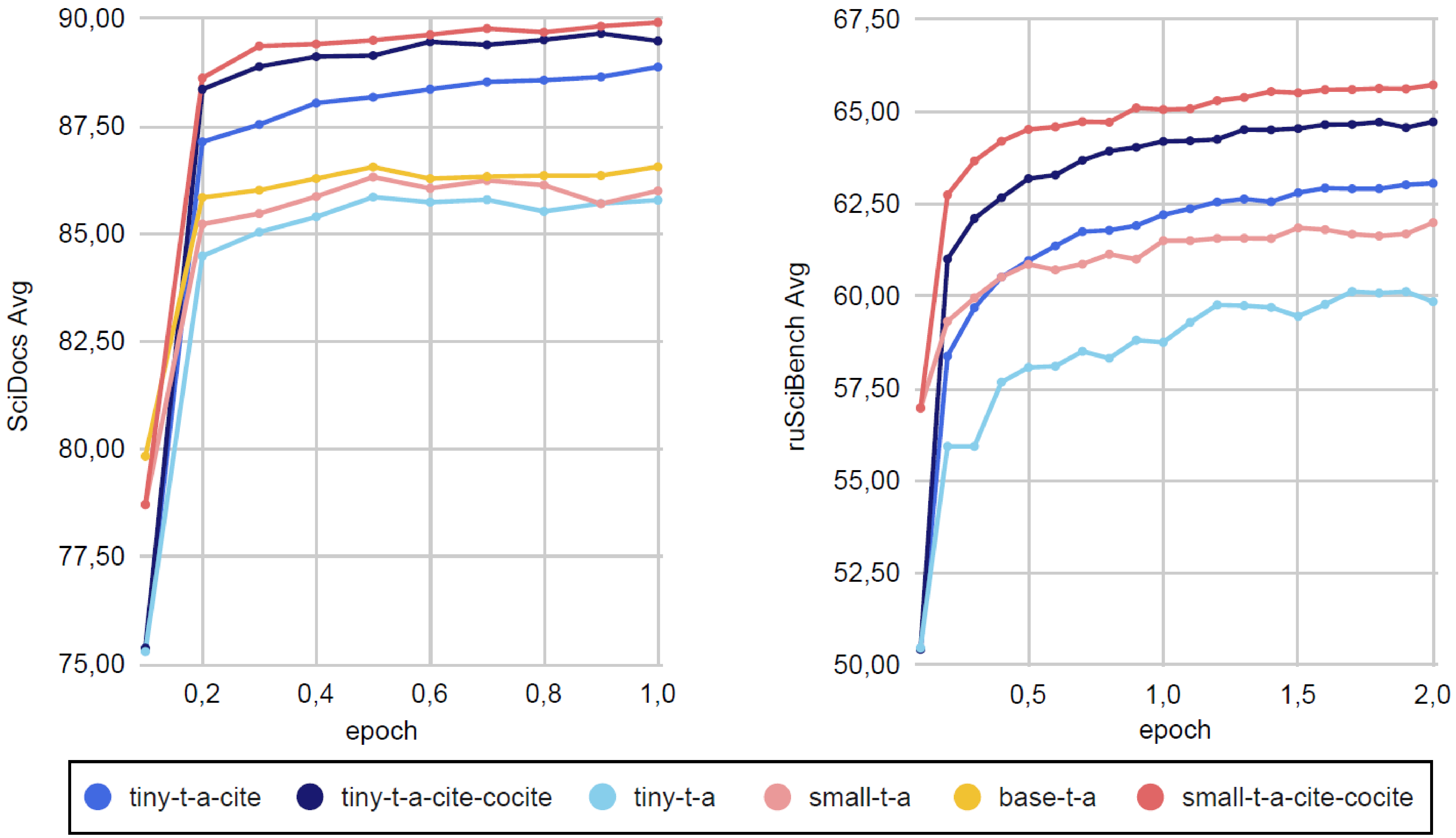
- «cite» — A cites B
- «co-cite» — a third paper C cites both A and B

S2AG:

- 13.3M cite pairs
- 62M co-cite pairs

eLibrary:

- 40M cite pairs
- 33.7M co-cite pairs



Models ranked by SciDocs averaged metrics

SOTA
(state of the art)

Model name	Model size	Avg
all-mpnet-base-v2	110M	**91,03**
Scincl	110M	90,84
scirus-tiny v3 (май 2024)	23M	90,10
e5-large-v2	335M	88,70
e5-base	109M	88,58
e5-base-v2	109M	88,43
multilingual-e5-large	560M	87,53
e5-small-v2	33.4M	86,99
multilingual-e5-base	278M	86,91
e5-mistral-7b-instruct 4byte	7.11B	86,03
scirus-tiny v2 (февраль 2024)	23M	84,21
sentence-transformers/LaBSE	471M	80,78
e5_pretrain_longer_240000_similarity_step_5581	23M	80,51
cointegrated/rubert-tiny2	29.4M	71,60
allenai/scibert_scivocab_uncased	110M	69,04
scirus-tiny v1 (ноябрь 2023)	23M	67,92
nreimers/MiniLM-L6-H384-uncased (e5-small-v2 pretrain)	33.4M	65,68

SciRus quality is better than models that are 5, 20 and even 200 times larger

Models ranked by ruSciBench averaged metrics

🏆 SOTA
(state of the art)

model_name	Model size	elibrary_oecd_full	translation_search	
		macro_f1	ru_en recall@1	en_ru recall@1
e5-mistral-7b-instruct	7.11B	67,28	3,65	18,11
multilingual-e5-large	560M	63,70	99,19	99,37
scirus-tiny3	23M	61,13	94,83	95,81
scirus-tiny2	23M	60,86	96,7	95,11
multilingual-e5-base	278M	62	97	98
LaBSE	471M	60,21	98,31	97,20
LaBSE-en-ru	128M	60,05	98,26	96,93
paraphrase-multilingual-mpnet-base-v2		60,03	66,33	78,18
FRED-T5-large	360M	59,80	22,25	0,79
distiluse-base-multilingual-cased-v1		58,69	92,04	90,83
paraphrase-multilingual-MiniLM-L12-v2		56,48	72,87	77,49
mfaq		**54,84**	**86,75**	**90,11**
scirus-tiny	**23M**	**54,83**	**88**	**88**

SciRus cross-language search quality is close to models that are 20 times larger

Conclusions from the comparison of models

1. **Model size and quality compared to SciNCL**
   — fewer parameters: 23M vs. 110M
   — fewer embedding dimensions: 312 vs. 768
   — longer context: 1024 vs. 512
   — comparable quality (SciDocs Avg): 90.10 vs. 90.84
2. **Contrastive training on title-abstract pairs**
   — significantly improves quality metrics,
   — especially the quality of cross-language search
3. **Contrastive training on cite / co-cite pairs**
   — compensates for the lack of cross-language data

N.Gerasimenko, A.Vatolin, A.Ianina, K.Vorontsov. SciRus: tiny and powerful multilingual encoder for scientific texts. 2024. (Doklady RAS, accepted, in print)
N.Gerasimenko, A.Vatolin, A.Ianina, K.Vorontsov. RuSciBench: open benchmark for russian and english scientific document representations. 2024. (Doklady RAS, accepted, in print)

Implementation

eLIBRARY.RU

«The model developed within the framework of this project is already widely used in the **Scientific Electronic Library** to solve a number of problems related to the assessment of thematic similarity of scientific documents. A useful service for scientists has already been tested by specialists, allowing for a given article or collection of articles to find thematically similar documents both among the entire eLIBRARY.RU dataset (more than 55 million of scientific publications) and only among new acquisitions. An important feature of this model for us is its multilingualism, since the Scientific Electronic Library (SEL) contains documents in many languages»

— Gennady Eremenko, General Director of the SEL

24-04-2024 eLIBRARY.RU Press Release: https://elibrary.ru/projects/news/search_similar_publ.asp

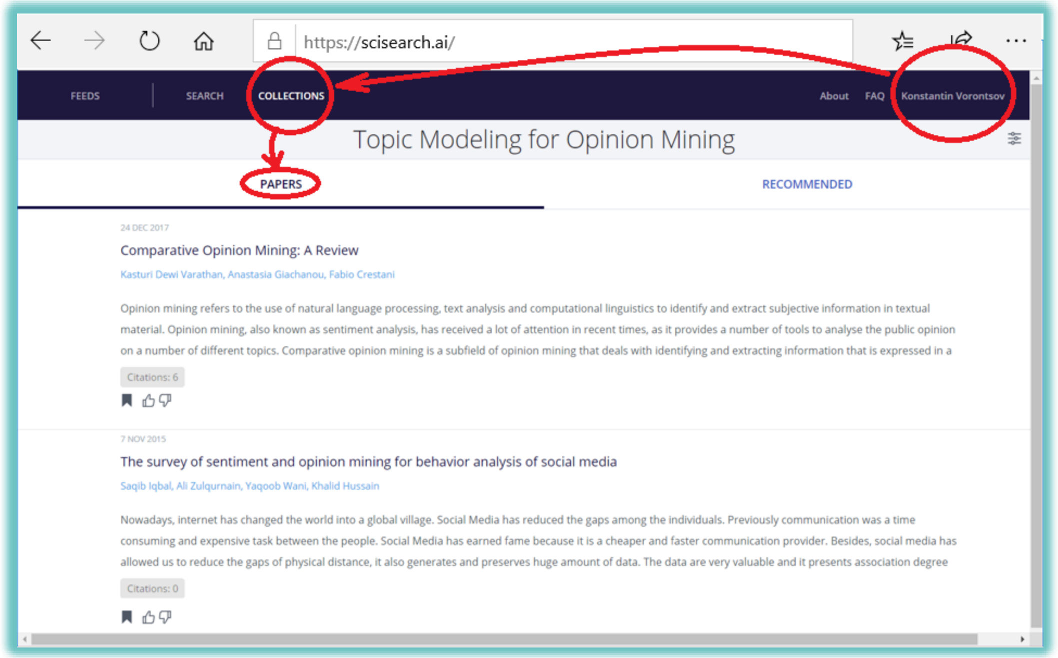
The (planned) services of Knowledge Factory

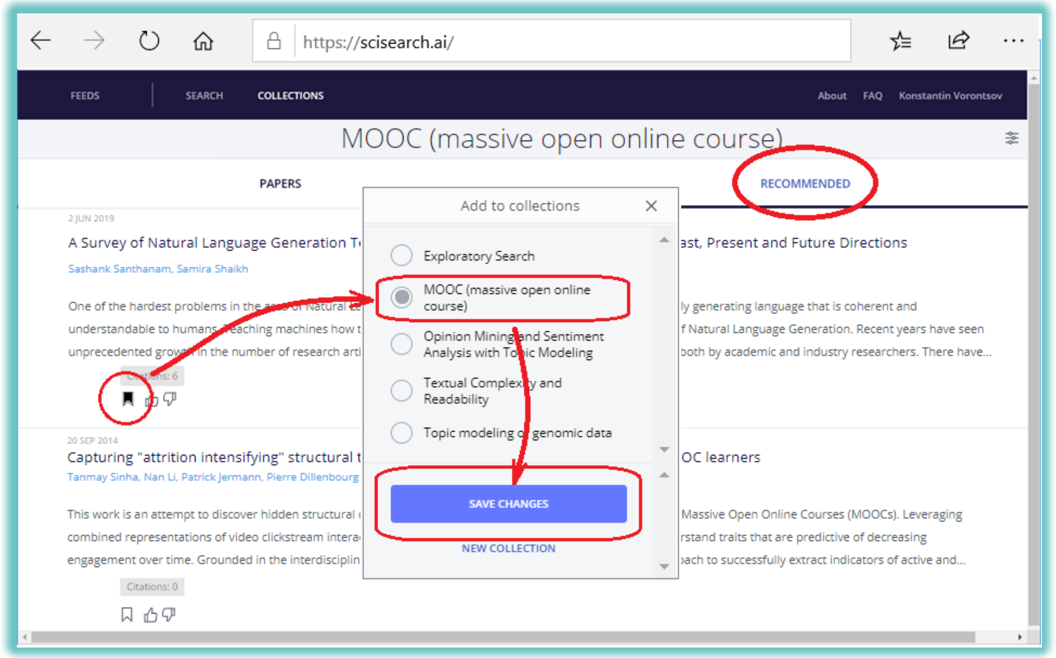
Collection is both a search query and a workspace of the user/group

Extended Collection is a collection joined with search result top-list

Services for search and recommendations:

- search for semantically similar documents by collection
- contextual search by a fragment of the document from the collection
- monitoring of new relevant documents by collection

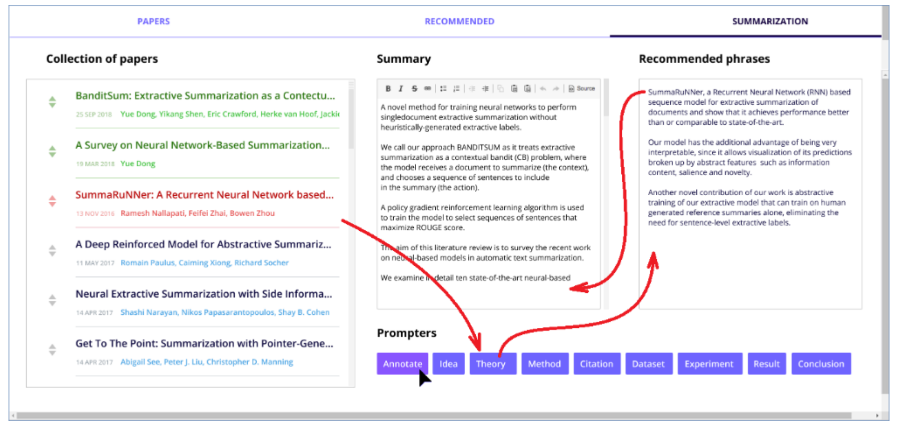


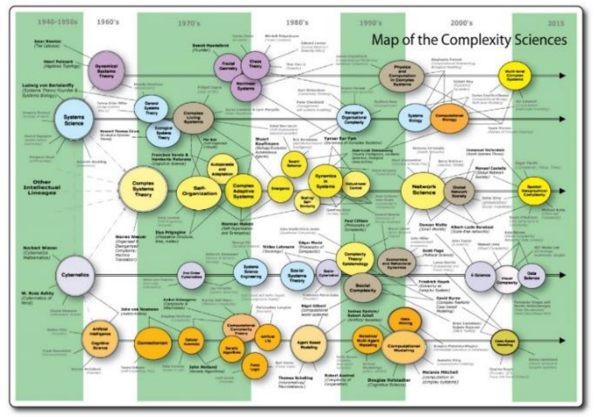


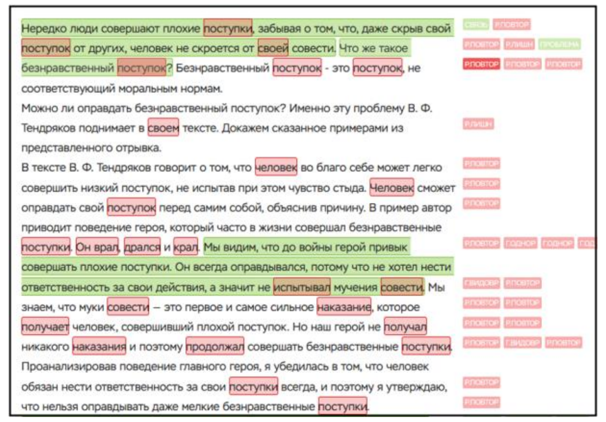
The (planned) services of Knowledge Factory

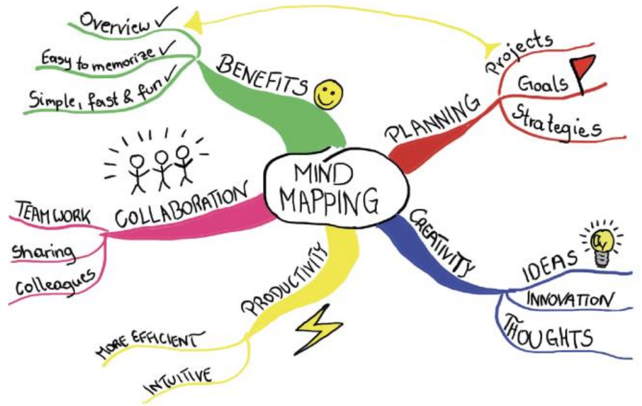
Services for knowledge understanding, analysis, and systematization:

- machine-aided human summarization (MAHS) of the collection
- thematization: extraction of topical clusters from the collection
- mind-mapping: extraction of ideas hierarchy from the collection
- ontologization: extraction of entities and relations from the collection
- chronologization: extraction longtime evolving topics from the collection
- identification of emerging trends from the collection
- content analysis, facts extraction and counting from the collection









Conclusions

Mission: to remove barriers between people and knowledge

Implemented: cross-language document-by-document semantic search

Hope: Large Language Models today (and near future) allow us to solve problems that were considered insurmountable five years ago

ToDo:

— **add services:** document-by-collection semantic search, monitoring, summarization, thematization, ontologization, chronologization, mind-mapping, personalization, trend analysis, content analysis

— **add sources:** project documentation, patents, news,…

— **add languages:** Russian—English—Chinese—…

Thanks!

Konstantin Vorontsov

Dr.Sc., professor of RAS;

Head of lab. Machine Learning and Semantic Analysis,
Institute of AI, MSU;

Head of chair Mathematical Methods of Forecasting,
faculty of CSC, MSU;

Chief Researcher of Intelligent Systems dept.,
FRC "Computer Science and Control" of RAS;

Head of chair Intelligent Systems,
Moscow Institute of Physics and Technology (MIPT)

k.vorontsov@iai.msu.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov