

Международная летняя школа молодых ученых
«Историческая информатика – 2024» 27-30 июня 2024 г.

Технологии искусственного интеллекта в анализе текстов, информационном поиске и социо-гуманитарных исследованиях

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. каф. математических методов прогнозирования МГУ им. М.В. Ломоносова,
зав. лаб. машинного обучения и семантического анализа ИИИ МГУ им. М.В. Ломоносова,
зав. каф. машинного обучения и цифровой гуманитаристики МФТИ, Москва, Россия

k.vorontsov@iai.msu.ru

Содержание

1. Технологии машинного обучения и искусственного интеллекта

- Обзор истории развития технологий ML/AI
- Обзор постановок задач машинного обучения
- Задачи анализа текстов и большие языковые модели

2. Проект «Мастерская знаний»

- Концепция мастерской знаний
- Большая языковая модель научных текстов
- Технологии анализа подборок текстовых документов

3. Языковые модели разметки текста

- Задачи автоматизации контент-анализа
- Методика оценивания моделей разметки
- Задача детекции ценностей социокультурного кода

Методология эмпирической индукции

От дедуктивного метода познания к индуктивному:

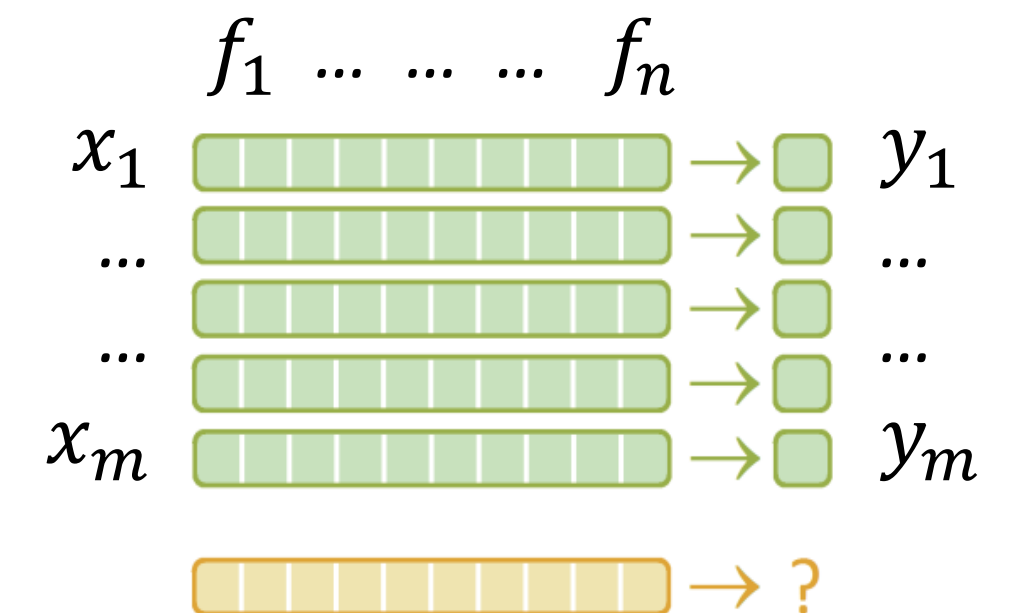
«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»



Фрэнсис Бэкон
(1561--1626)

«Таблица открытия»: множество объектов $\{x_1, \dots, x_m\}$:

- $f_j(x_i)$ – измеряемое значение j -го признака объекта x_i
- y_i – измеряемое значение целевого свойства x_i , либо $y_i \in \{0,1\}$ – отсутствие или наличие целевого свойства



Задачи машинного обучения с учителем

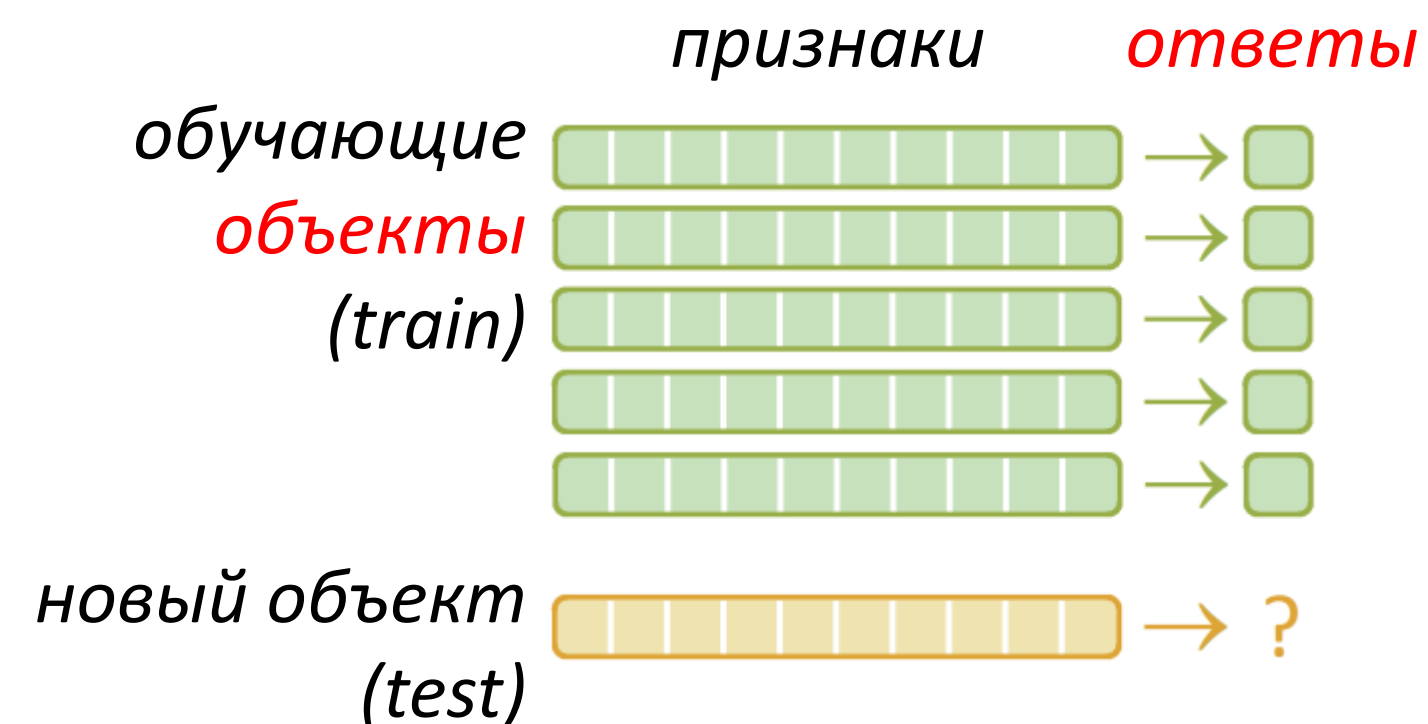
Этап №1 – обучение с учителем

- **На входе:**
данные – выборка прецедентов «**объект** → **ответ**»,
каждый объект описывается набором *признаков*
- **На выходе:**
модель, предсказывающая ответ по объекту

Если нет данных,
то нет
и машинного
обучения

Этап №2 – применение

- **На входе:**
данные – новый **объект**
- **На выходе:**
предсказание **ответа** на новом объекте

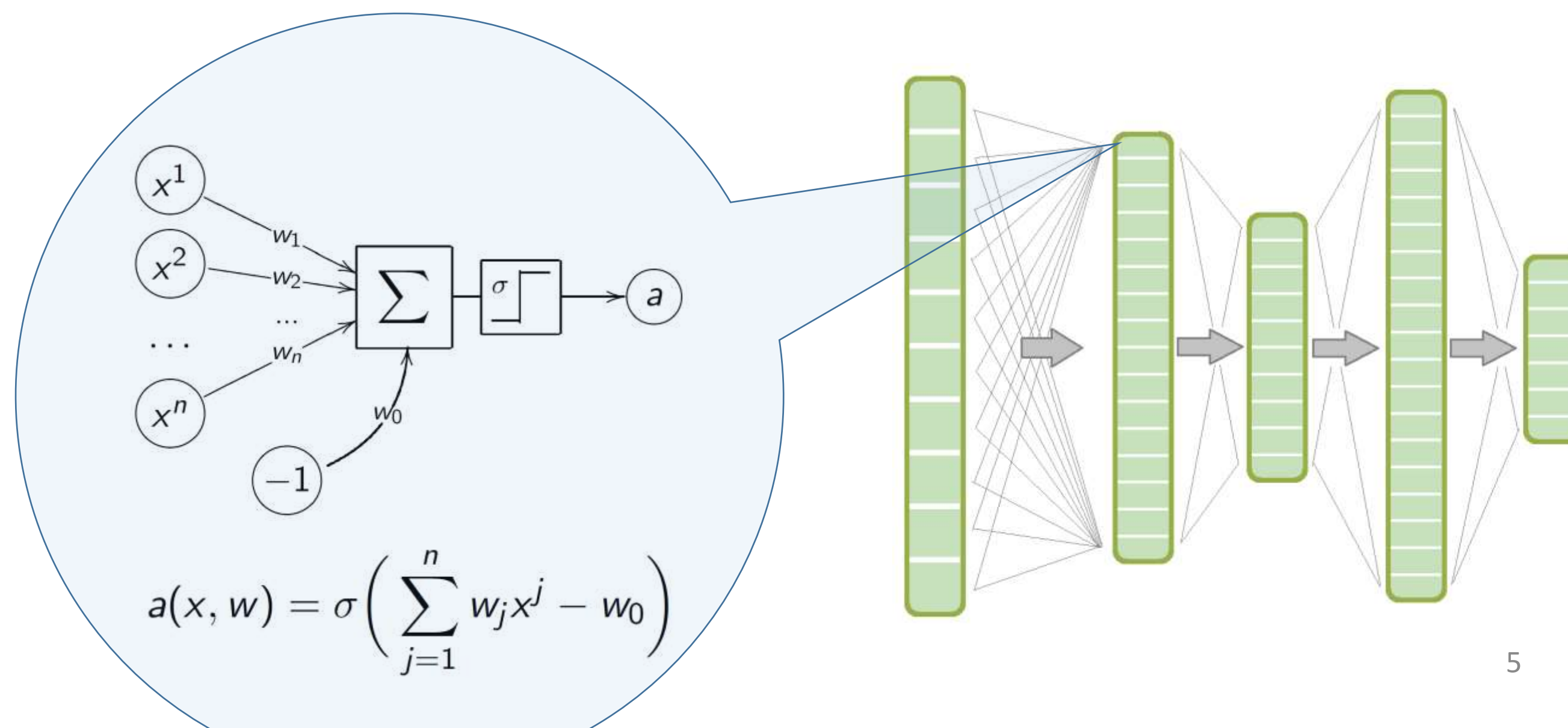
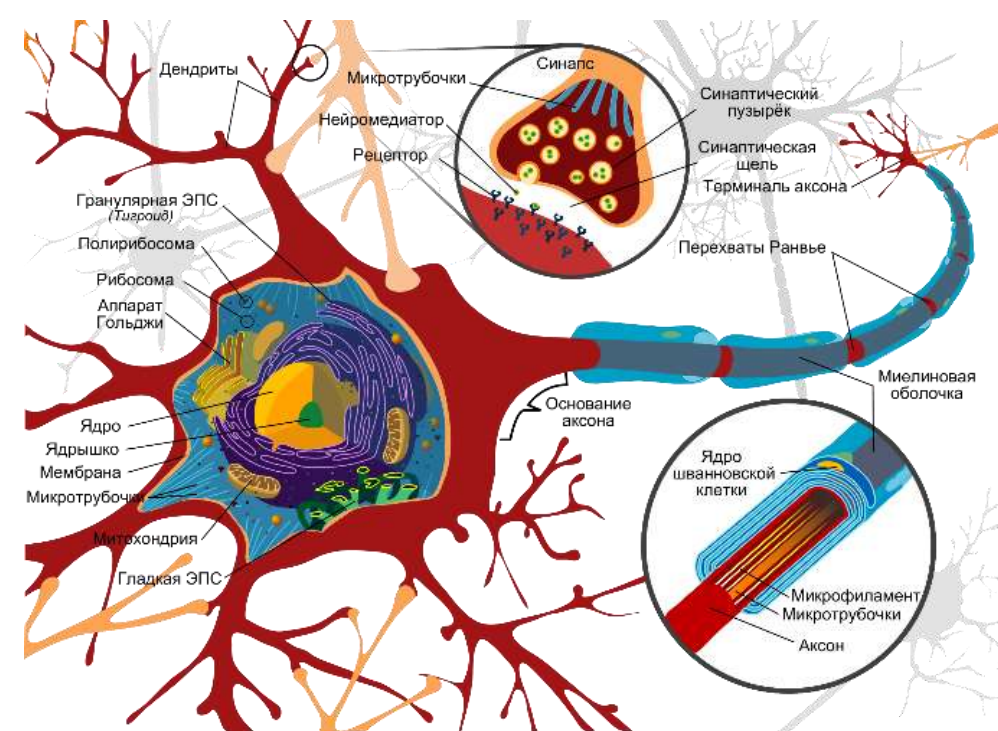


Модели искусственных нейронных сетей

На каждом слое сети вектор объекта преобразуется в новый вектор

Каждое преобразование (нейрон) – линейная модель $a(x, w)$

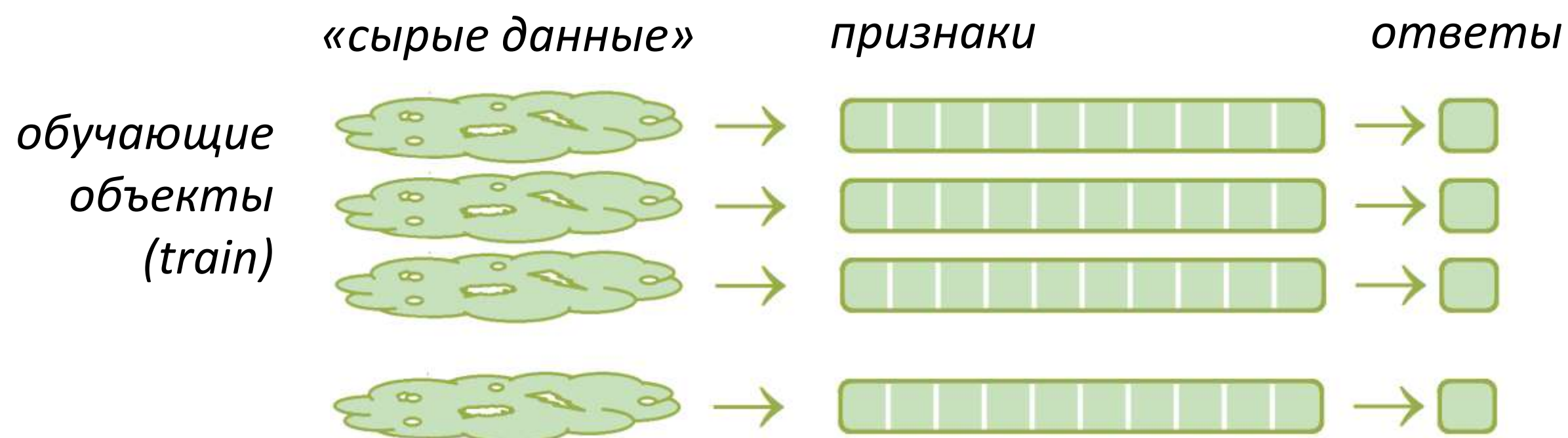
Веса w являются обучаемыми параметрами модели



Модели глубоких искусственных нейронных сетей

Вход: сложно структурированные «сырые» данные объектов

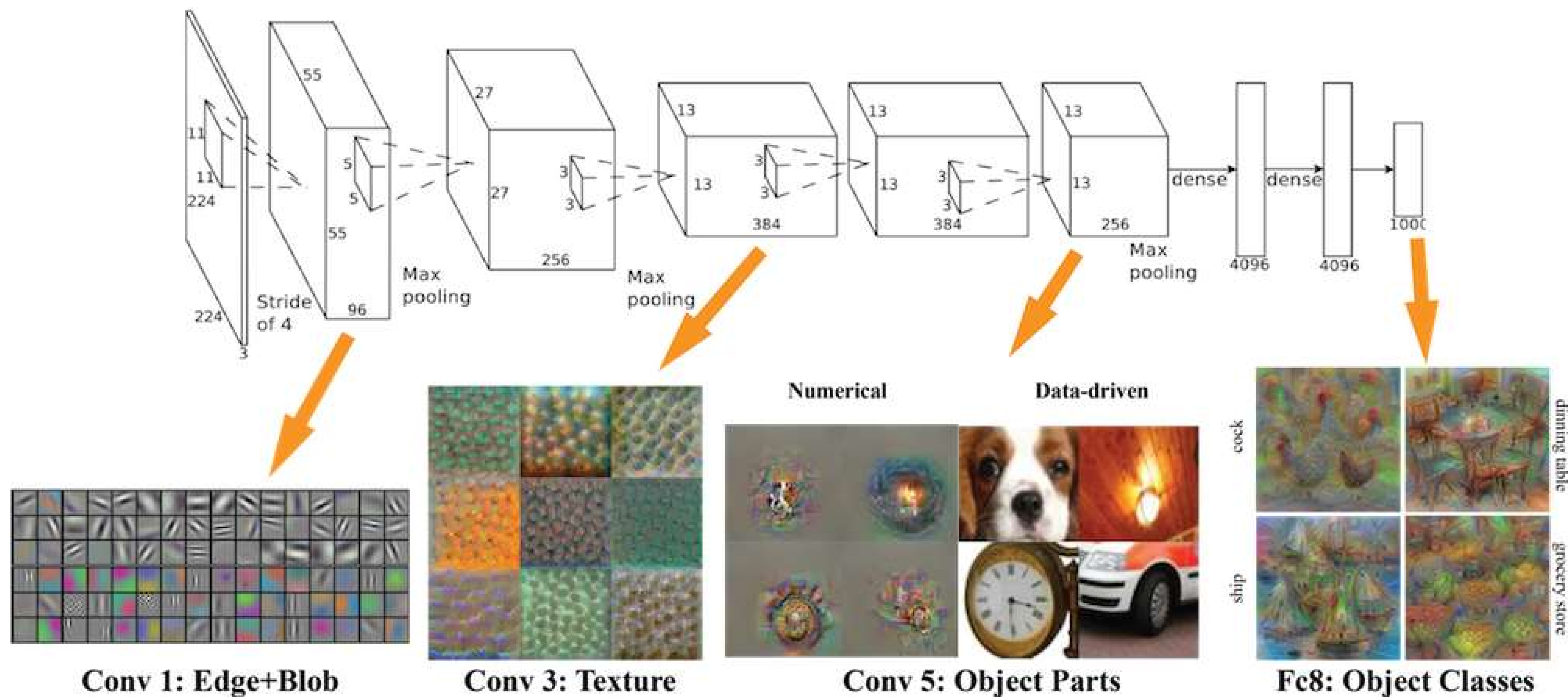
Выход: векторные признаковые представления объектов, затем ответы



*Deep Learning – это
всего лишь обучаемая
векторизация
сложных объектов*

Примеры сложно структурированных объектов:
изображения, видео, временные ряды, тексты, транзакции, графы, ...

Глубокие свёрточные нейронные сети для классификации объектов на изображениях

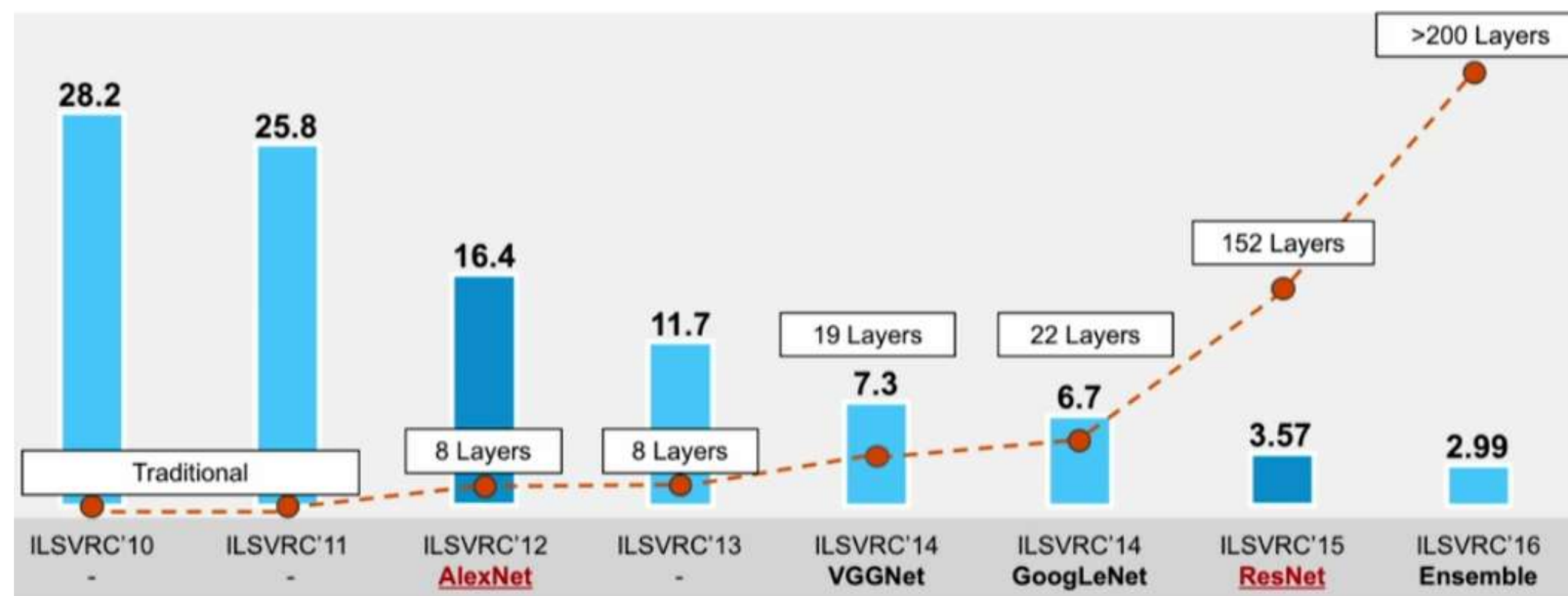


Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. 2012.

Роль больших данных

ImageNet: открытая выборка 14М изображений, 20К категорий

IMAGENET



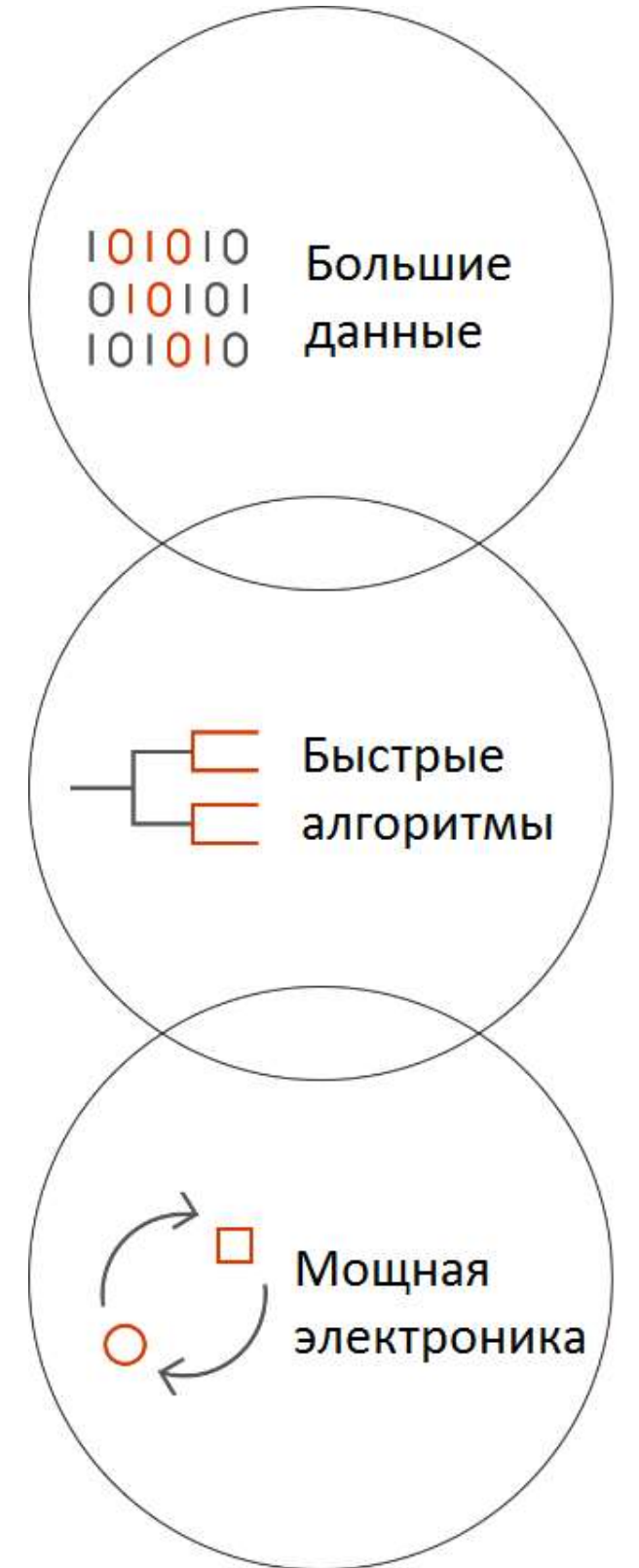
Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

Li Fei-Fei et al. ImageNet: A large-scale hierarchical image database. 2009.

Li Fei-Fei et al. Construction and analysis of a large scale image ontology. 2009.

Три составляющих успеха Deep Learning

- Повсеместное применение компьютерных технологий
→ *накопление больших выборок данных*
в частности, ImageNet
- Развитие математических методов и алгоритмов
→ *накопление критической массы опыта*
методы оптимизации для больших размерностей
- Достижения микроэлектроники
→ *рост вычислительных мощностей, закон Мура*
в частности, графические ускорители (GPU)



Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки

$a(x, w)$ – предсказательная модель

w – параметры модели

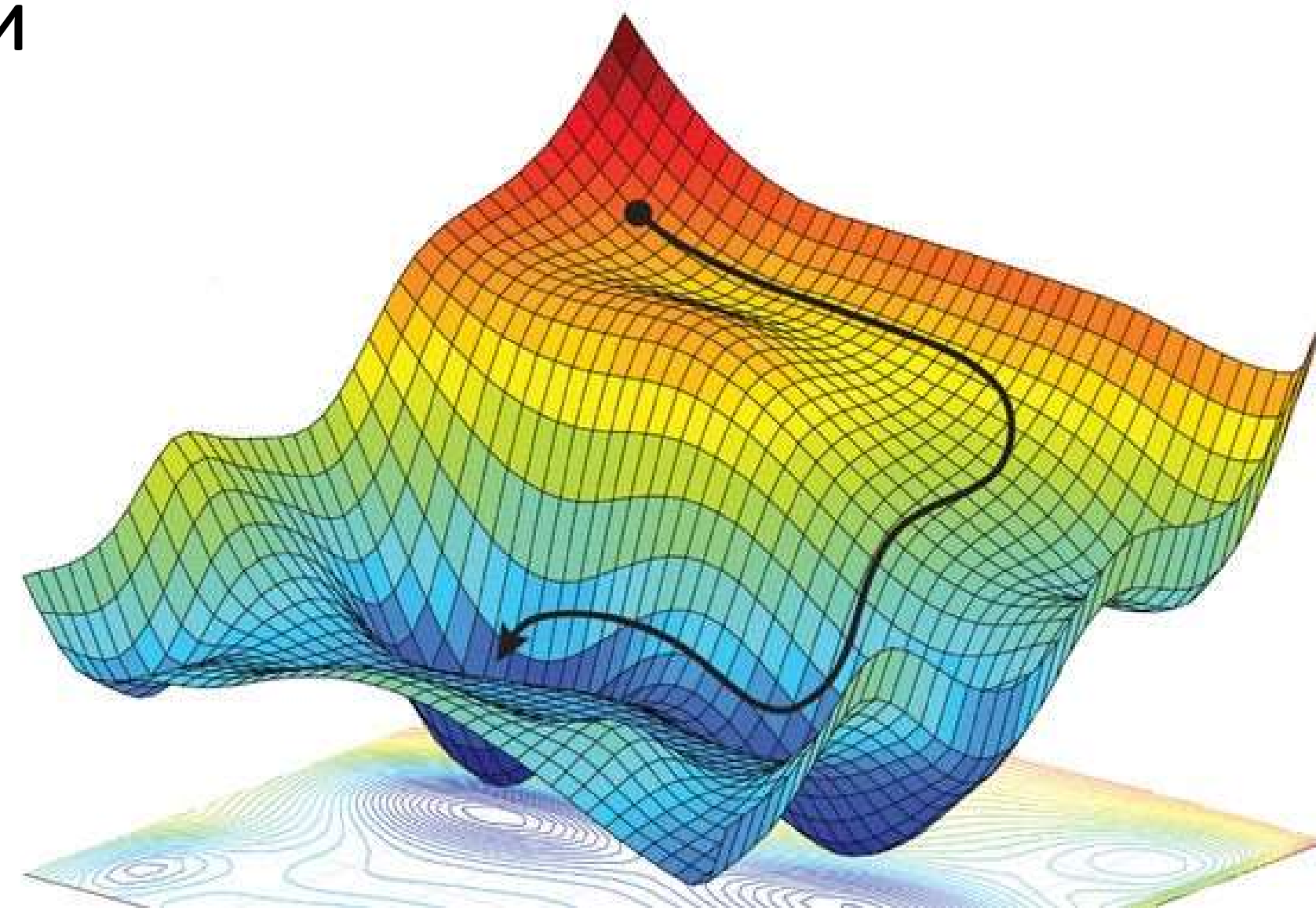
$\text{Loss}(x, w)$ – функция потерь

$Q(w)$ – критерий качества модели

Задача обучения параметров модели:

$$Q(w) = \sum_x \text{Loss}(x, w) \rightarrow \min$$

Способ решения – численные методы оптимизации



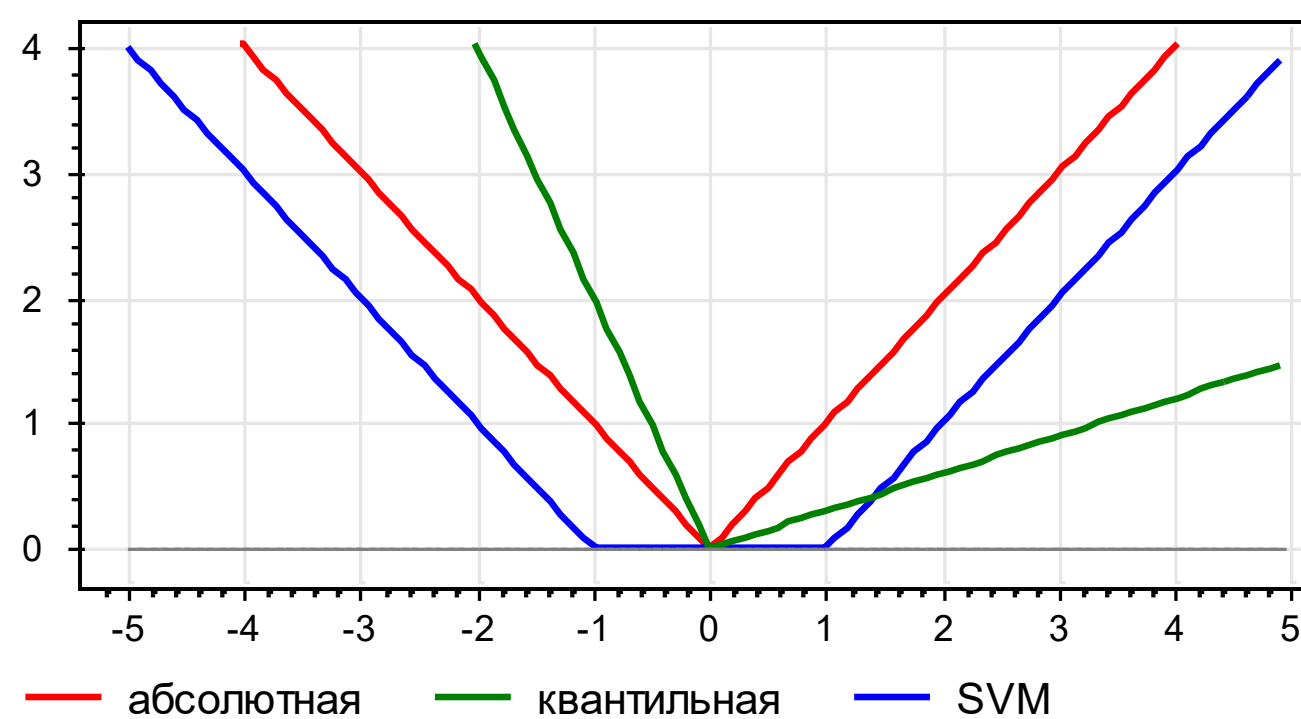
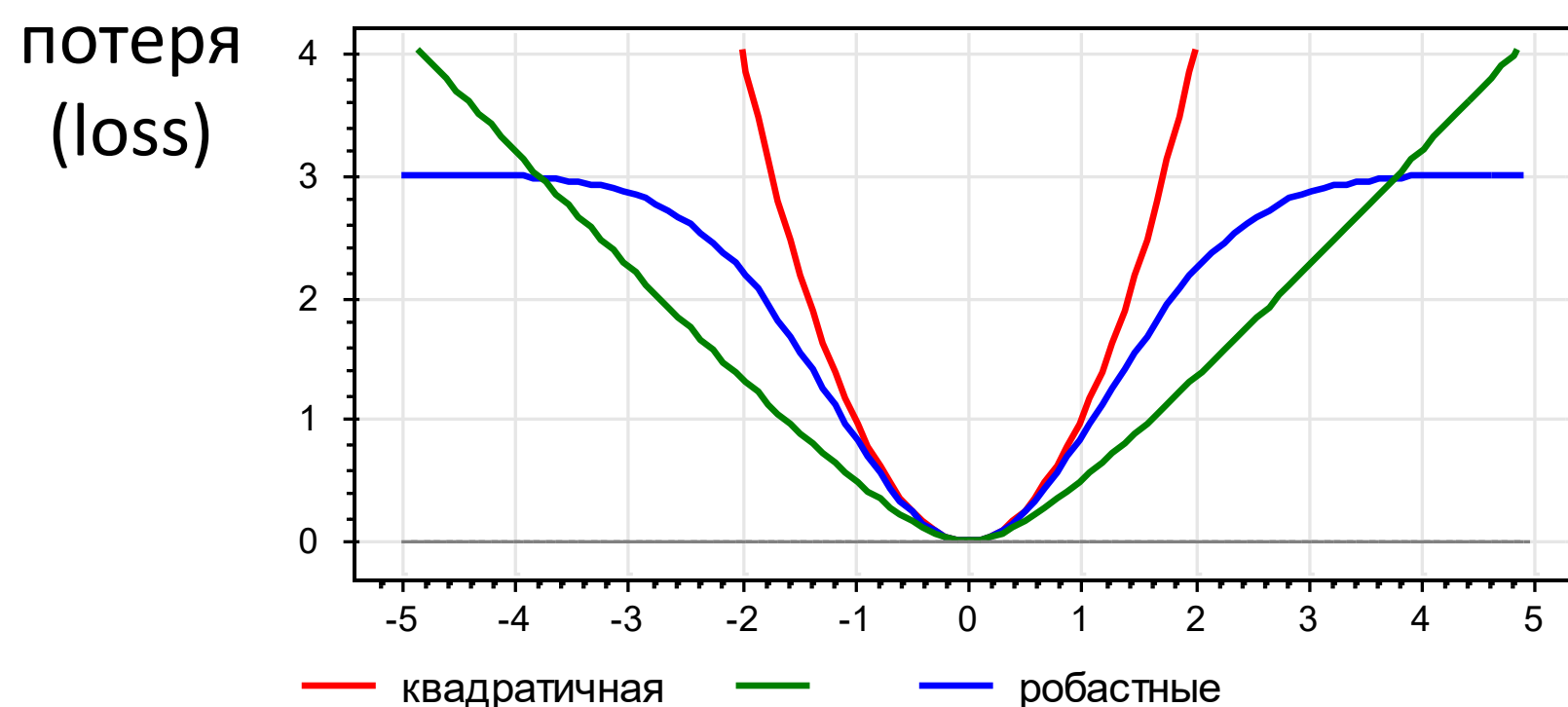
Обучение с учителем (supervised learning): восстановление регрессии (regression)

x — вектор объекта обучающей выборки, y — числовой ответ

$a(x, w)$ — модель регрессии с параметрами w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель регрессии

$\text{Loss}(x, w) = (a(x, w) - y)^2$ — квадратичная функция потерь



НЕВЯЗКА
(error)

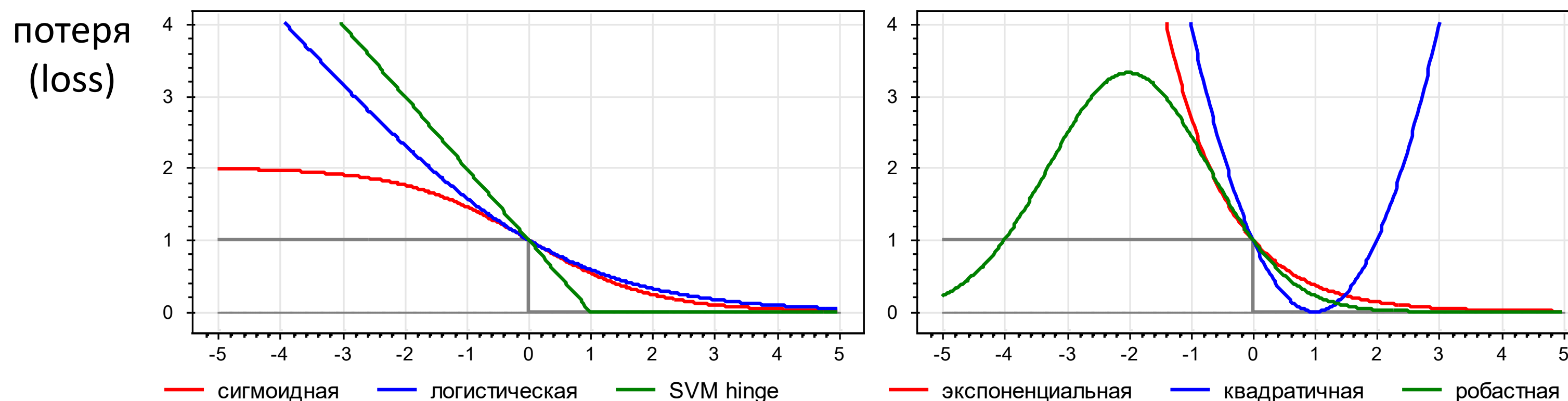
Обучение с учителем (supervised learning): классификация (classification)

x — вектор объекта обучающей выборки, y — ответ (+1 или -1)

$a(x, w)$ — модель классификации с параметрами w

Например, $a(x, w) = \text{sign}(\sum_j w_j x_j)$ — линейная модель

$\text{Loss}(x, w) = \max(0, 1 - y \sum_j w_j x_j)$ — функция потерь SVM hinge



отступ
(margin)

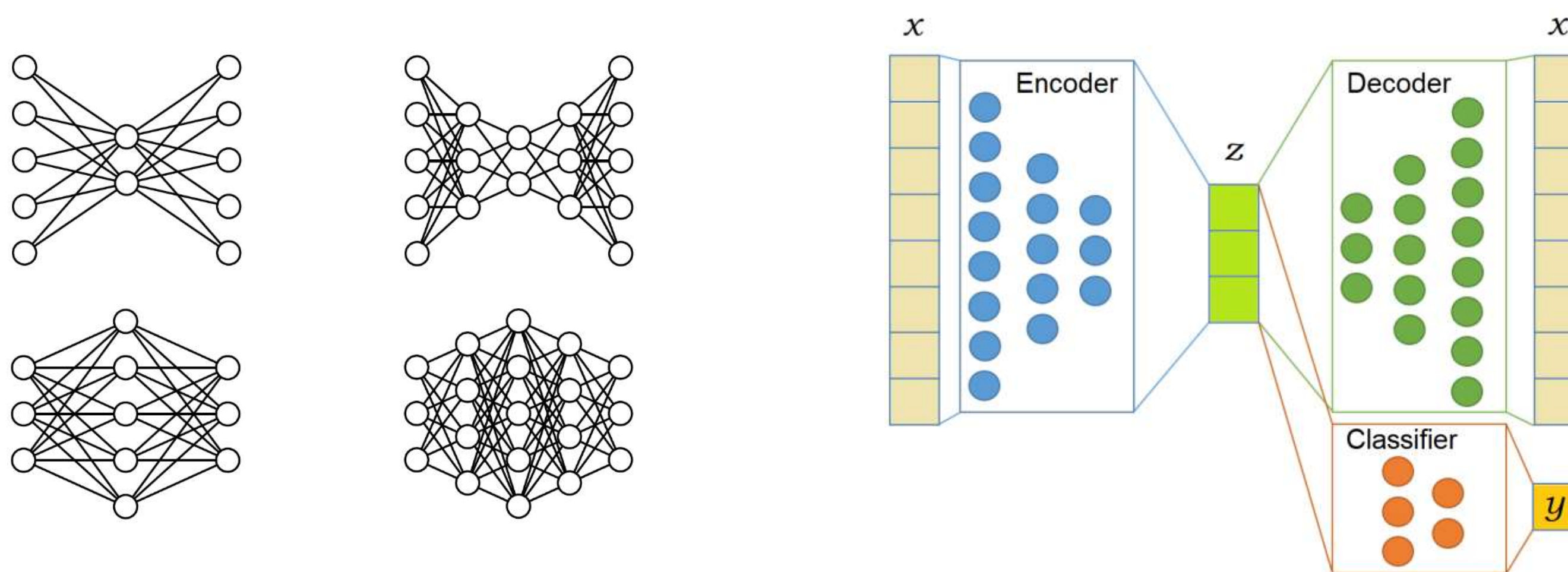
Обучение без учителя (unsupervised learning): векторизация, автокодировщик (autoencoder)

x — описание объекта обучающей выборки, ответов не дано

$z = f(x, w)$ — модель кодирования x в векторное представление z

$x' = g(z, w')$ — модель декодирования z в реконструкцию x'

$\text{Loss}(x, w) = \|g(f(x, w), w') - x\|$ — точность реконструкции объекта



*обучаемая
векторизация
сложных
объектов*

Перенос обучения (transfer learning), предобучение модели векторизации

$z = f(x, w)$ — модель векторизации, универсальная для многих задач

$y = g(z, w')$ — часть модели, специфичная для своей задачи

$\min_{w, w'}: \sum_x \text{Loss}_1(g_1(f(x, w), w'))$ — обучение по большим данным

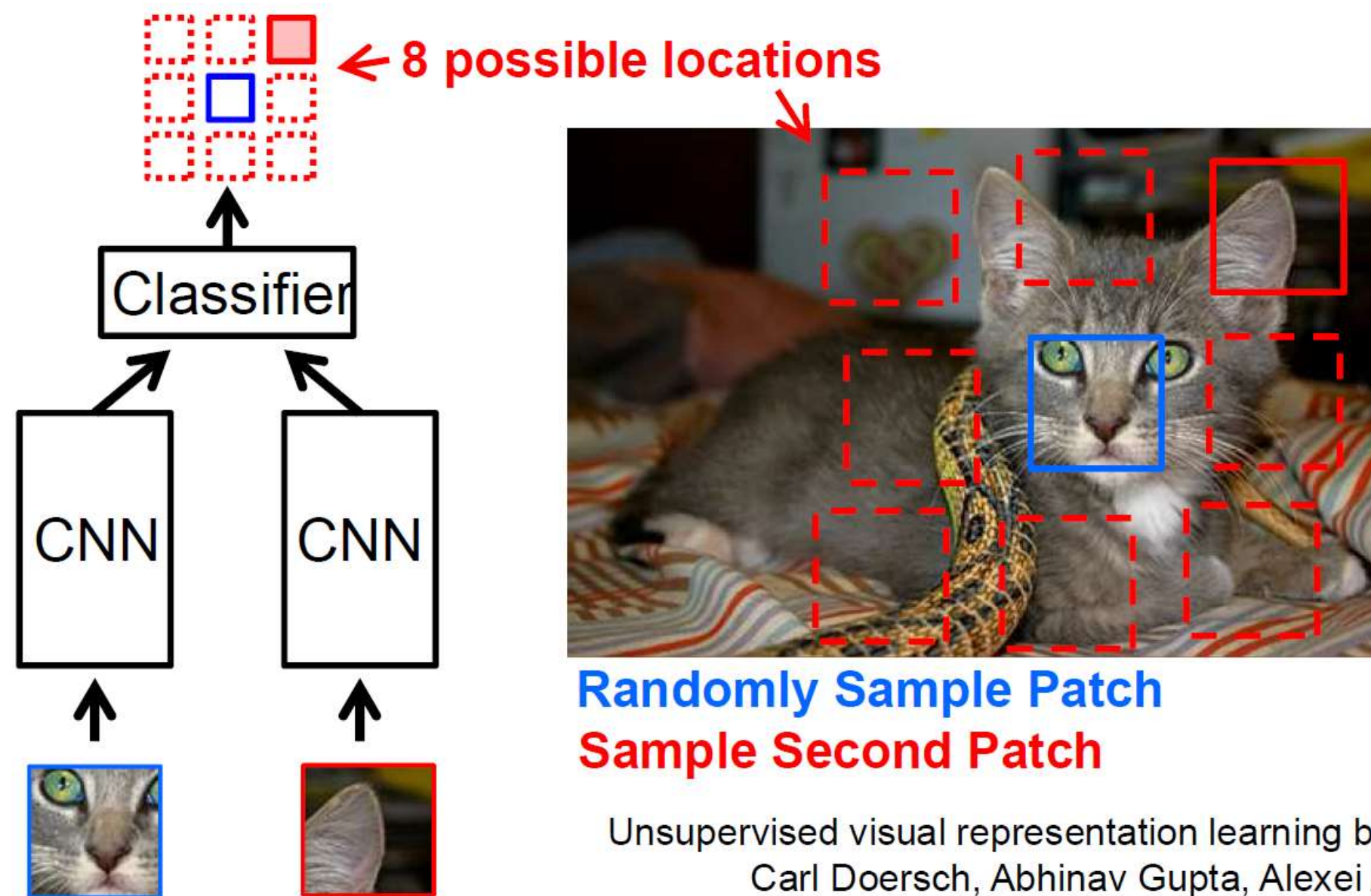
$\min_{w'}: \sum_{x'} \text{Loss}_2(g_2(f(x', w), w'))$ — обучение по своим данным



Самостоятельное обучение (self-supervised learning)

x — изображение

$z = f(x, w)$ — модель векторизации, обучается предсказывать взаимное расположение пар фрагментов одного изображения



Преимущество:
сеть выучивает векторные представления объектов без размеченной обучающей выборки

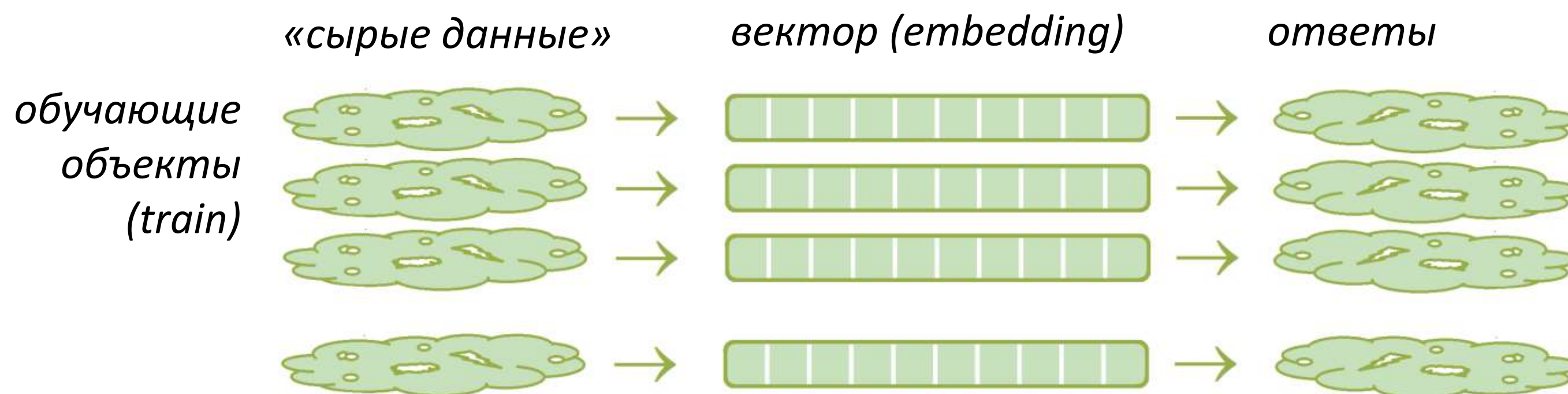
Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Нейронные сети для синтеза объектов

Вход: x — сложно структурированный объект

$z = f(x, w)$ — его векторное представление (embedding)

Выход: $y = g(z, w')$ — сложно структурированный ответ



Примеры: синтез изображений, перенос стиля, распознавание речи, машинный перевод, суммаризация текстов, ответы на вопросы

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT и др.

Эволюция подходов в обработке естественного языка

Как решали задачи анализа текстов 10 лет назад

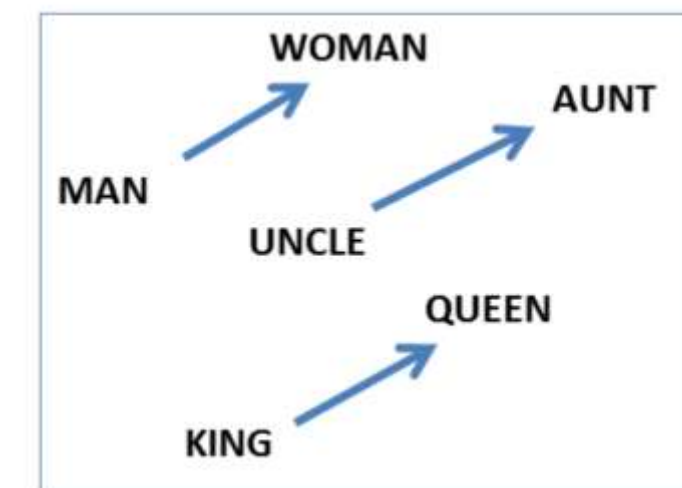
- морфологический анализ, лемматизация, опечатки, ...
- синтаксический анализ, выделение терминов, NER, ...
- семантический анализ, выделение фактов, тем, ...

Модели векторизации слов (эмбединги слов)

- модели дистрибутивной семантики:
word2vec [Mikolov, 2013], FastText [Bojanowski, 2016], ...
- тематические модели LDA [Blei, 2003], ARTM [2014], ...

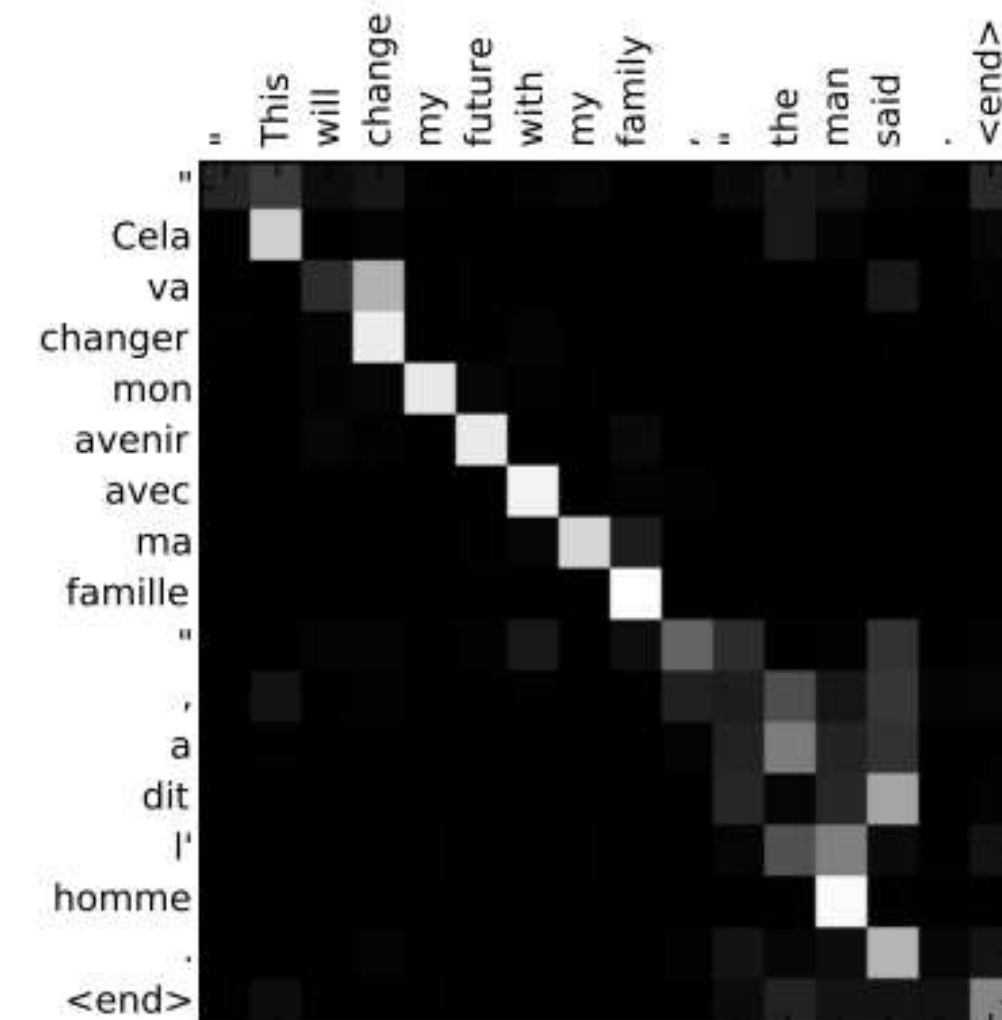
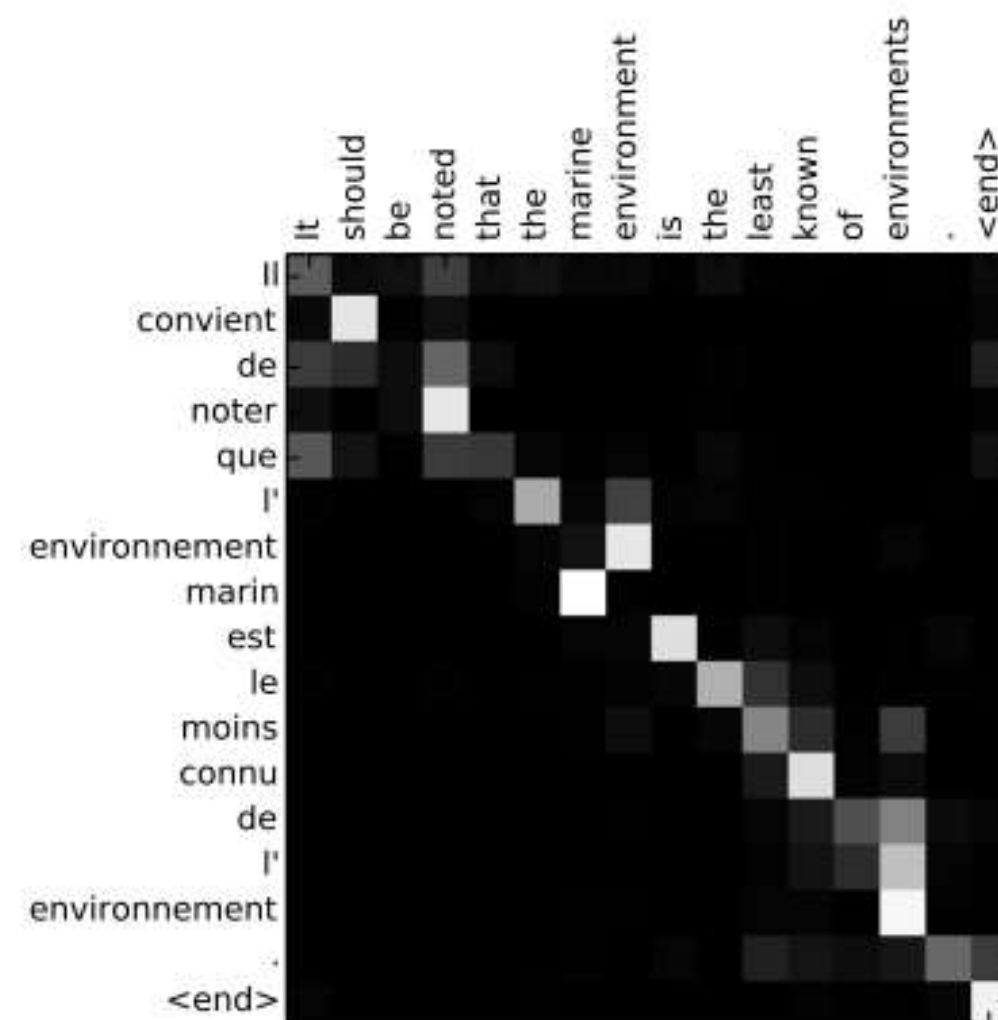
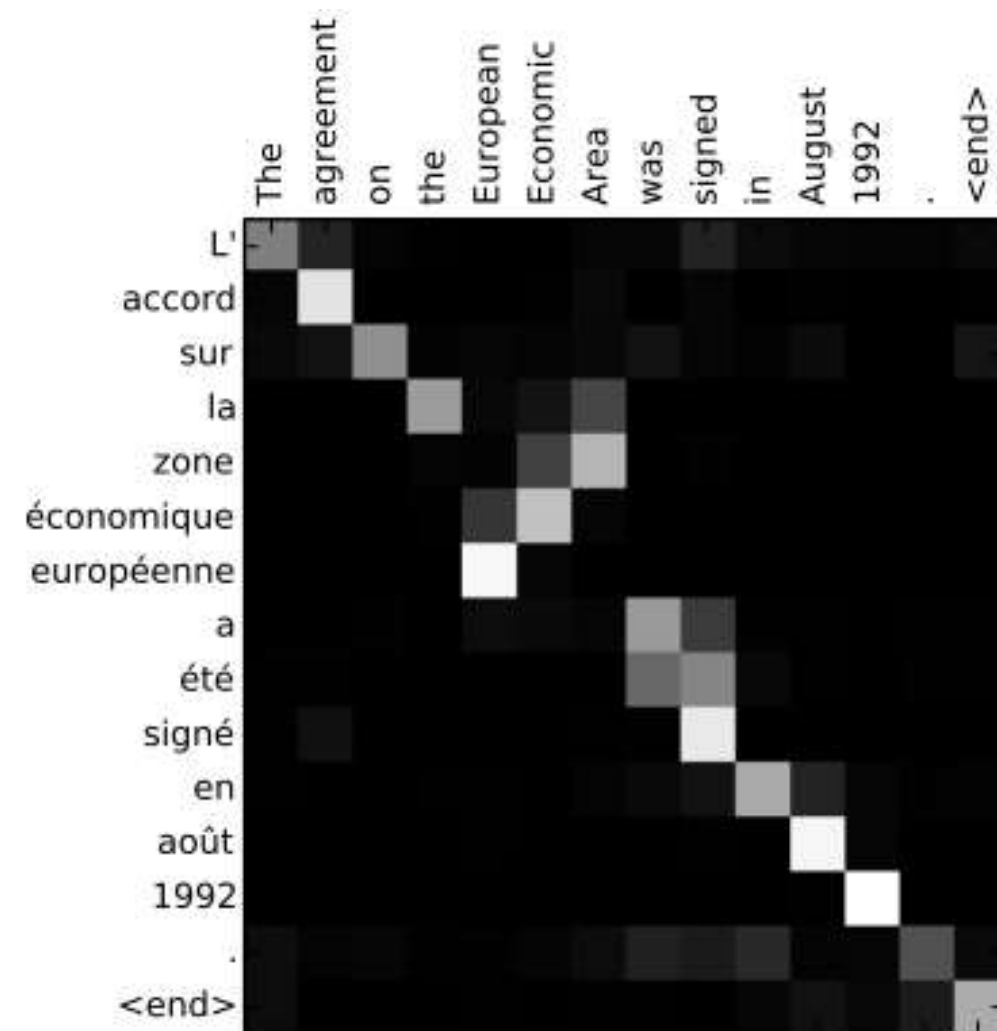
Большие модели (LLM) контекстной векторизации

- рекуррентные нейронные сети: LSTM, GRU, ...
- «end-to-end» модели внимания и трансформеры:
машинный перевод [2017], BERT [2018], GPT-4 [2023], ...



$$\text{softmax} \left(\frac{\begin{matrix} Q \\ \text{grid} \end{matrix} \times \begin{matrix} K^T \\ \text{grid} \end{matrix}}{\sqrt{d}} \right) \begin{matrix} V \\ \text{grid} \end{matrix}$$

Модели внимания: машинный перевод



Интерпретация моделей внимания: *матрица семантического сходства* $A[t,i]$ показывает, на какие слова $x[i]$ входного текста модель обращает внимание, когда генерирует слово перевода $y[t]$

Модели внимания: аннотирование изображений



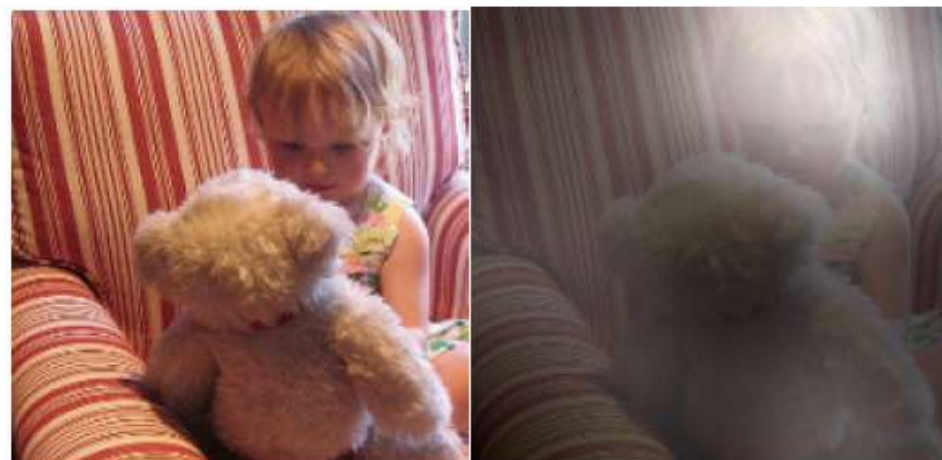
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Интерпретация: на какие области модель обращает внимание, когда генерирует подчёркнутое слово в описании изображения

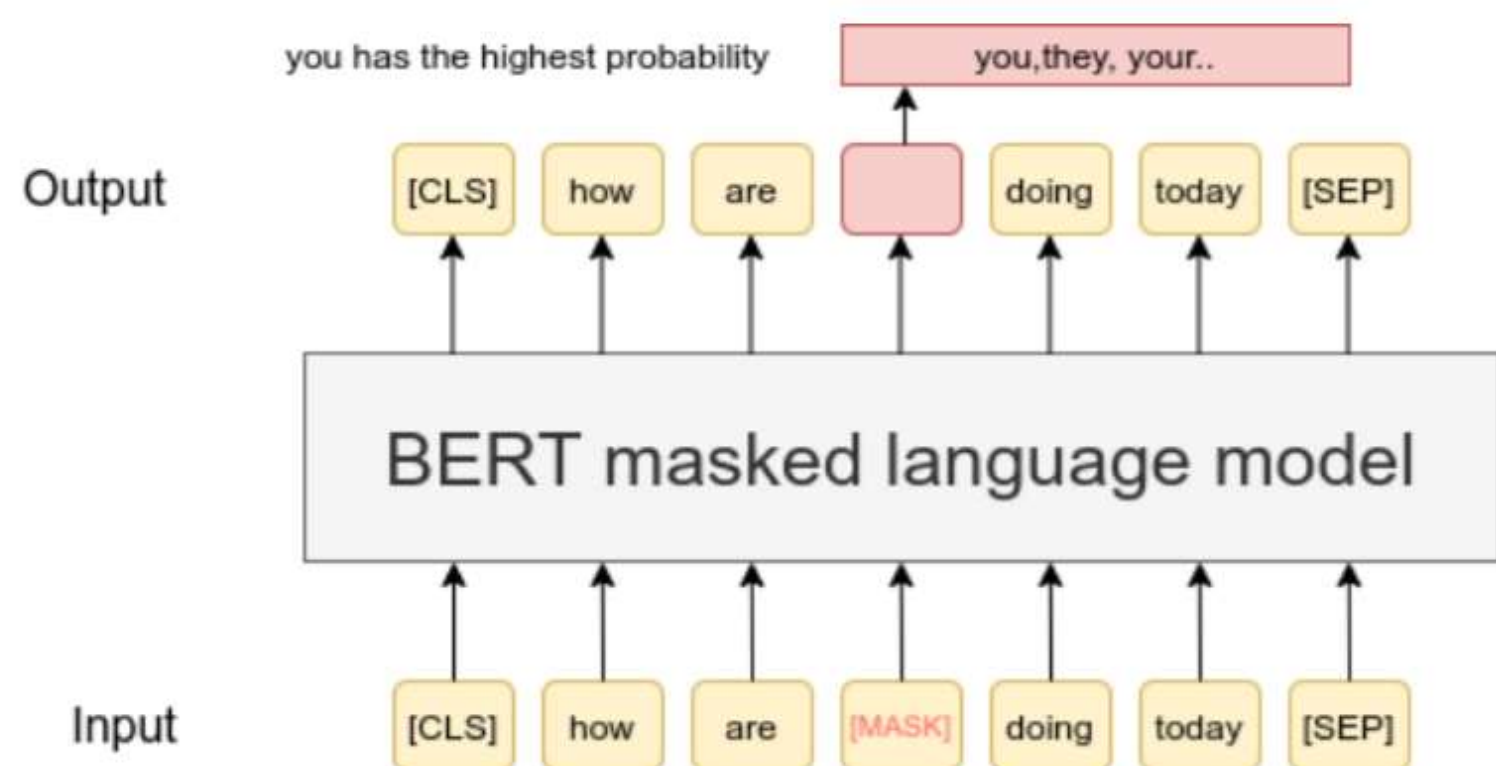
Обучение контекстной векторизации слов

x_i — слово на i -й позиции в коллекции текстовых документов

$z_i = f(x_i, C_i, w)$ — модель векторизации слова x_i по контексту C_i

$p(x|i, z, w')$ — вероятностная модель предсказания слова по вектору z

$\text{Loss}(x_i, w) = -\ln p(x_i|i, f(x_i, C_i, w), w')$ — потеря от предсказания слова на i -й позиции по его контексту (Masked Language Model)



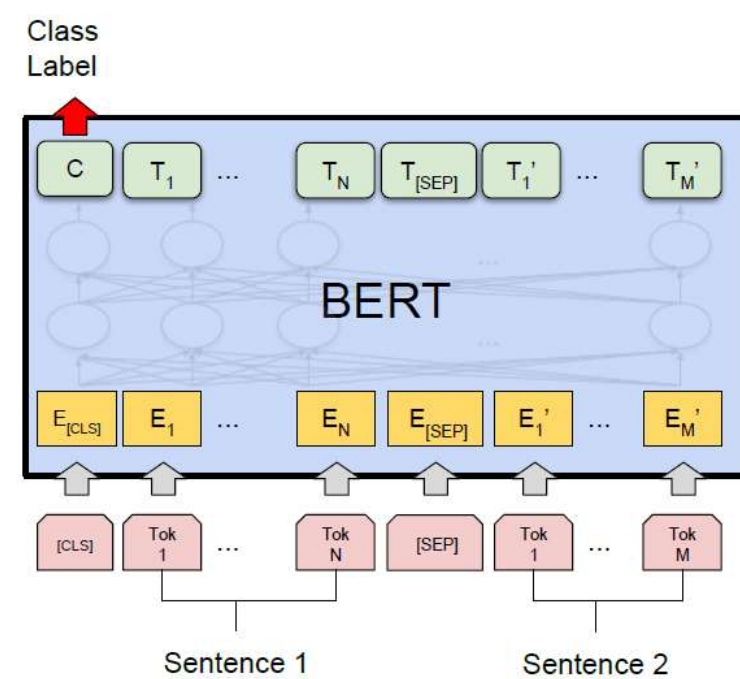
Vaswani et al. (Google) Attention is all you need. 2017.

Jacob Devlin et al. (Google AI Language)

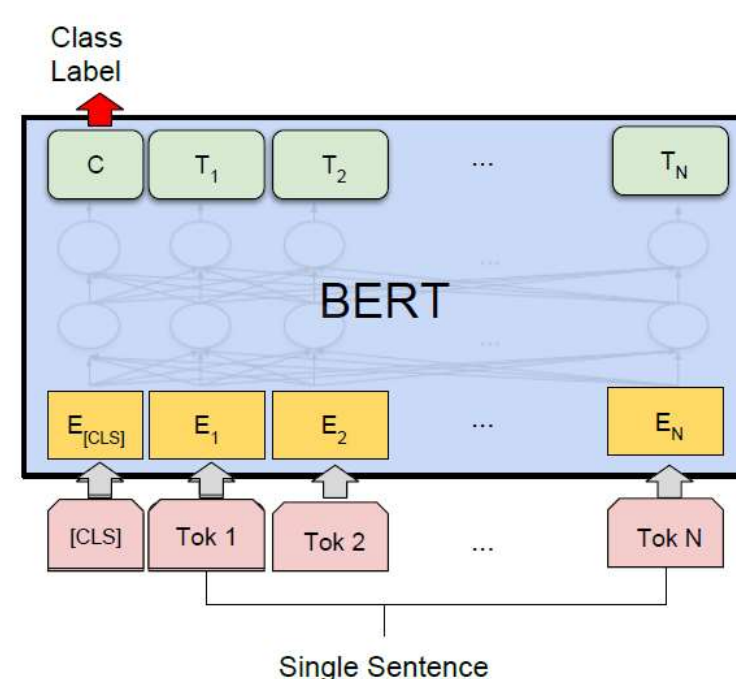
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Трансформеры: большие языковые модели (LLM)

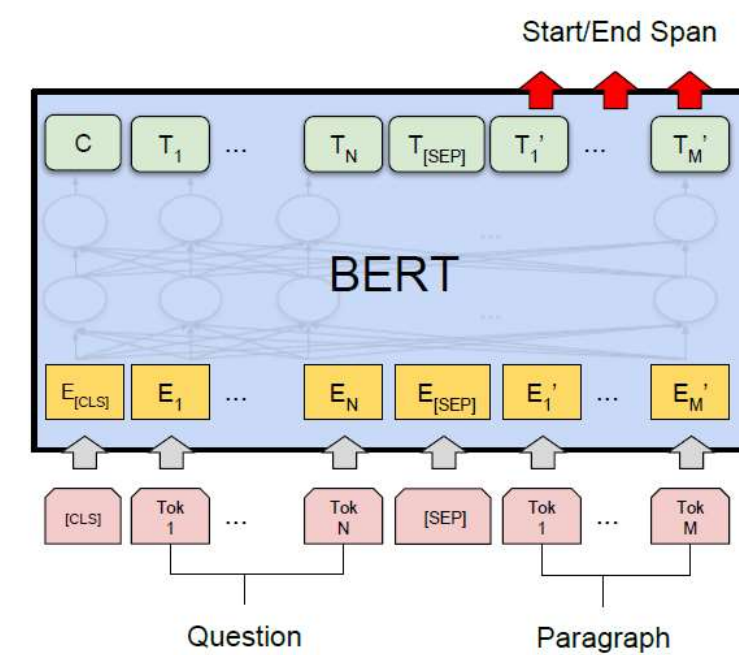
- Обучаются векторизовать и предсказывать слова по контексту
- Обучаются по терабайтам текстов, «они видели в языке всё»
- Мультиязычны: обучаются на десятках языков
- Мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



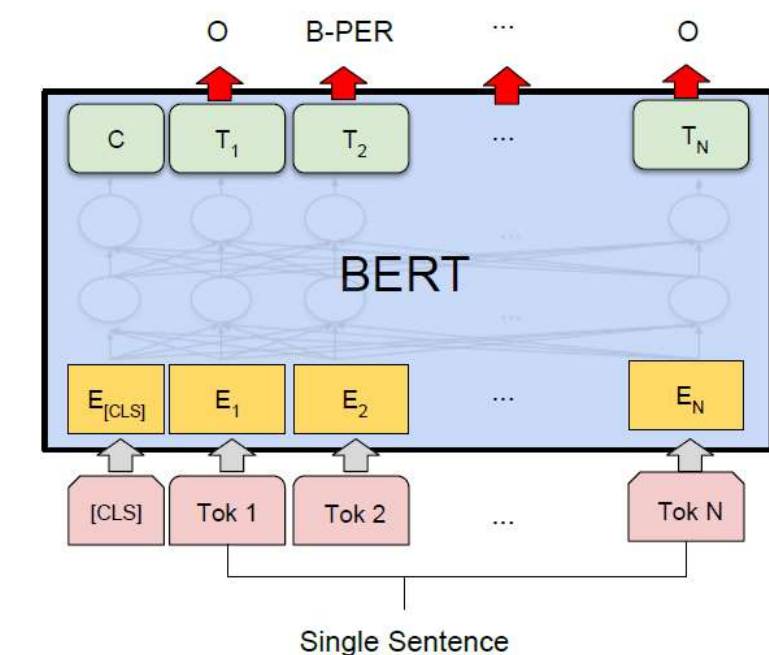
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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



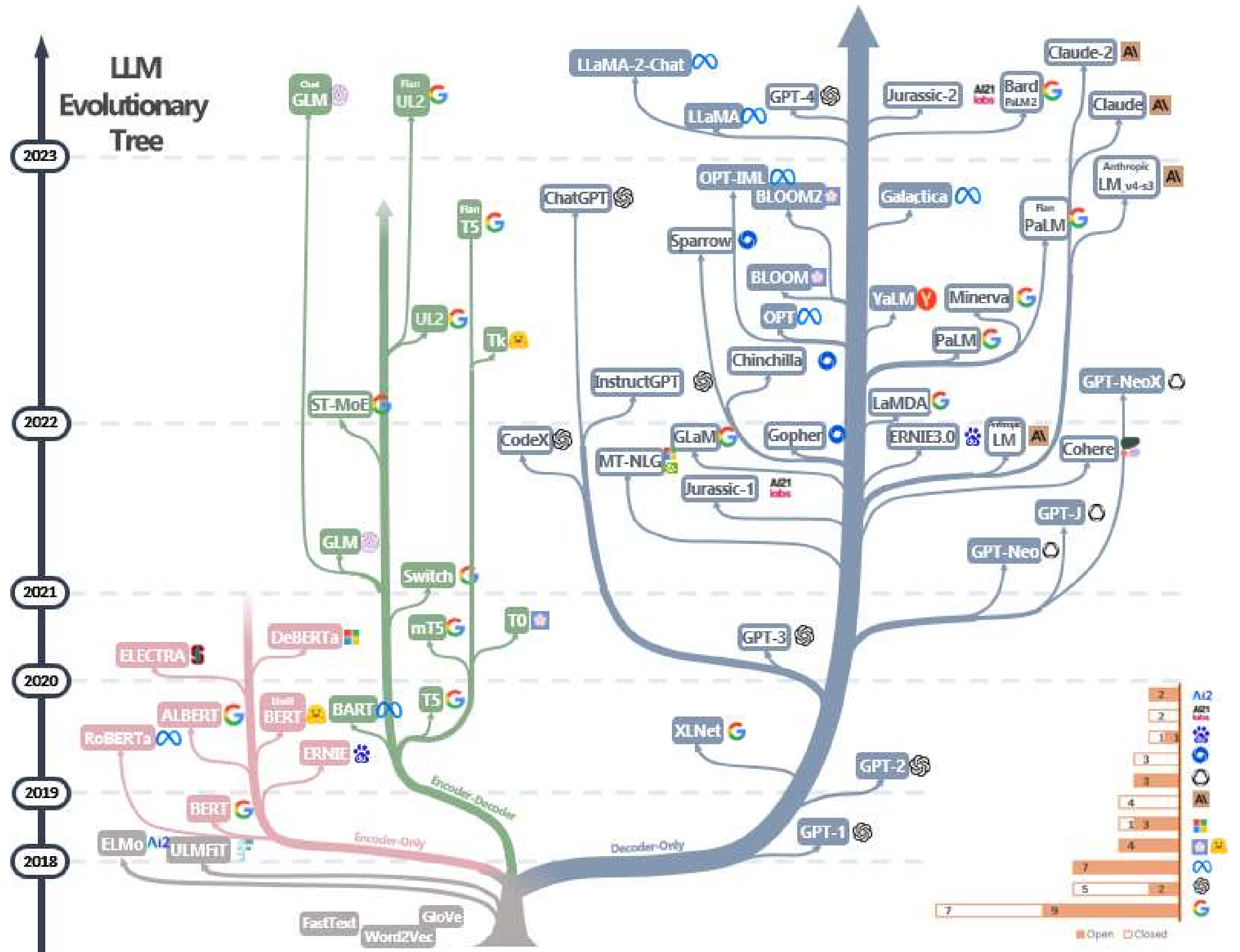
(c) Question Answering Tasks:
SQuAD v1.1



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

Трансформеры: нейросетевые модели языка

Рост числа параметров
больших языковых
моделей



Проблески общего искусственного интеллекта

Sparks of Artificial General Intelligence: Early experiments with GPT-4

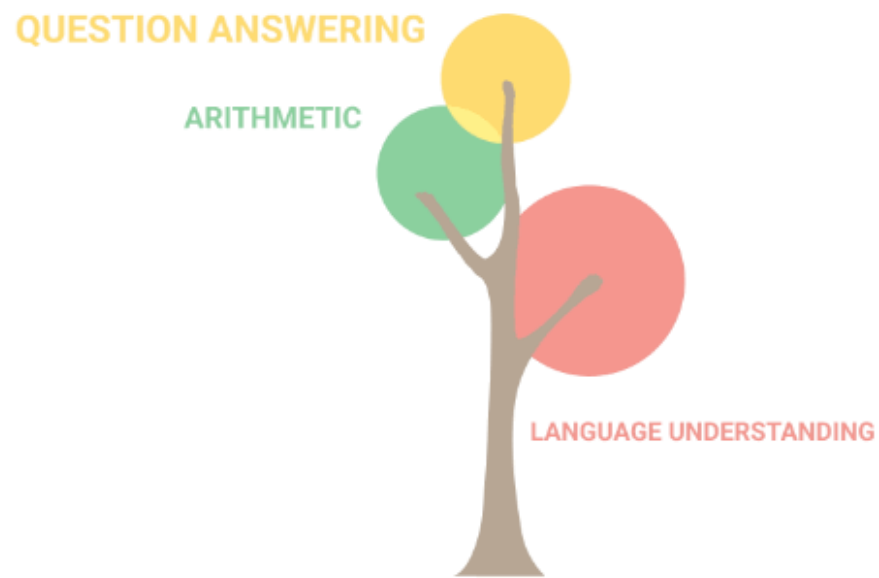
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research (27 March 2023)

Новые способности модели, не закладывавшиеся при обучении:

- объяснять свои ответы, перефразировать, переводить на другие языки
- реферировать, генерировать планы, сценарии, шаблоны
- строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

Эмерджентные (не ожидавшиеся) способности модели

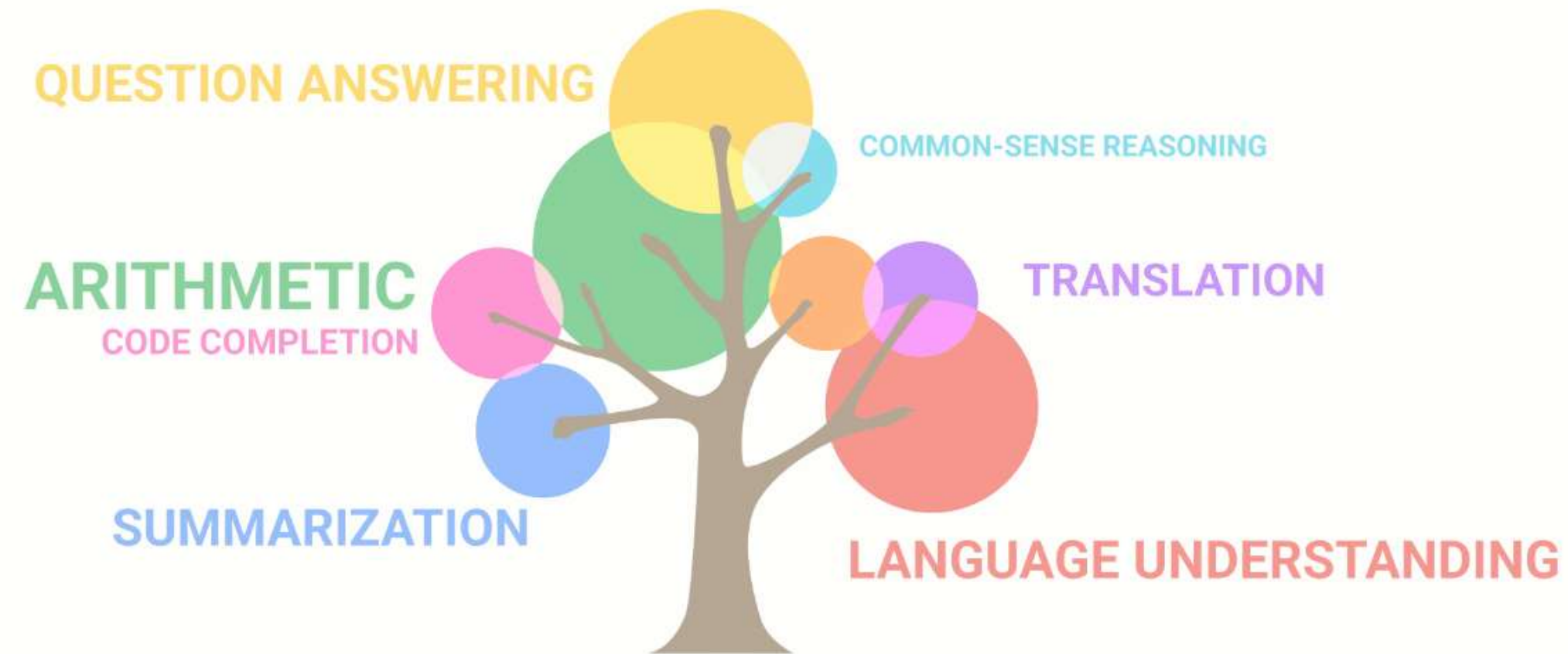


GPT-2: 14-Feb-2019

1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb), контекст 768 слов (1,5 стр.)

- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Эмерджентные (не ожидавшиеся) способности модели

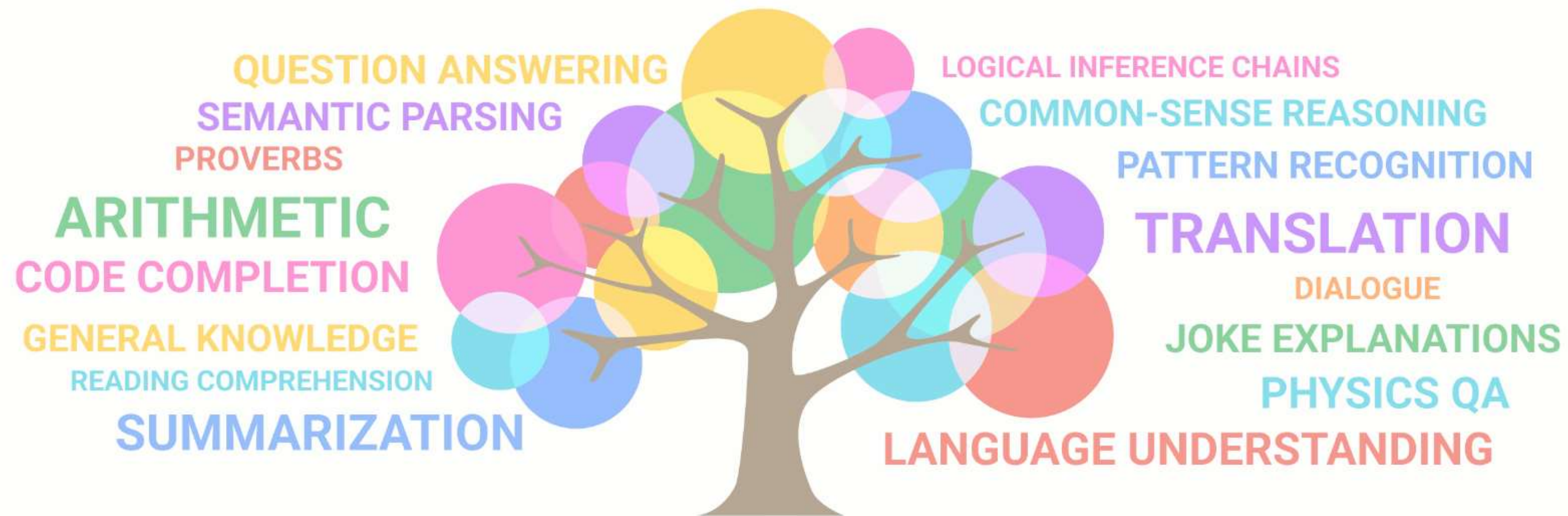


GPT-3: 11-Jun-2020

175 млрд. параметров, корпус 500 млрд. токенов, контекст 1536 слов (3 стр.)

- способность делать перевод на другие языки
- способность решать логические и простейшие математические задачи
- способность генерировать программный код по текстовому описанию

Эмерджентные (не ожидавшиеся) способности модели

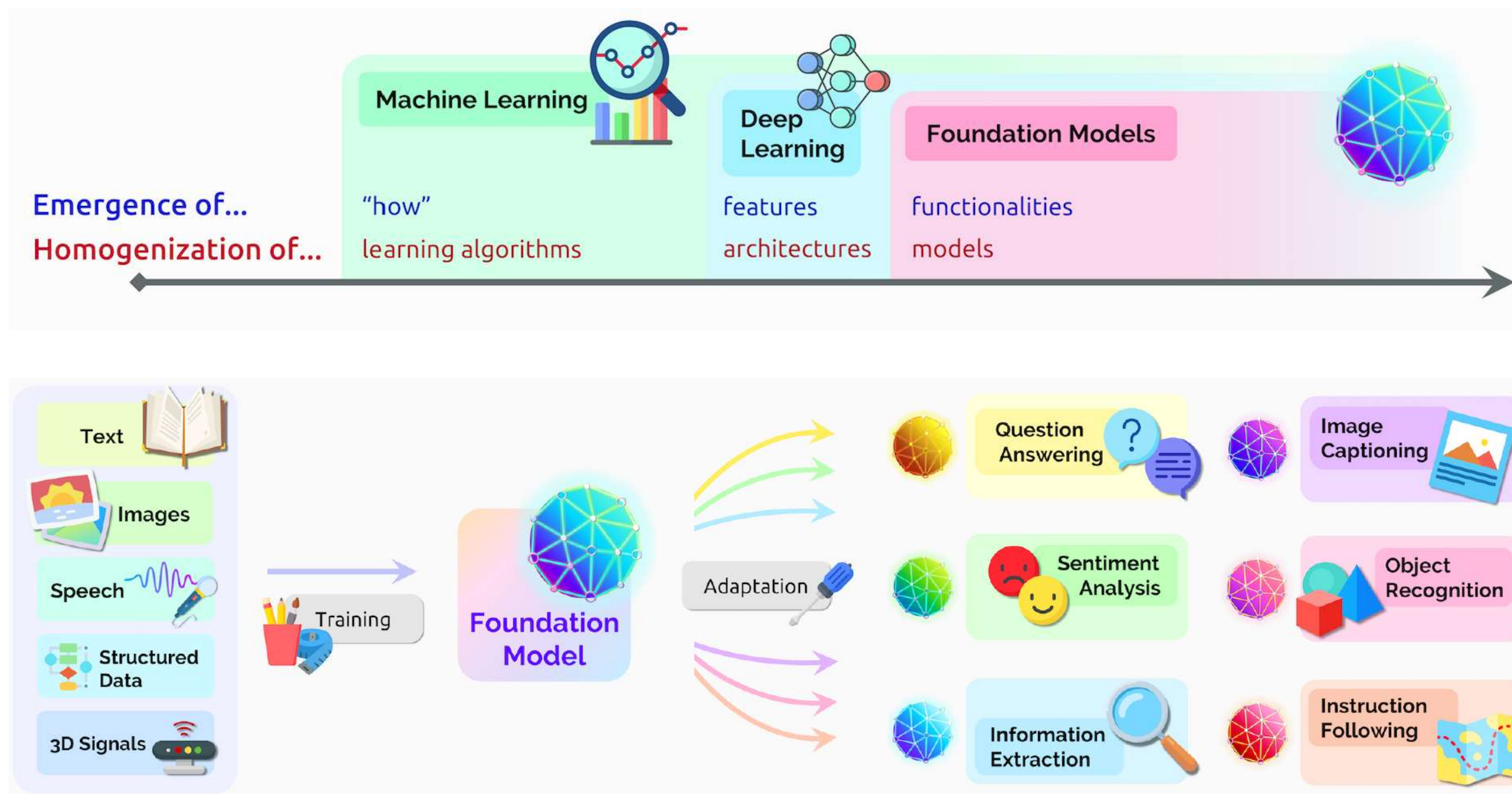


GPT-4: 14-Mar-2023

>1 трл. параметров, корпус >1Тb, контекст 24 000 слов (48 страниц)

- способность описывать и анализировать изображения
- способность реагировать на подсказки вроде «Let's think step by step»
- способность решать качественные физические задачи по картинке

Фундаментальные модели (Foundation Models)



Выводы (технологические): о составных частях успеха ИИ

- принцип эмпирической индукции Фрэнсиса Бэкона
- минимизация (и аппроксимация) эмпирического риска
- регуляризация некорректно поставленных задач
- коннекционизм и глубокие нейросетевые архитектуры
- векторизация сложно структурированных данных
- самостоятельное обучение вместо обучения по разметке
- увеличение скорости и параллелизма вычислителей

Выводы (гуманитарные): как относиться к ИИ и его развитию

- ИИ = Имитация Интеллекта, набор технологий, не объект, не субъект
- ИИ начинается с постановки задачи *Дано-Найти-Критерий*
- *Люди* ставят задачу и несут ответственность за её решение, за чистоту и достоверность данных — тоже
- *Глубокие нейронные сети* — не аналог мозга человека, а обучаемая векторизация сложно структурированных данных
- *Генеративные модели текста* — не интеллект, а новый языковой интерфейс к ~~знанию человечества~~ содержимому Интернета, с его избыточностью, неточностью, противоречивостью

Содержание

1. Технологии машинного обучения и искусственного интеллекта

- Обзор истории развития технологий ML/AI
- Обзор постановок задач машинного обучения
- Задачи анализа текстов и большие языковые модели

2. Проект «Мастерская знаний»

- Концепция мастерской знаний
- Большая языковая модель научных текстов
- Технологии анализа подборок текстовых документов

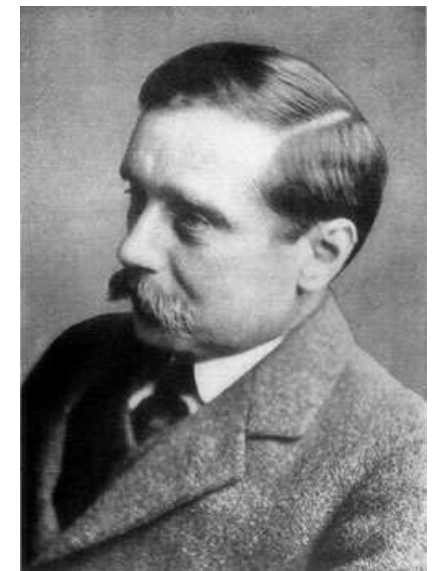
3. Языковые модели разметки текста

- Задачи автоматизации контент-анализа
- Методика оценивания моделей разметки
- Задача детекции ценностей социокультурного кода

Концепция «Мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этим знаниям, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в **своеобразной мастерской**, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи.» – Герберт Уэллс, 1940

(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)



Сегодня технологии IR/ML/NLP/NLU позволяют решать такие задачи

От поиска информации к «Мастерской знаний»

Недостатки обычного поиска:

- как искать новые знания?
- что делать с найденным?



Мастерская знаний – инструментарий для автоматизации *последующих этапов* работы с текстовыми источниками:

- ищу – чтобы накапливать
- накапливаю – чтобы анализировать
- анализирую – чтобы понимать
- понимаю – чтобы применять и передавать

Сегодня технологии IR/ML/NLP/NLU позволяют решать такие задачи

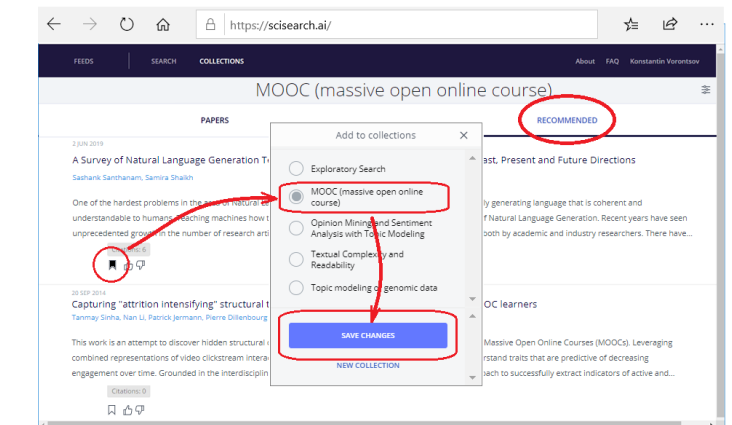
Функции «Мастерской знаний»

Подборка текстов – поисковый интерес и рабочее пространство пользователя/группы

Расширенная подборка — подборка + семантически близкие тексты

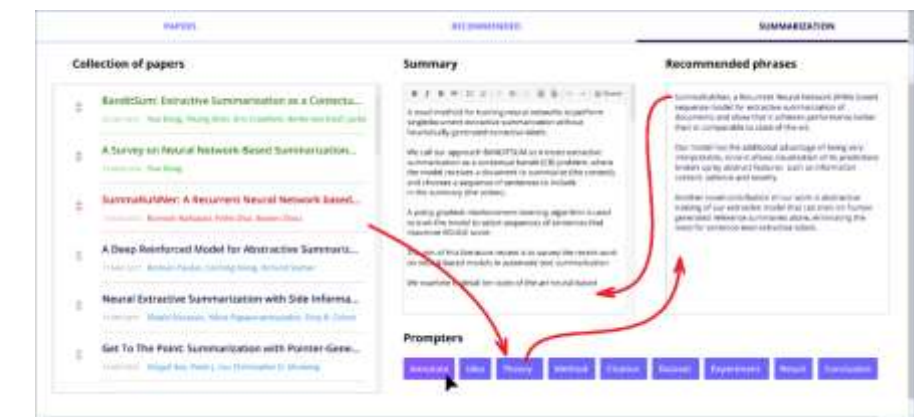
Поисково-рекомендательные сервисы:

- поиск семантически близких документов по **подборке**
- мониторинг новых документов по тематике **подборки**
- контекстные рекомендации по тексту документа из **подборки**



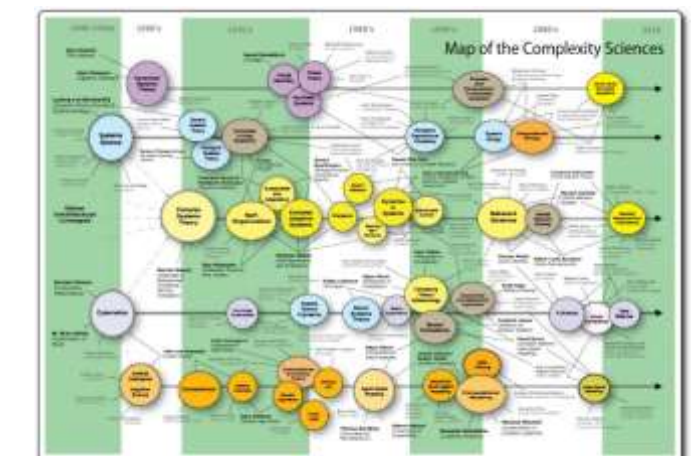
Аналитические сервисы:

- полуавтоматическая суммаризация **подборки**
- тематизация, контент-анализ, визуализация **подборки**
- выявление новых научных трендов по тематике **подборки**
- рекомендация порядка чтения документов **подборки**



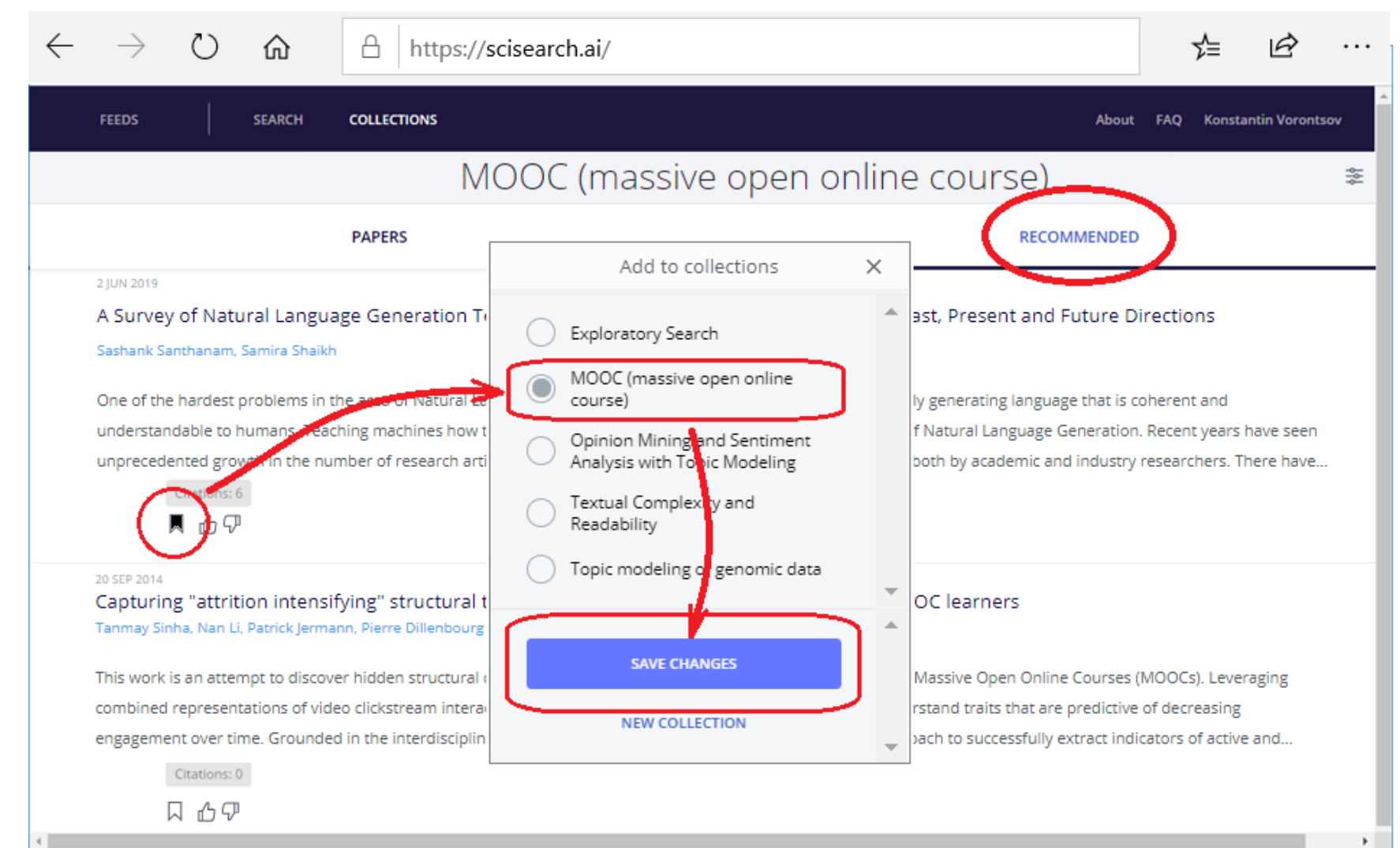
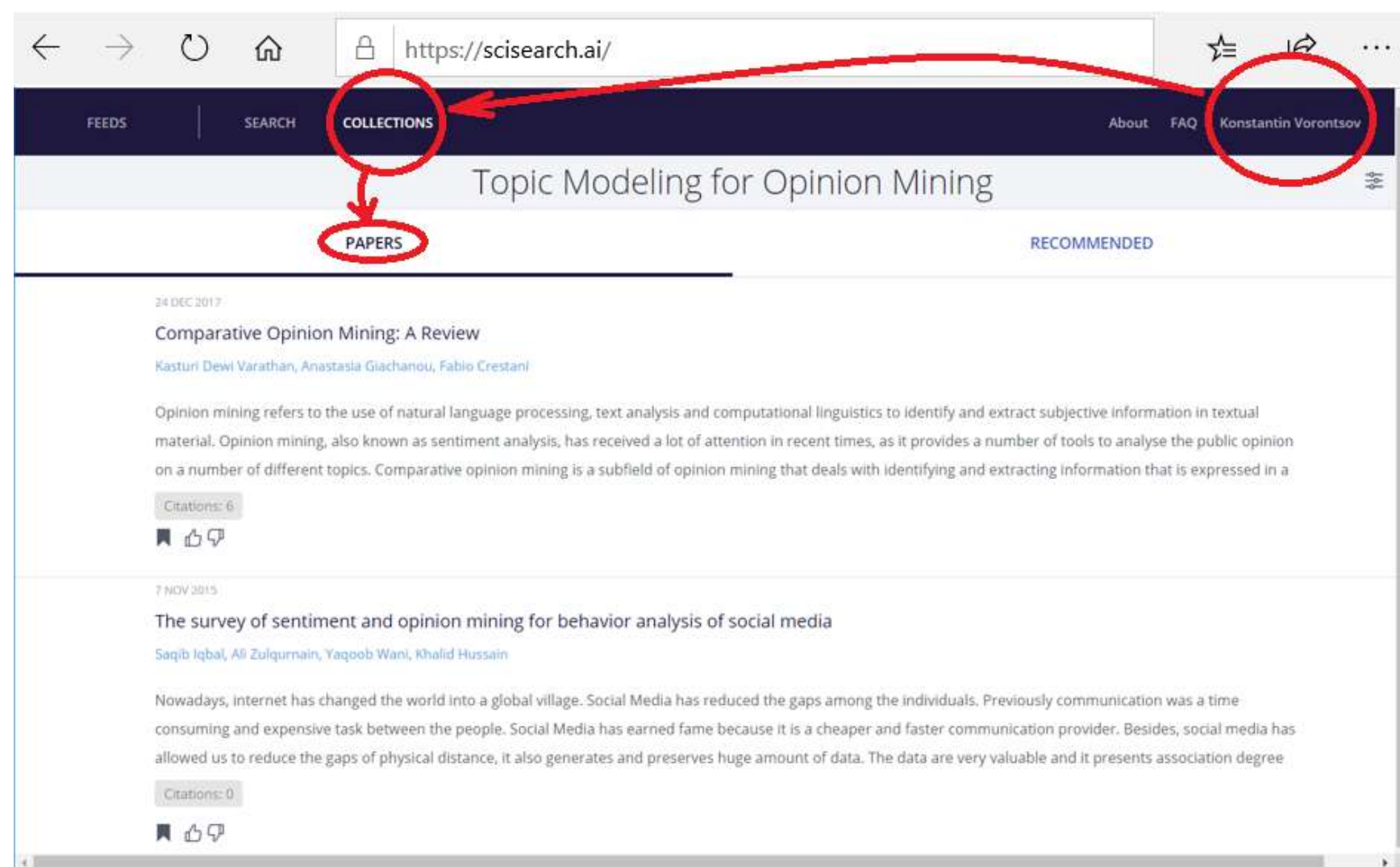
Коммуникативные сервисы, поддержка коллективного разума:

- совместное составление, обсуждение, использование **подборок**
- поддержка проектной деятельности по тематике **подборки**

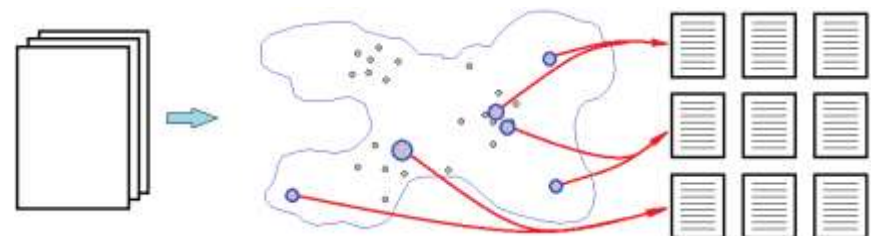
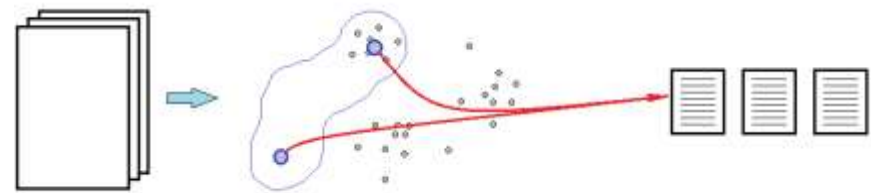
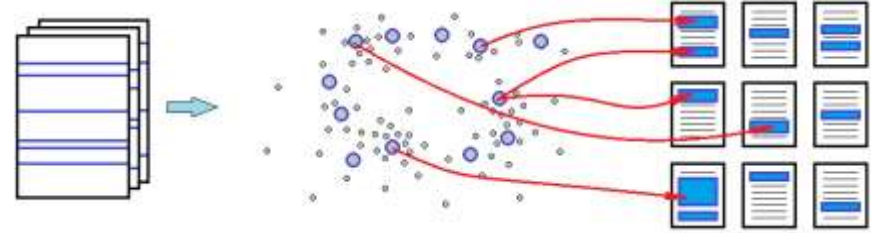
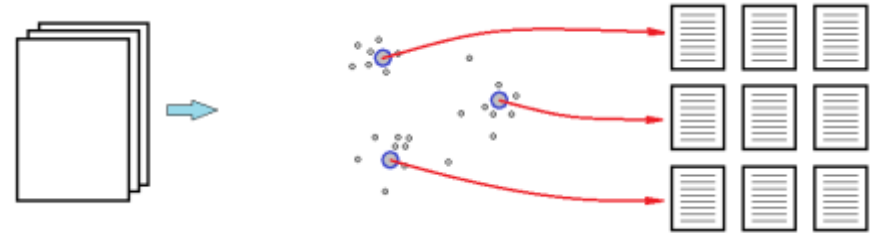
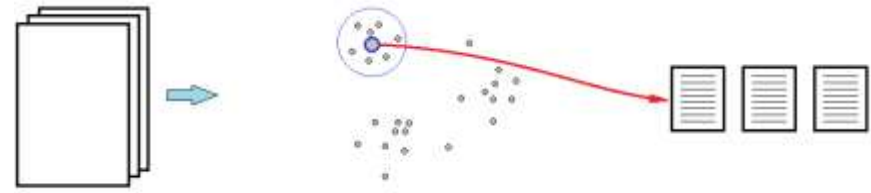
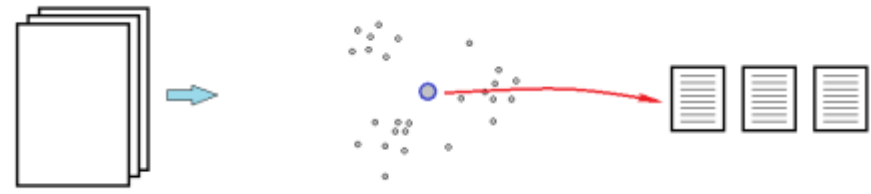


Поиск и рекомендации (прототип интерфейса)

Подборка играет роль поискового запроса и поисковой выдачи одновременно



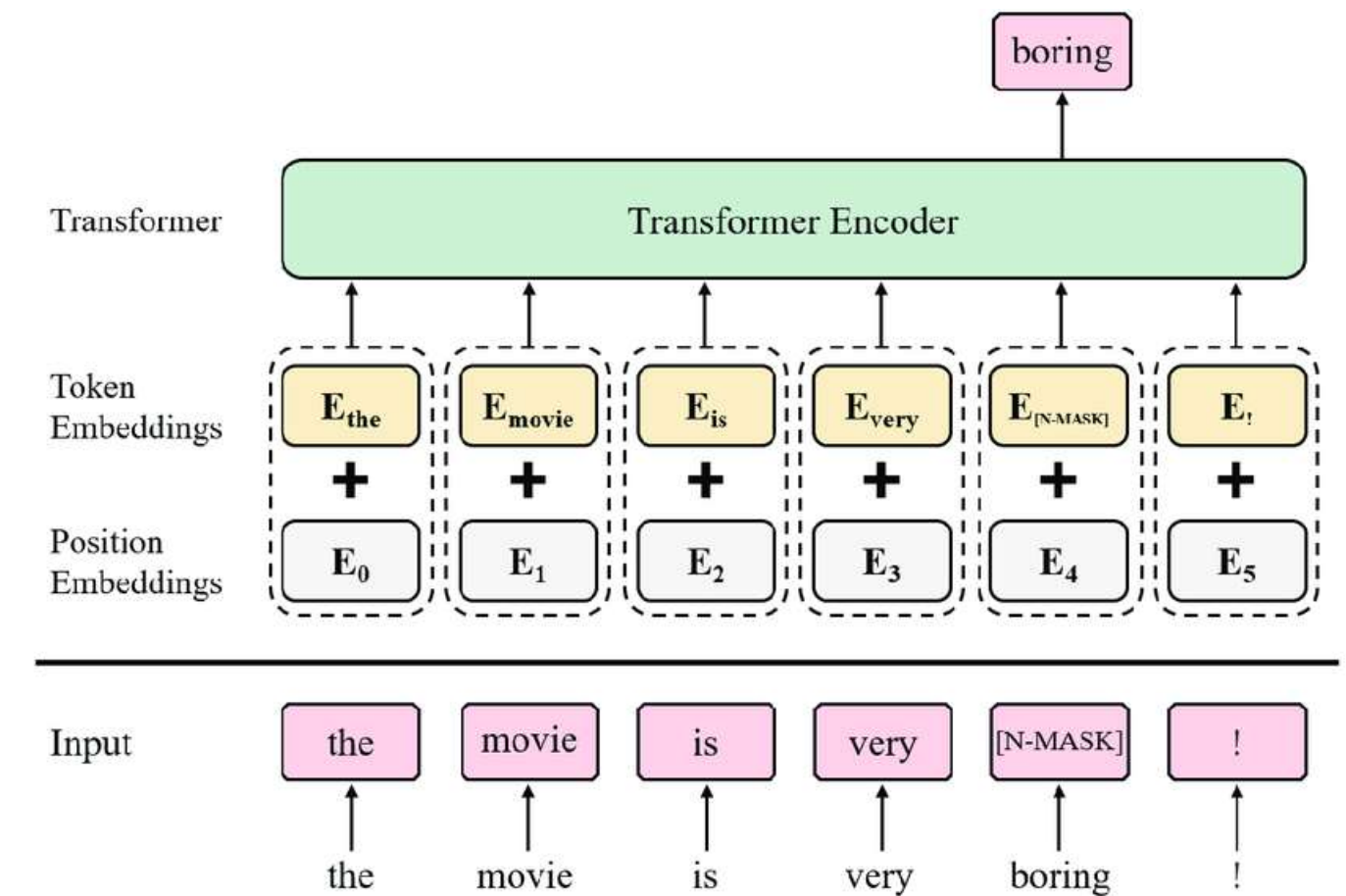
Стратегии векторного документного поиска



1. Поиск по среднему вектору **подборки** (самая простая, но не самая удачная стратегия)
2. Поиск по документу из **подборки** или нескольким близким к нему документам
3. Разбиение **подборки** на кластеры и поиск по центральным документам кластеров
4. Разбиение документов **подборки** на сегменты и поиск по сегментов документов
5. Поиск по документам смежной тематики для документа или части документов **подборки**
6. Поиск по тематике, смежной для всей **подборки**

Большие языковые модели научных текстов

- **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.



Мотивации нашего исследования

Модель должна быть применима в русскоязычных сервисах для поиска, рекомендации, классификации, анализа научных публикаций («Мастерская знаний», eLibrary.ru, научные электронные библиотеки)

Требования к модели:

- минимизация размера модели (23М параметров)
- при качестве, сопоставимом с лучшими (SOTA) моделями
- возможность вычисления эмбедингов без GPU
- мультиязычность: английский, русский, и др.
- возможность дообучения модели по данным о цитировании
- оценивание качества — по стандартным + новым benchmark-ам

Данные для обучения модели научных текстов

Данные для обучения:

- **S2ORC — Semantic Scholar Open Research Corpus**
205М публикаций, 121М авторов
30М (12В токенов) отобрано для обучения модели,
title+abstract, 85% на английском, 2% на русском
- **eLibrary**, заголовки и аннотации (title+abstract):
8.6М (2В токенов) на русском
8.8М (1.2В токенов) на английском

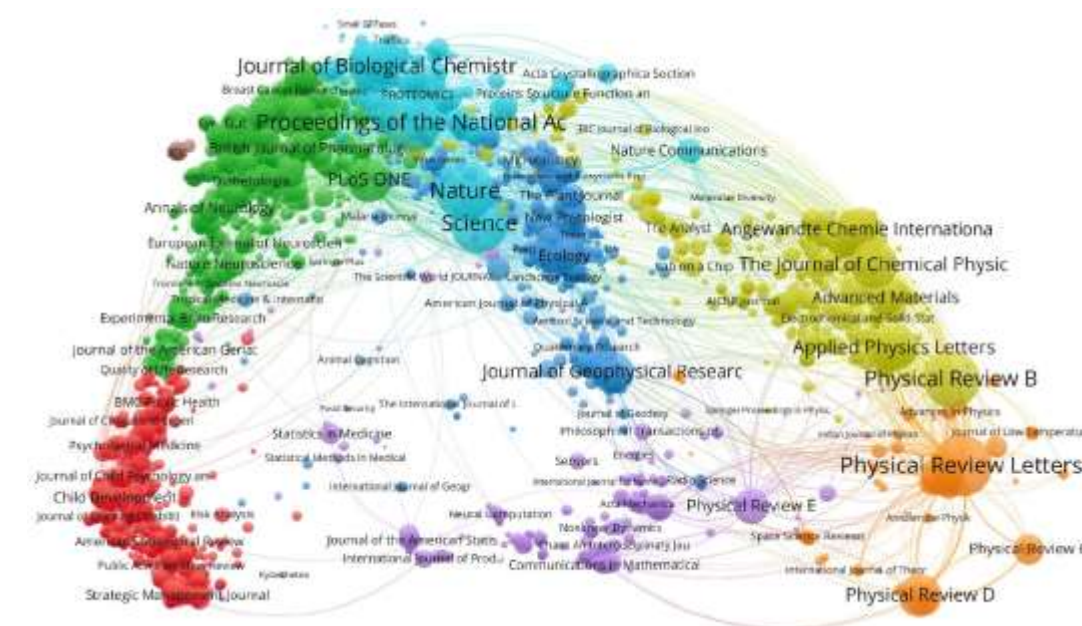


SEMANTIC SCHOLAR

eLIBRARY.RU

Данные для дообучения:

- **S2AG — Semantic Scholar Academic Graph**
источники: Crossref, PubMed, Unpaywall и др.
2.5В связей цитирования



Методики оценивания моделей (benchmarks)

SciDocs: 6 задач

- классификация статей по MeSH / по тематике
- предсказание цитирования / со-цитирования
- предсказание пользовательской активности, рекомендации статей

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

- классификация, регрессия, сходство, поиск,
- подбор рецензента для статьи, разрешение неоднозначности авторов

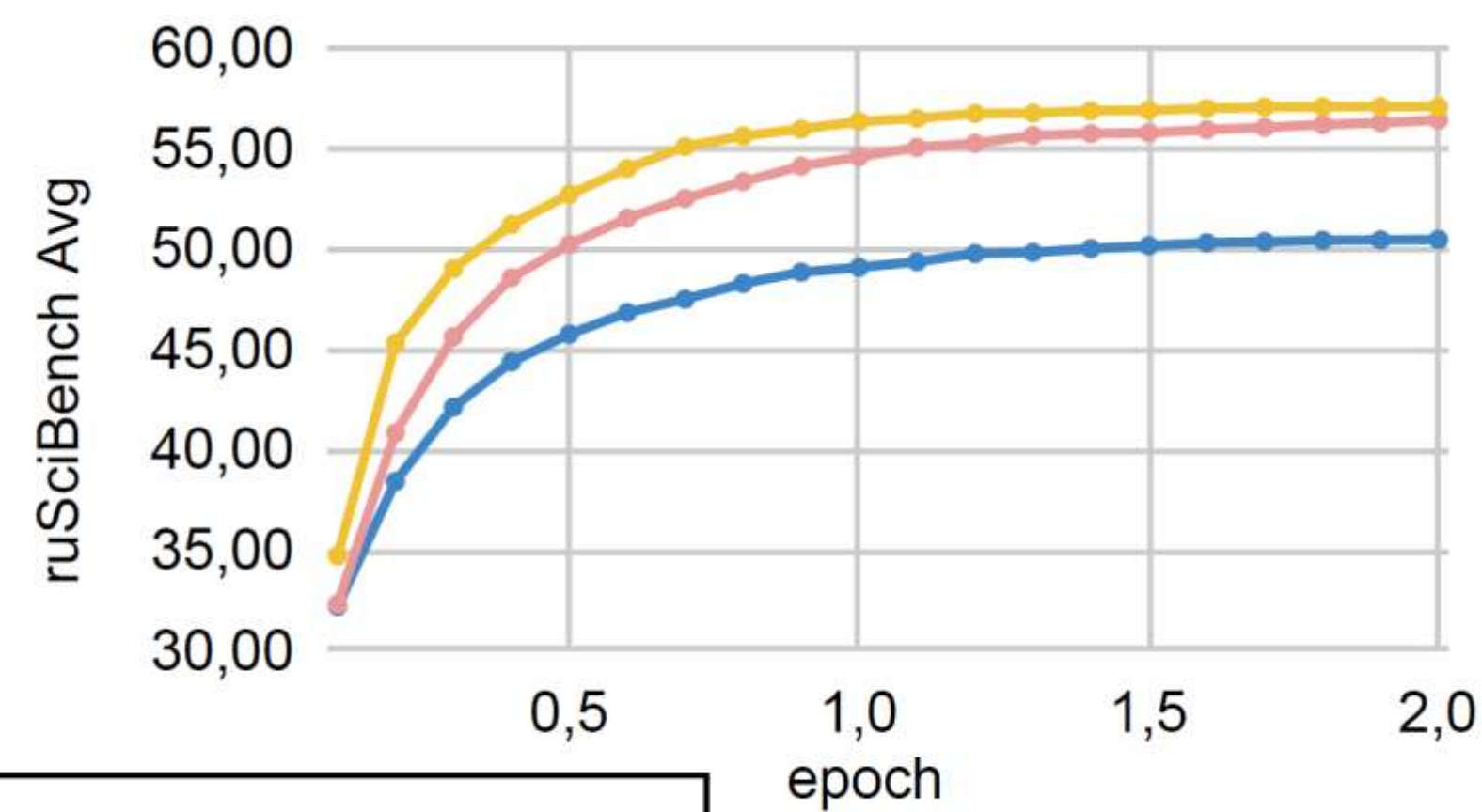
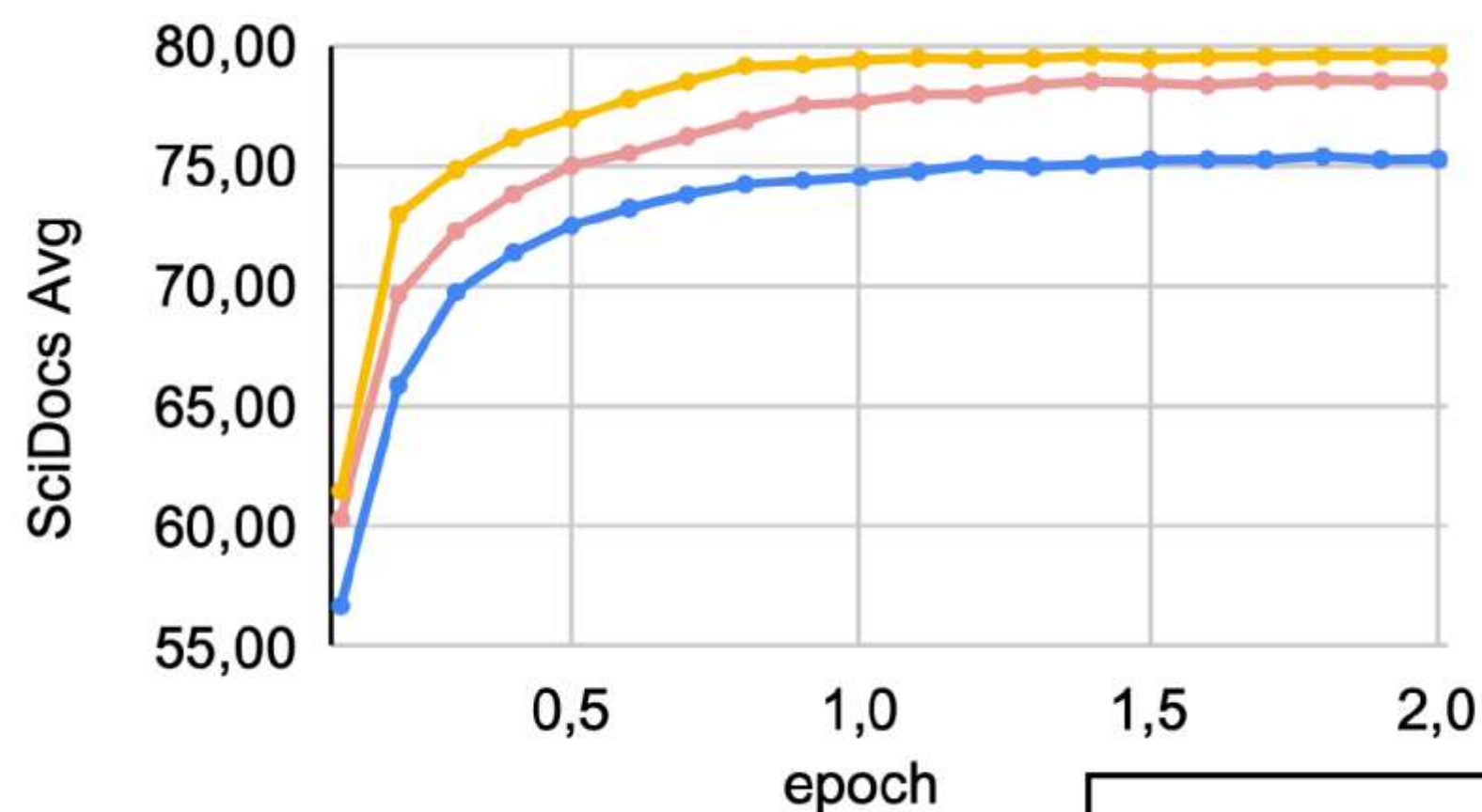
RuSciBench: 8 задач

- классификация OECD/ГРНТИ по аннотации ru / en / ru+en
- поиск аннотации по её переводу ru→en / en→ru

Этап 1: предобучение модели SciRus-tiny (MSU)

Архитектура RoBERTa (Y.Liu et al., 2019), случайная инициализация:
tiny (sz=23M, dim=312), **small** (sz=61M, dim=768), **base** (sz=85M, dim=1024)

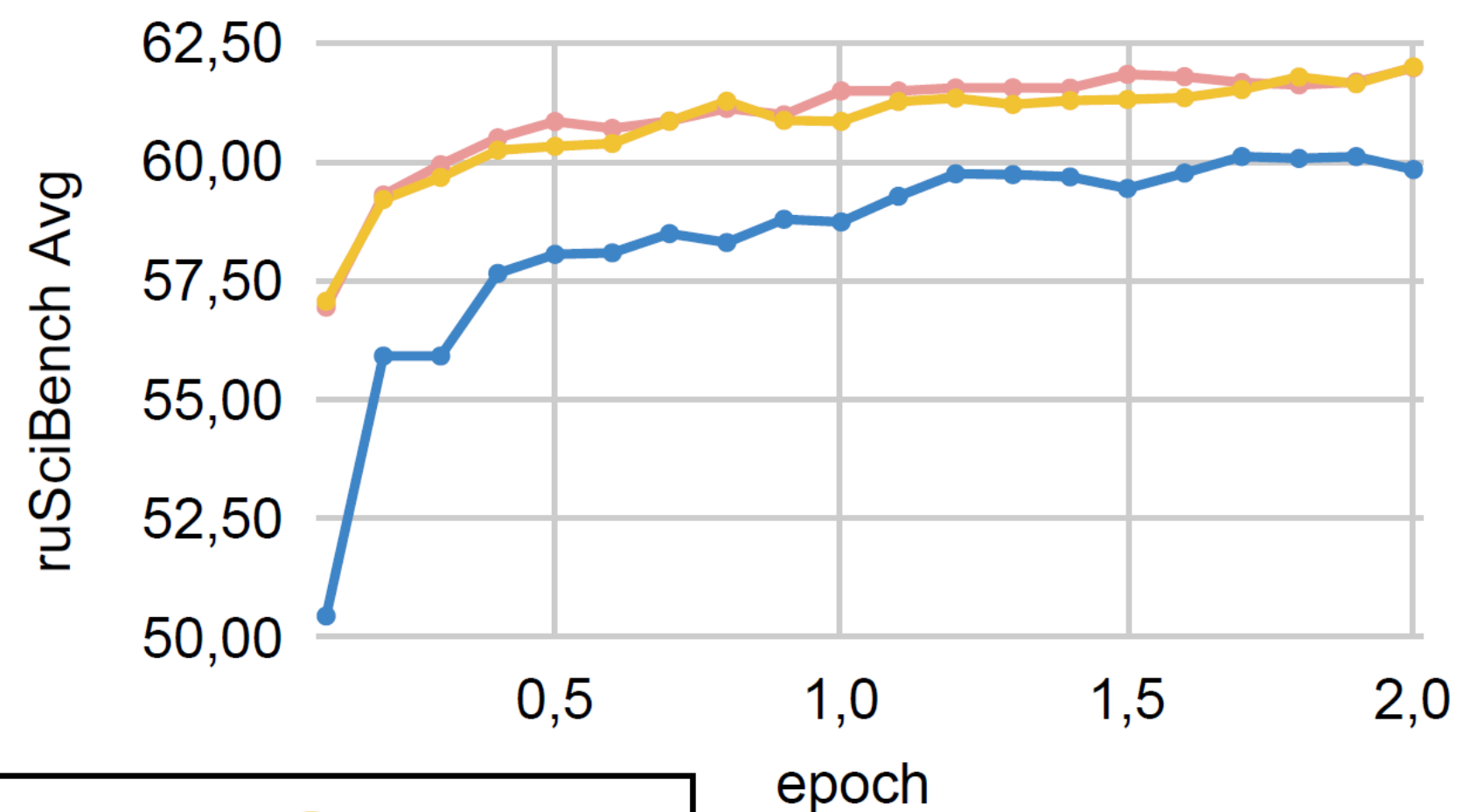
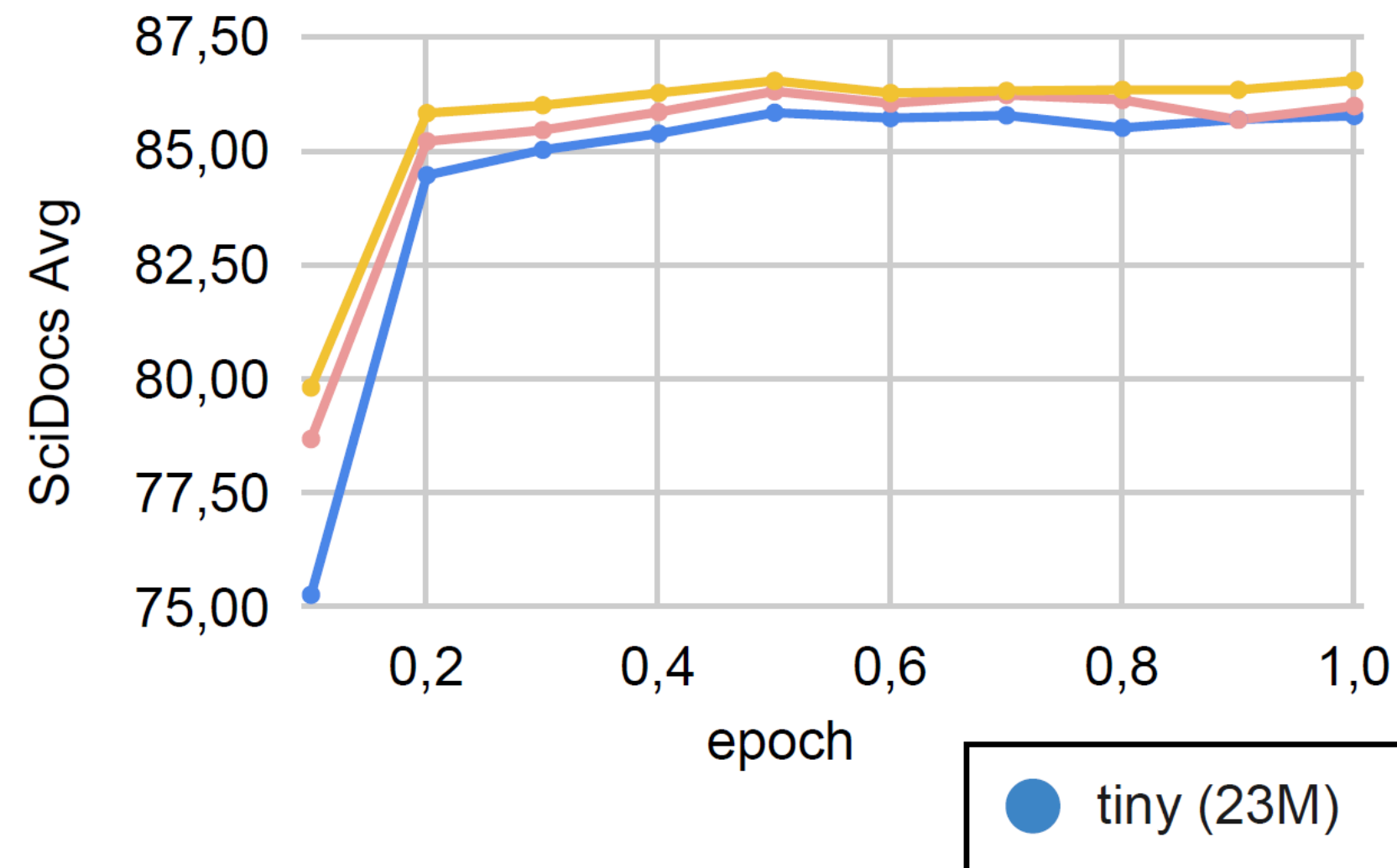
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



Этап 2: дообучение на парах title-abstract

Критерий: сблизать эмбединги в контрастных парах название/аннотация, ru/en

- 30.6M пар из S2AG
- 17.8M пар из eLibrary



Этап 3: дообучение на парах cite-cocite

Критерий: сблизать эмбединги пары документов (А,В) при цитировании:

«cite» — статья А цитирует статью В

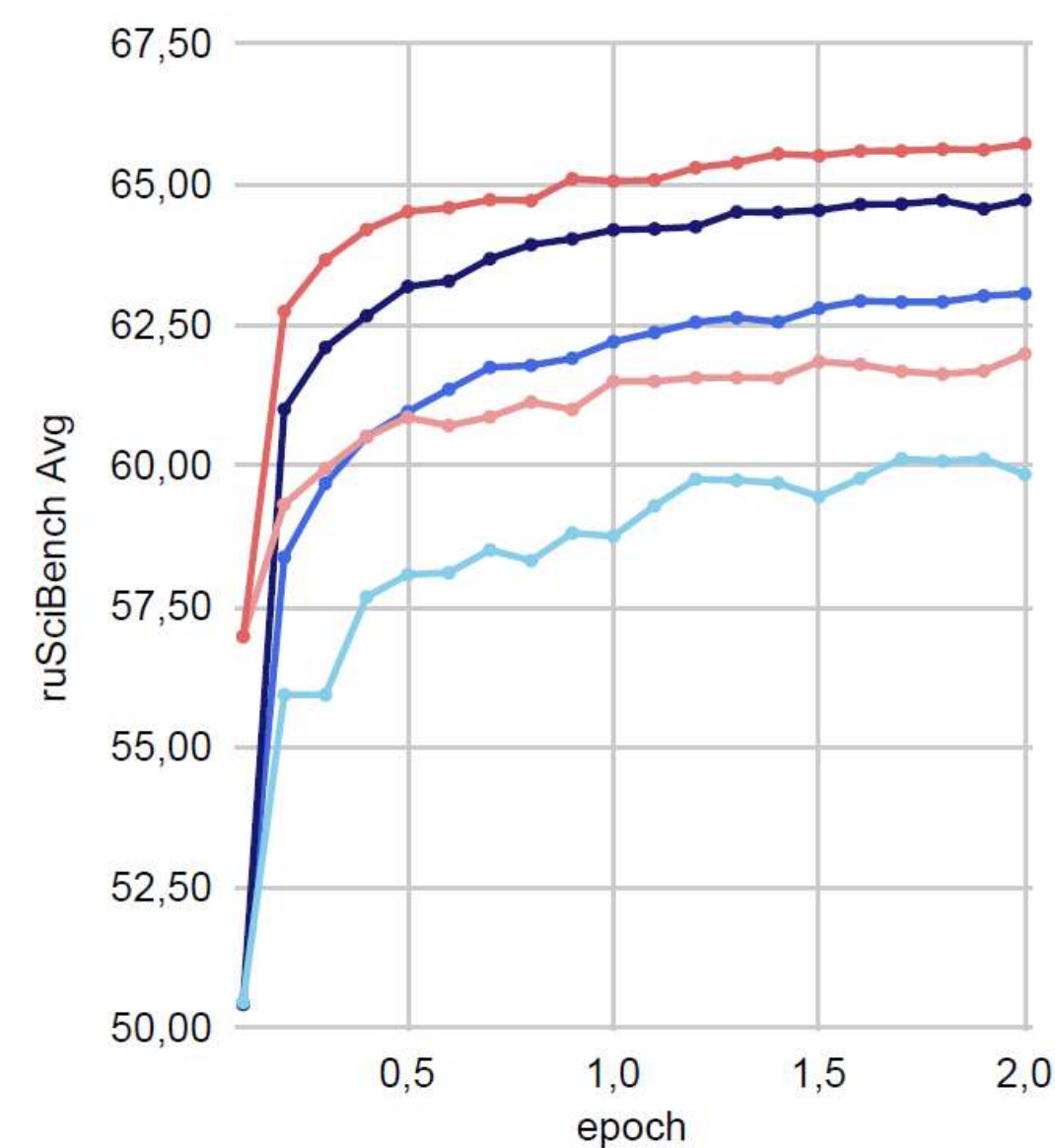
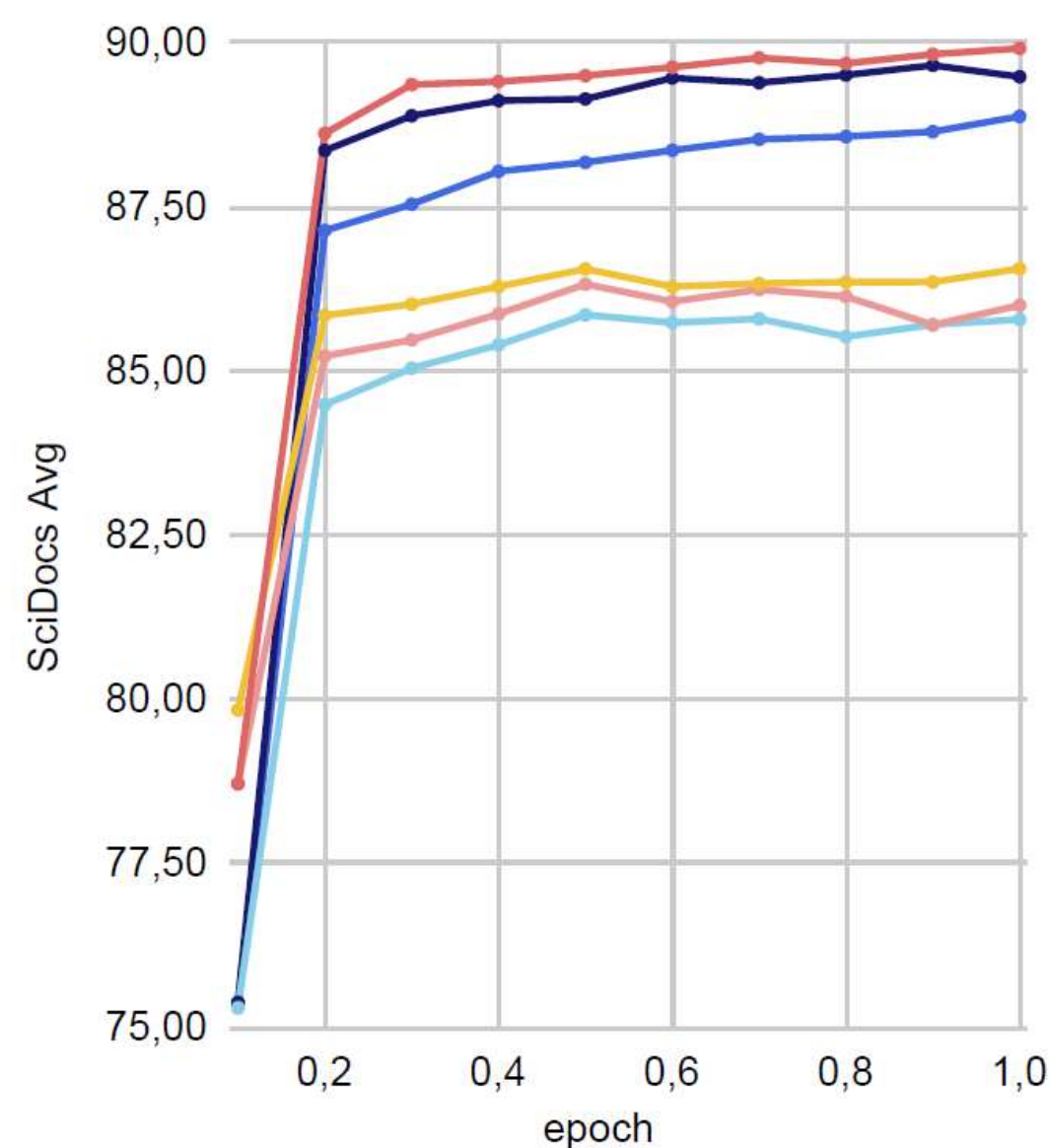
«co-cite» — третья статья С цитирует статьи А и В

S2AG:

- 13.3М пар cite
- 62М пар co-cite

eLibrary:

- 40М пар cite
- 33.7М пар co-cite



Сравнение моделей по метрикам SciDocs

 **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 <small>43</small>	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

В среднем качество лучше, чем у моделей, которые в 5, 20 и даже 200 раз больше

Сравнение моделей по метрикам ruSciBench

 **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

Качество кросс-языкового поиска близко к моделям, которые в 20 раз больше

Выводы по результатам сравнения моделей

1. Размер и качество модели в сравнении с SciNCL

- меньше параметров: 23М против 110М
- меньше размерность: 312 против 768
- больше контекст: 1024 против 512
- сопоставимое качество (SciDocs Avg): 90.10 против 90.84

2. Контрастивное дообучение на парах title-abstract

- существенно улучшает метрики качества,
- особенно качество кросс-языкового поиска

3. Контрастивное дообучение на парах cite / cocite

- компенсирует недостаточность кросс-языковых данных

Первое внедрение



«Разработанная в рамках данного проекта модель уже широко используется в **Научной электронной библиотеке** для решения целого ряда задач, связанных с оценкой тематической близости научных документов. Уже протестирован специалистами полезный сервис для ученых, позволяющий для заданной статьи или подборки статей найти тематически похожие документы, как среди всего массива [eLIBRARY.RU](https://elibrary.ru) (более 55 млн. научных публикаций), так и только среди новых поступлений. Важной для нас особенностью данной модели является ее мультязычность, поскольку **Научная электронная библиотека** содержит документы на различных языках.»

— *Геннадий Еременко, генеральный директор НЭБ*

Научная электронная библиотека, портал eLIBRARY.RU. Пресс-релиз 24-04-2024: «Открыт поиск близких по тематике публикаций с применением нейросети МГУ для анализа научных текстов.»

https://elibrary.ru/projects/news/search_similar_publ.asp

Полуавтоматическое реферирование подборки

PAPERS

Collection of papers

- BanditSum: Extractive Summarization as a Contextu...**
25 SEP 2018 Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, Jacki...
- A Survey on Neural Network-Based Summarization...**
19 MAR 2018 Yue Dong
- SummaRuNNer: A Recurrent Neural Network based...**
13 NOV 2016 Ramesh Nallapati, Feifei Zhai, Bowen Zhou
- A Deep Reinforced Model for Abstractive Summariz...**
11 MAY 2017 Romain Paulus, Caiming Xiong, Richard Socher
- Neural Extractive Summarization with Side Informa...**
14 APR 2017 Shashi Narayan, Nikos Papasarakantopoulos, Shay B. Cohen
- Get To The Point: Summarization with Pointer-Gener...**
14 APR 2017 Abigail See, Peter J. Liu, Christopher D. Manning

RECOMMENDED

Summary

B I S [Rich text editor toolbar]

A novel method for training neural networks to perform singledocument extractive summarization without heuristically-generated extractive labels.

We call our approach BANDITSUM as it treats extractive summarization as a contextual bandit (CB) problem, where the model receives a document to summarize (the context), and chooses a sequence of sentences to include in the summary (the action).

A policy gradient reinforcement learning algorithm is used to train the model to select sequences of sentences that maximize ROUGE score.

The aim of this literature review is to survey the recent work on neural-based models in automatic text summarization.

We examine in detail ten state-of-the-art neural-based

Promoters

Annotate Idea Theory Method Citation Dataset Experiment Result Conclusion

SUMMARIZATION

Recommended phrases

SummaRuNNer, a Recurrent Neural Network (RNN) based sequence model for extractive summarization of documents and show that it achieves performance better than or comparable to state-of-the-art.

Our model has the additional advantage of being very interpretable, since it allows visualization of its predictions broken up by abstract features such as information content, salience and novelty.

Another novel contribution of our work is abstractive training of our extractive model that can train on human generated reference summaries alone, eliminating the need for sentence-level extractive labels.

А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

С. Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

Полуавтоматическое реферирование подборки

Концепция MAHS (Machine Aided Human Summarization)

1. Система рекомендует *сценарий реферата* — список статей **подборки**, ранжированный в рекомендуемом порядке их упоминания (цитирования)
2. Пользователь может скорректировать сценарий в соответствии со своими целями
3. В цикле по статьям сценария, в порядке их упоминания:
 - пользователь запрашивает аспекты статьи, кликая на кнопки *суфлёров*: «как другие авторы обычно ссылаются на эту статью», «цель исследования», «основная идея», «метод», «результат», «вывод», «недостаток» и т.д.
 - *алгоритм суфлёра* строит ранжированный список релевантных фраз
 - пользователь добавляет фразу из предложенного списка в текст реферата
 - при необходимости пользователь корректирует текст реферата

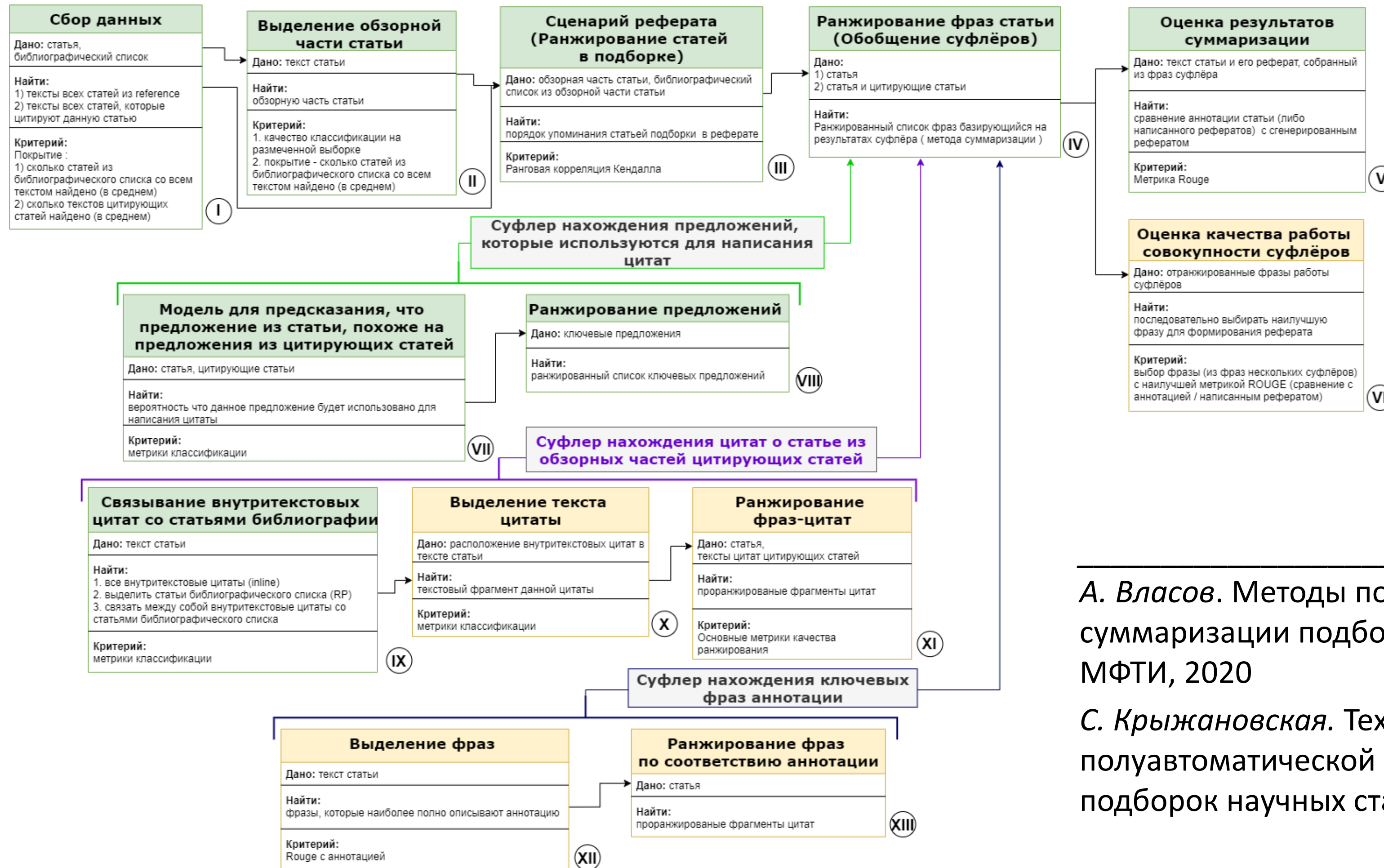
Полуавтоматическое реферирование подборки

Основные задачи машинного обучения:

- Формирование обучающей выборки: **paper** → **(refs, survey)**
- Ранжирование статей для сценария реферата
- Выбор релевантных фраз из текста статьи для каждого суфлёра
- Ранжирование выбранных фраз для каждого суфлёра
- Выбор релевантного контекста по данной ссылке, например:

Few contextual citation graphs are publicly available. The ACL Anthology Network (AAN) (Radev et al., 2009) is one such contextual citation graph built from the ACL Anthology corpus (Bird et al., 2008), consisting of 24.6K papers manually augmented with citation information. CiteSeer (Giles et al., 1998) provides a large corpus consisting of 1.0M papers with full text and bibliography entries parsed from PDFs. Saier and Farber (2019) introduces a contextual citation graph of approximately 1.0M arXiv papers with full text LaTeX parses where citations are linked to papers in the Microsoft Academic Graph.

Полуавтоматическое реферирование подборки



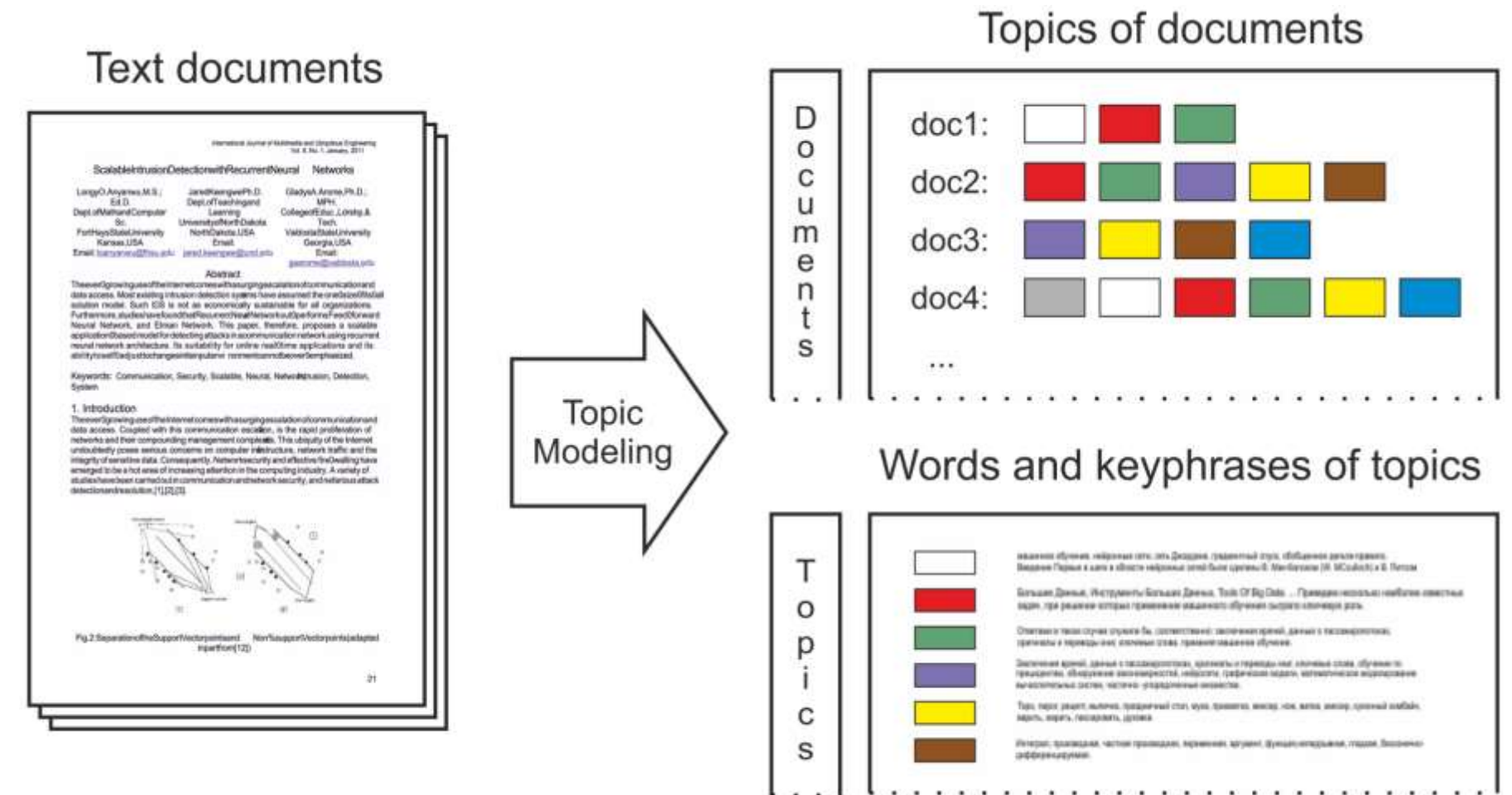
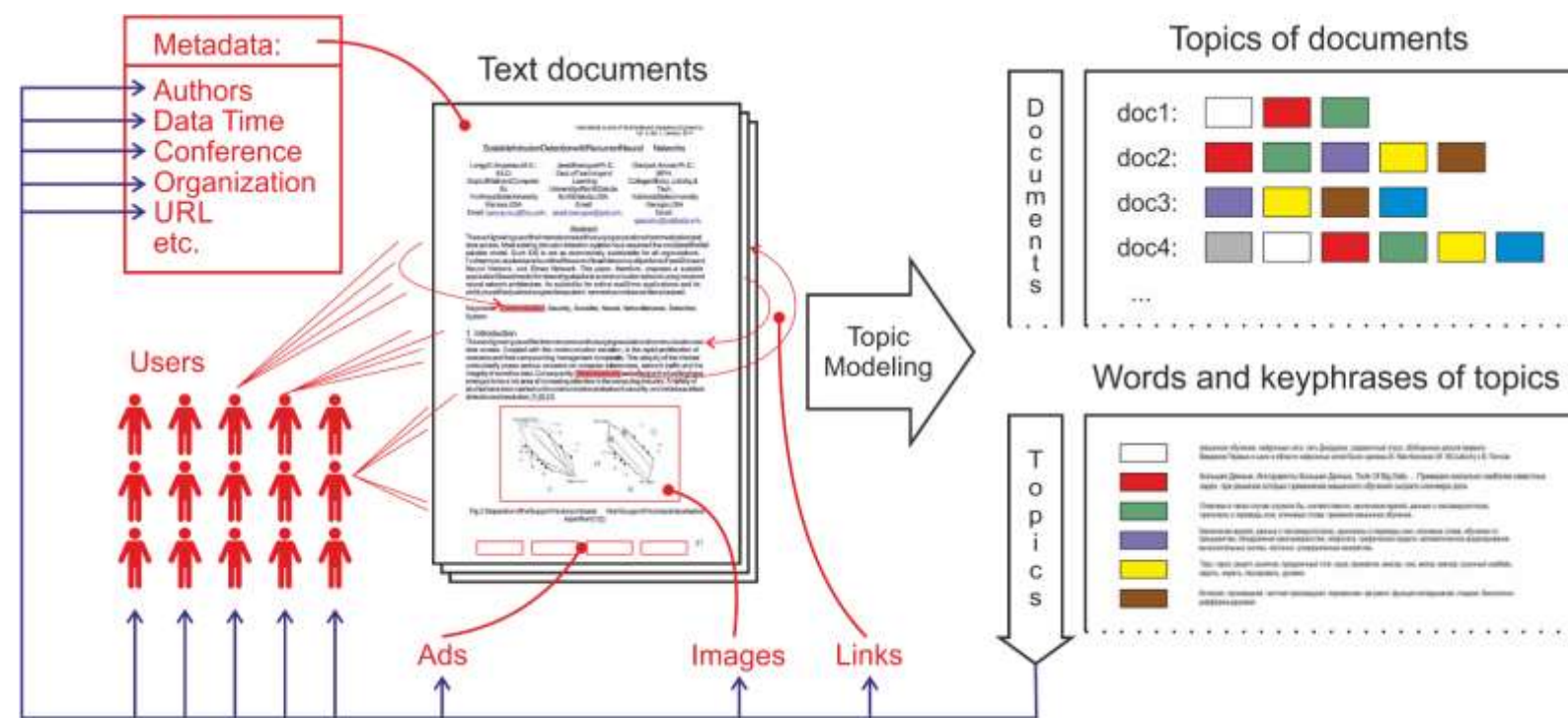
А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

С. Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

Тематическое моделирование подборки

Тематическая модель (ТМ) коллекции текстовых документов определяет

- какие темы есть в каждом документе
- из каких слов состоит каждая тема



Мультимодальная ТМ определяет также,

- какие ещё нетекстовые токены содержатся в каждой теме

Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#). AINL 2017.

Vorontsov K. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023.

<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

Технология тематического моделирования BigARTM

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

3.7М статей Википедии, 100К слов: время min (перплексия)

проц.	T	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

Мурат Апишев. Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией.

Диссертация к.т.н., ФИЦ ИУ РАН. 2020.

Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#). AINL 2017.

Vorontsov K. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023.

<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

Технология тематического поиска

Схема эксперимента:

- длинные запросы (1 стр. А4)
- 100 запросов на коллекцию
- 3 ассессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:



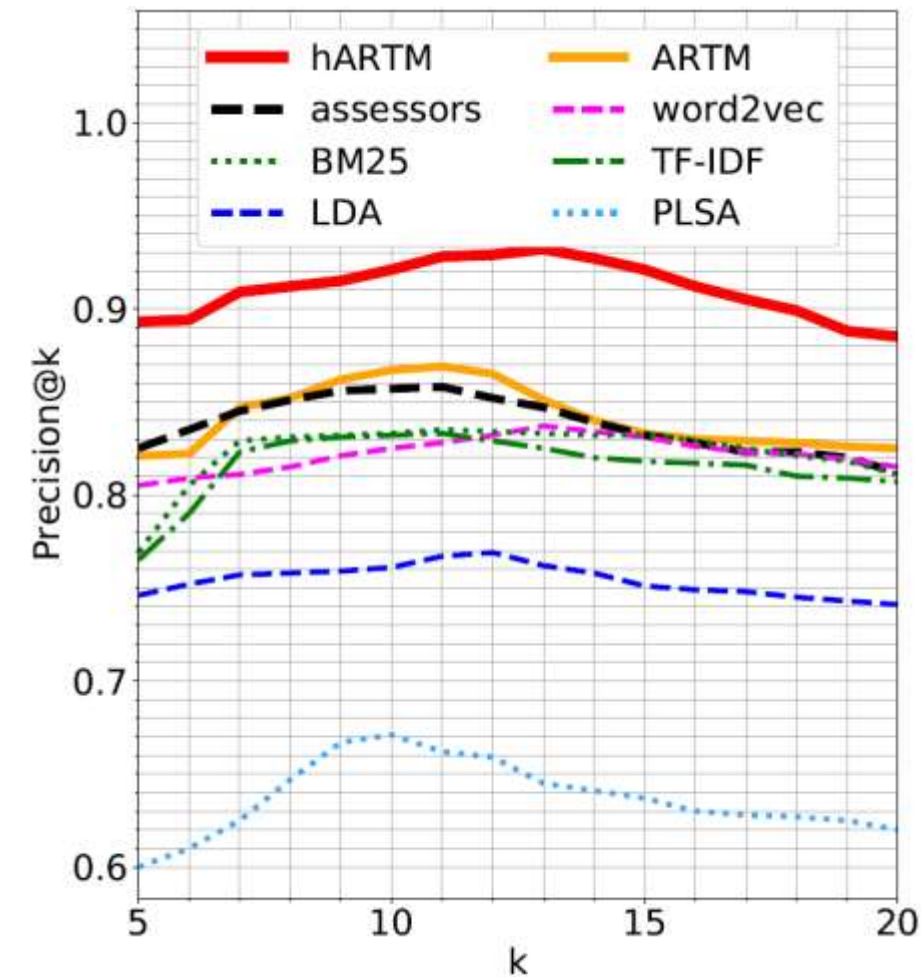
(170K Russian docs)



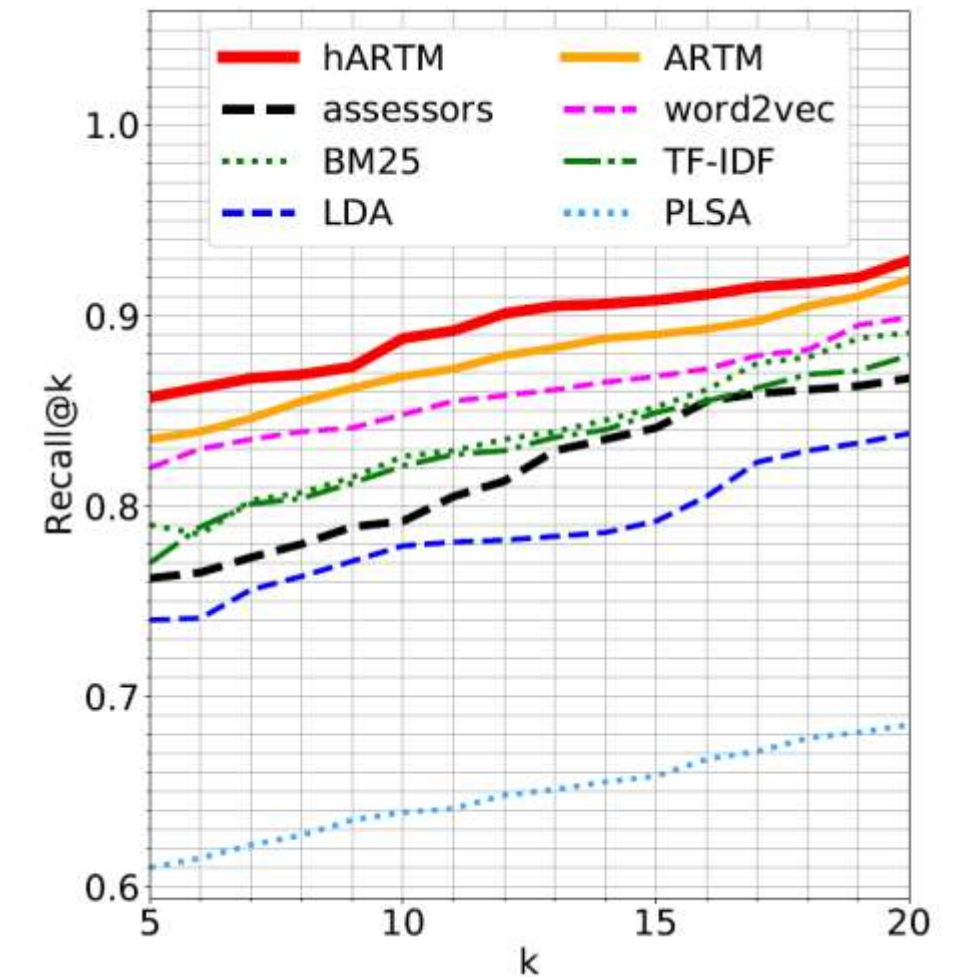
(750K English docs)

Оценки качества поиска:

точность (precision@k)



полнота (recall@k)



Анастасия Янина. Тематические и нейросетевые модели языка для разведочного информационного поиска.

Диссертация к.ф.-м.н., МФТИ. 2022.

Ianina A., Vorontsov K. [Regularized multimodal hierarchical topic model for document-by-document exploratory search](#). 2019.

Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#). AINL 2017.

Технология мультиязычного поиска и категоризации

Данные:

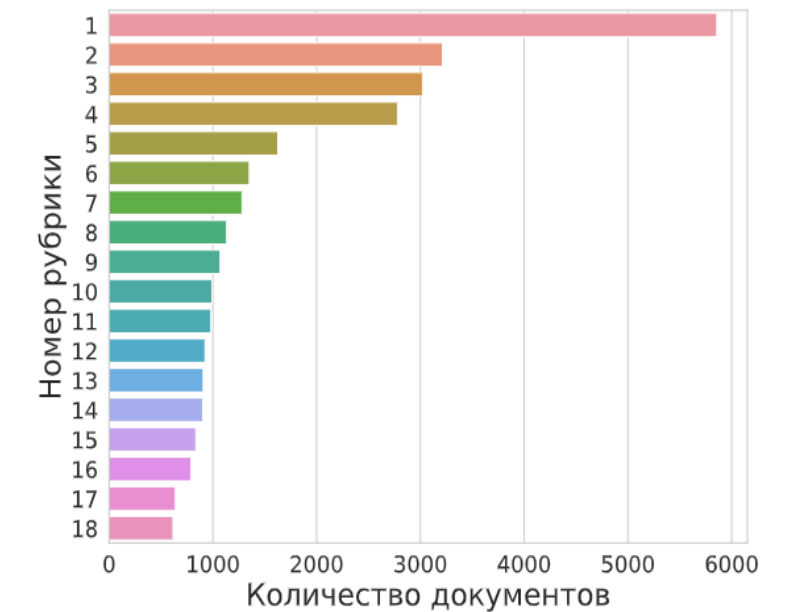
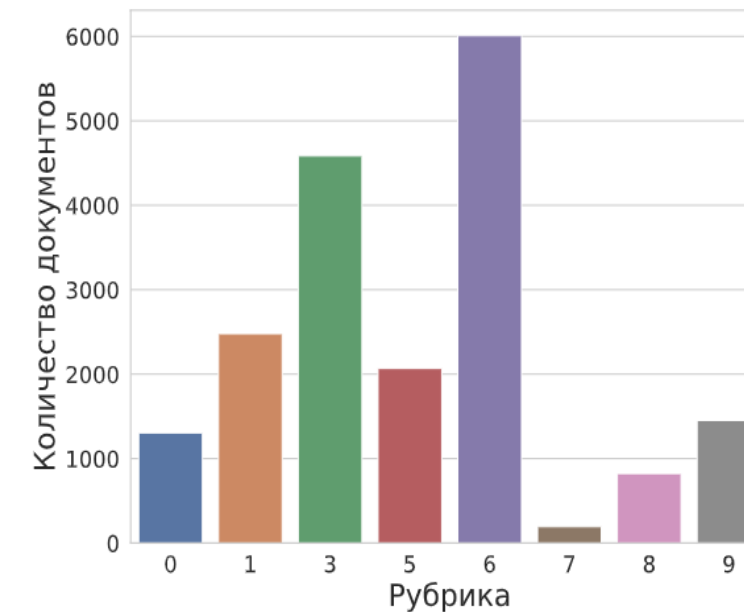
- научные статьи eLibrary и статьи Wikipedia (100 языков)
- рубрики ГРНТИ, ВАК, УДК, ОЭСР

Две задачи, одна модель:

- тематический поиск документов по документам
- категоризация документов

Особенности решения:

- модальности: языки, рубрики
- редукция словарей (VPE-токенизация) до 11 тыс. токенов на каждый язык
- сокращение модели с 128 Гб до 4.8 Гб



94%
Точность
поиска

Рубрикатор	ГРНТИ	ВАК	УДК	ОЭСР
Точность	81%	70%	86%	80%

Технология автоматического выделения терминов

Объединение трёх технологий: TopMine & SyntaxNet & BigARTM

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого случайного подмножества (2000 n -грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
логистическая регрессия, градиентный бустинг

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	0.95	0.38	0.91	0.97	0.41	0.99

$$\boxed{\text{Стат}} < \boxed{\text{Син}} < \boxed{\text{Син+Стат}} < \boxed{\text{Тем}} < \boxed{\begin{matrix} \text{Стат+Тем} \\ \text{Син+Тем} \end{matrix}} < \boxed{\text{Стат+Син+Тем}}$$

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

В.Полушин. Тематические модели для ранжирования рекомендаций текстового контента. 2017. ВМК МГУ.

Соревнование RuTermEval-2024



Поиск и классификация терминов в русскоязычных научных статьях:

- specific term – термины, специфичные доменно и лексически
- common term – общеизвестные термины, специфичные только доменно
- nomen – номенклатурные наименования доменно специфичных объектов

Метрики качества:

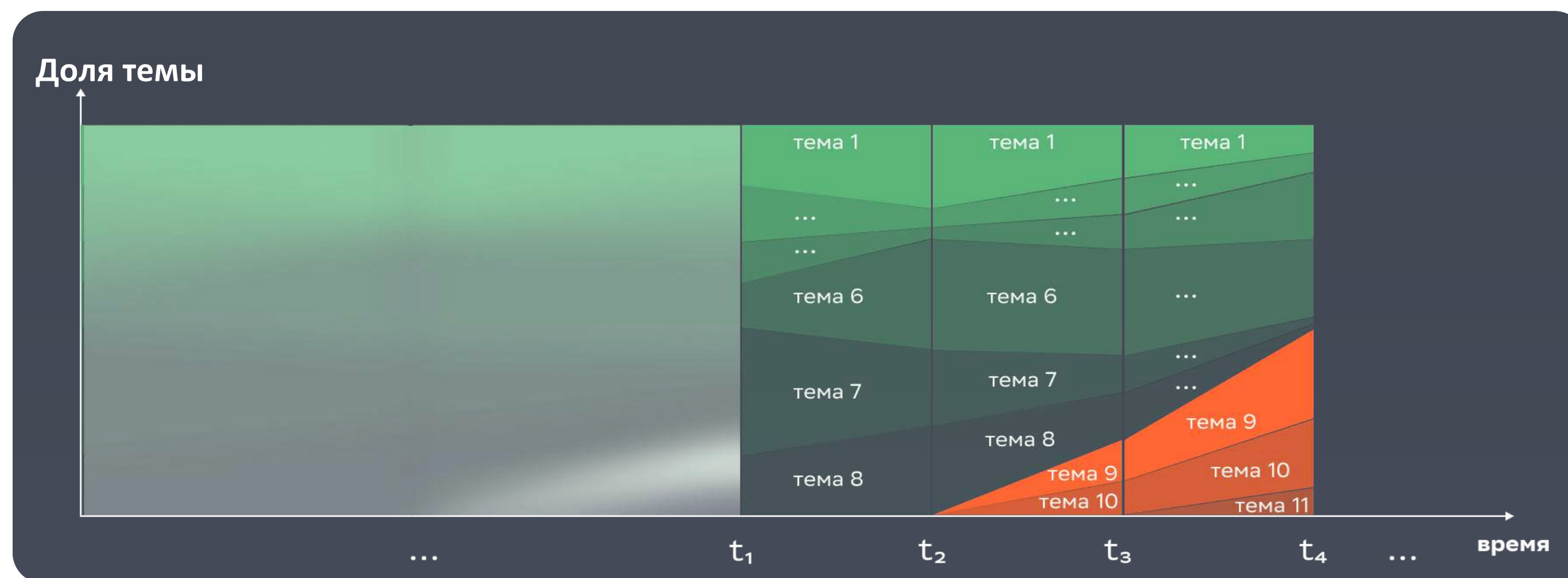
- полное/частичное совпадение выделенных терминов

Особенности соревнования:

- вложенные термины, мультидоменная и мультижанровая постановка задачи
- разметка: 1150 русскоязычных аннотаций,
20 статей конференции Диалог 2000-2023 (домен компьютерной лингвистики)
250 аннотаций статей пяти других доменов

Технология тематического поиска научных трендов

- *Темпоральная тематическая модель* дообучается последовательно без учителя (т.е. без размеченных данных) на статьях, вышедших за 30 дней
- Удаётся детектировать >60% из 87 трендовых тем (из области Data Science), выделенных экспертами в течение года после появления темы



Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В.

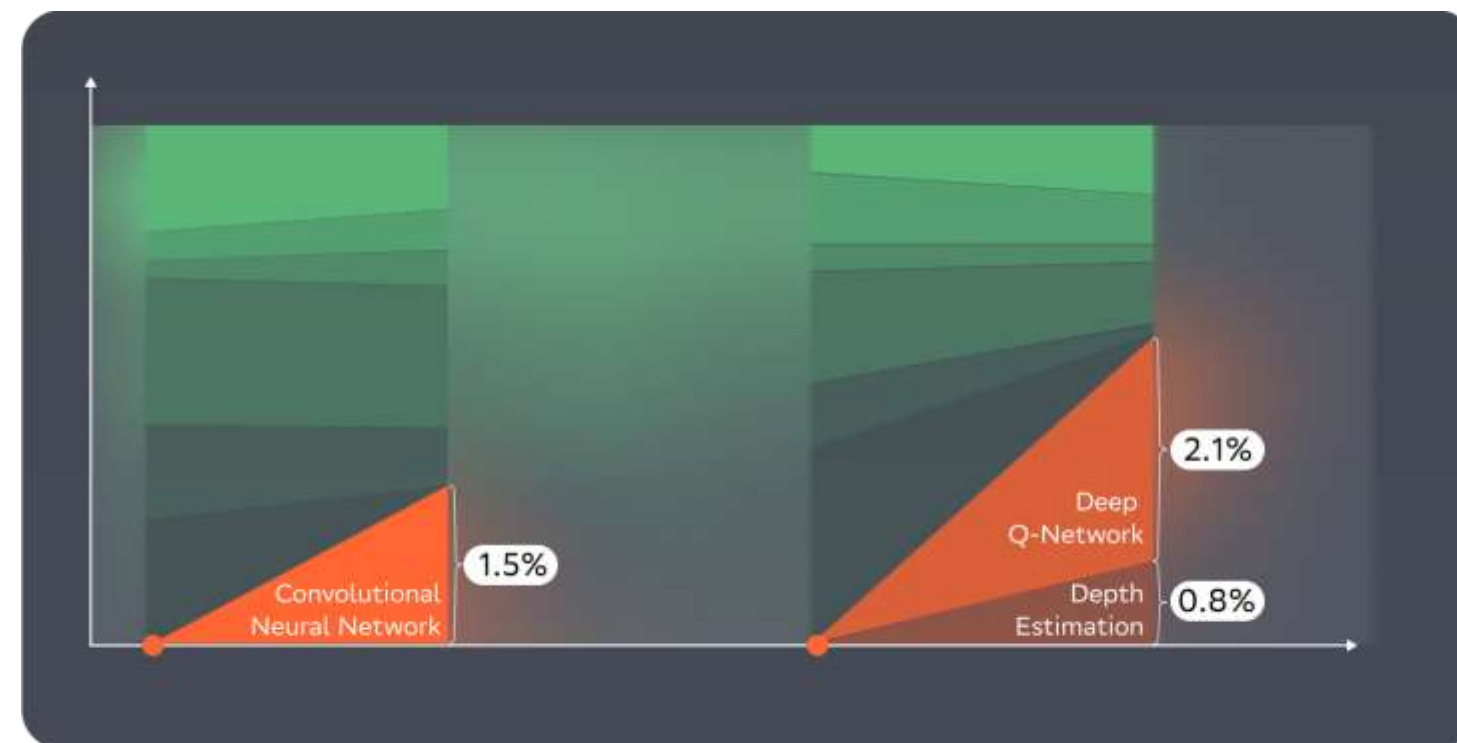
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. Математика, информатика, процессы управления, 2022, том 508, С.106–108

Тематический поиск научных трендов

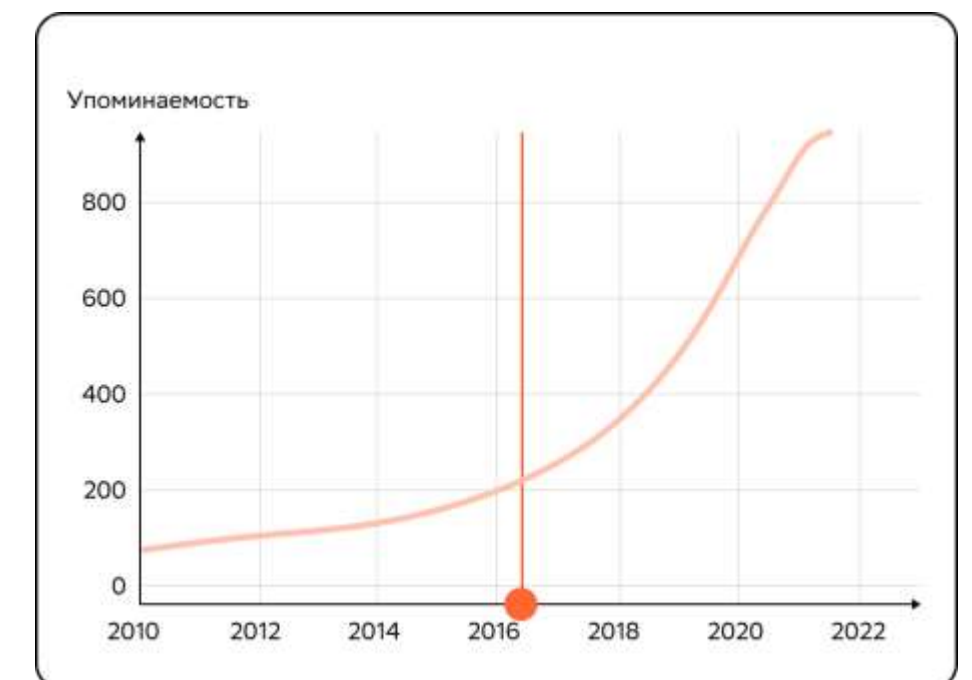
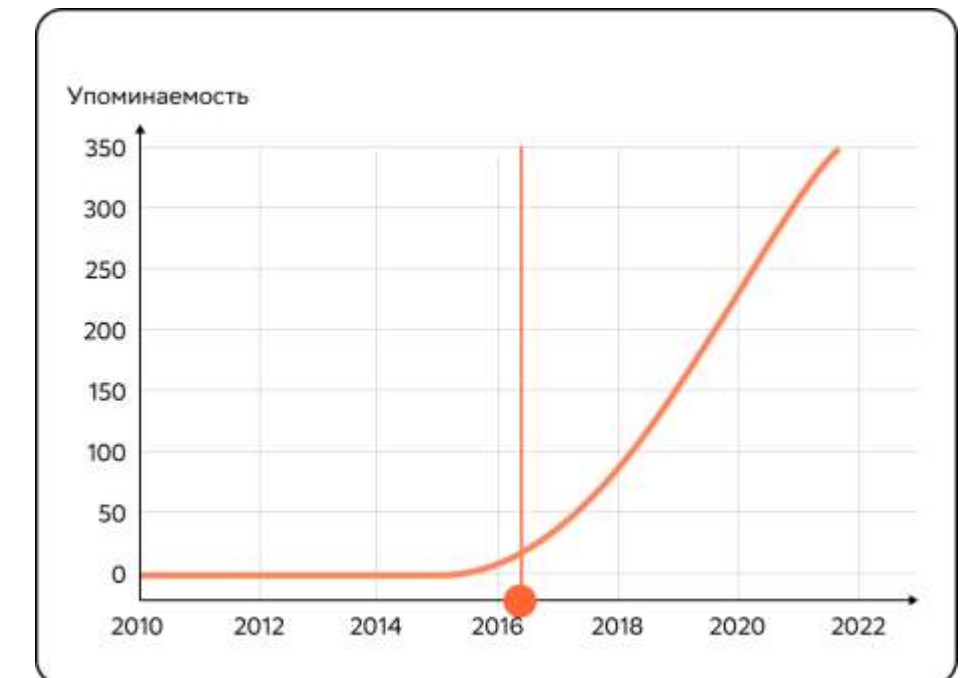
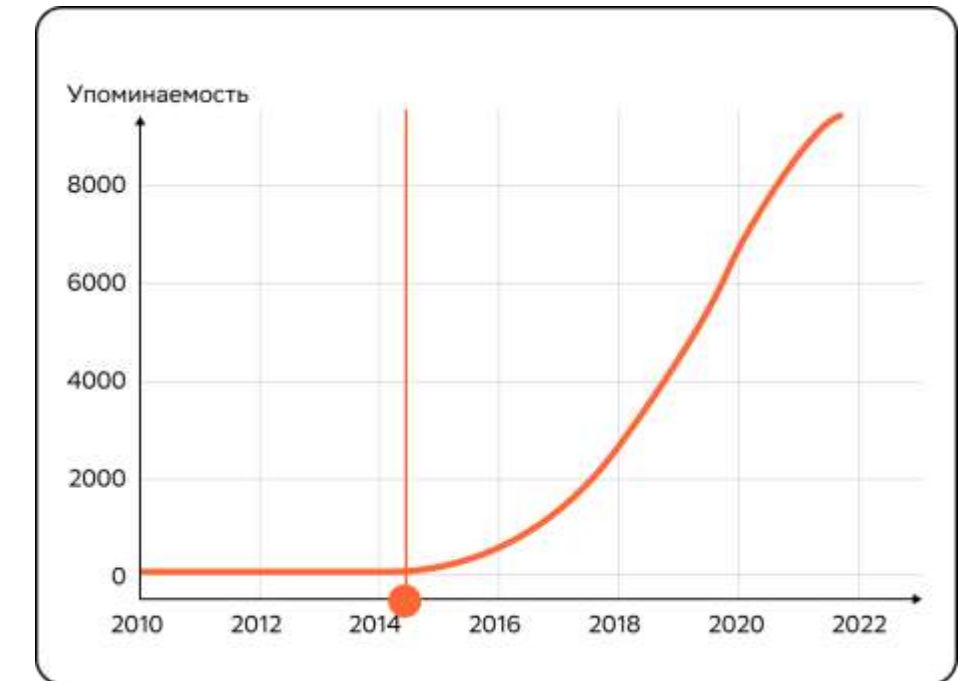
Трендовая тема:

- наличие семантического ядра
- наличие быстрого (обычно экспоненциального) роста

Примеры: динамика упоминаний трендовых тем



Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В.
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. Математика, информатика, процессы управления, 2022, том 508, С.106–10822222



Тематический поиск научных трендов: примеры тем

Topic modeling	Speech recognition	Collaborative filtering	Machine translation
latent variable	prosodic feature	web page	word alignment
mixture model	speech signal	search result	target language
topic model	eye gaze	recommender system	bleu score
mixture component	audio signal	collaborative filtering	parallel corpus
Gibbs sampling	spontaneous speech	word sense	source sentence
multinomial distribution	topic segmentation	ranking model	translation model
Gibbs sampler	acoustic feature	web search	machine translation
generative process	ASR output	user preference	sentence pair
Dirichlet distribution	switchboard corpus	user profile	source language
Dirichlet process	audio data	ranking score	best list

Тематический поиск научных трендов: примеры тем

StyleGAN

stylegan

latent code

mapping network

ablation study

text generation

generation quality

generator architecture

mask

encoder

gan model

Meta Learning

meta model

meta train

meta optimization

meta update

meta testing

training task

continual learning

previous task

catastrophic forgetting

ablation study

NERF

neural radiance field

accurate depth estimation

additional qualitative result

novel loss function

optical flow prediction

image reconstruction loss

monocular depth prediction

geometric consistency loss

depth estimation method

optical flow network

Примеры тематического моделирования в исторических исследованиях (газетные архивы)

[1] Корпус *Pennsylvania Gazette* 1728--1800, 25М слов:

- выделение последовательности событийных тем, изучение синхронности событий
- комбинирование автоматического анализа и ручного.

[2] Газеты Техаса от гражданской войны до наших дней:

- выделение динамики всех тем, связанных с хлопком;

[3] Газеты и периодика Финляндии (1854--1917):

- выделение тем о церкви, религии, образовании;
- тренды модернизации и секуляризации финского общества.

1. *D.Newman, S.Block*. Probabilistic topic decomposition of an eighteenth-century American newspaper. 2006.

2. *Tze-I Yang, A.J.Torget, R.Mihalcea*. Topic modeling on historical newspapers. 2011.

3. *J.Marjanen et al*. Topic modelling discourse dynamics in historical newspapers.

Примеры тематического моделирования в исторических исследованиях (летописи и дневники)

[4] Двужычный корпус книг на английском и немецком языке:

--- все темы, связанные с эпистемологией

[5] Корпус текстов на китайском языке (1644--1912):

--- все темы, связанные с бандитизмом, преступлениями;

--- анализ контекста для установления типа преступления.

[6] Дневник Martha Ballard (1735–1812), охватывает 27 лет:

--- выделение тем событийных и перманентных, персональных и исторических

--- специфичный английский XVIII века.

4. *M.Erlin*. Topic modeling, epistemology, and the English and German novel. 2017.

5. *Ian Matthew Miller*. Rebellion, crime and violence in Qing China, 2013.

6. *Cameron Blevins*. <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary>.

Примеры тематического моделирования в исторических исследованиях (журнальная периодика)

Статьи коллекции JSTOR доступны в виде «мешков слов».

[7] Научные журналы XX века:

- различия тематики на английском и немецком языках;
- особенно исследовались различия, связанные со 2МВ;
- для объединения тем использовались интервики Википедии.

[8] Более 100 лет литературно-художественной периодики:

- как менялись темы;
- как менялись значения слов внутри каждой темы;
- как менялась тема насилия (violence, power, fear, blood, death, murder, act, guilt).

7. *D.Mimno*. Computational historiography: Data mining in a century of classics journals. 2012.

8. *A.Goldstone, T.Underwood*. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. 2014.

Примеры тематического моделирования в политологии (анализ публичных выступлений)

[9] Выступления (210К) в Европарламенте, 1999--2014:

- выявление событийных тем и эволюции перманентных тем;
- как члены и комитеты ЕП влияют на формирование тем

[10] Модель контрастных мнений (Contrastive Opinion Modeling):

- выступления в Сенате США (www.votesmart.org);
- СМИ: New York Times, Xinhua News, The Hindu, 2009—2010

[11] Выступления в Совбезе ООН по Афганистану, 2001--2017:

- динамика отношения разных стран к проблемам Афганистана

9. *D.Greene, J.P.Cross*. Unveiling the political agenda of the European Parliament plenary: a topical analysis. 2015.

10. *Fang Y. et al*. Mining contrastive opinions on political texts using cross-perspective topic model. 2012.

11. *M.Schonfeld*. Discursive landscapes and unsupervised topic modeling in IR: a validation of text-as-data approaches through a new corpus of UN Security Council speeches on Afghanistan. 2018.

Примеры тематического моделирования в политологии (анализ СМИ и социальных медиа)

[12] Тематика изменения климата в СМИ Пакистана, 2010--2021

--- выявление, группирование и динамика тем

[13] Выявление поляризации новостей (AULIEN COVID-19):

--- 1,5М новостей, 440 источников СМИ, 11.2019--07.2020

[14] Выявление политических взглядов пользователей Twitter

[15] Что пишет NYT о ядерных технологиях с 1945 по н/в

12. *W.Ejaz et al.* Politics triumphs: A topic modeling approach for analyzing news media coverage of climate change in Pakistan. 2023

13. *Zihao He.* Detecting polarized topics using partisanship-aware contextualized topic embeddings. 2021

14. *R.Cohen, D.Ruths.* Classifying Political Orientation on Twitter: It's Not Easy! 2013.

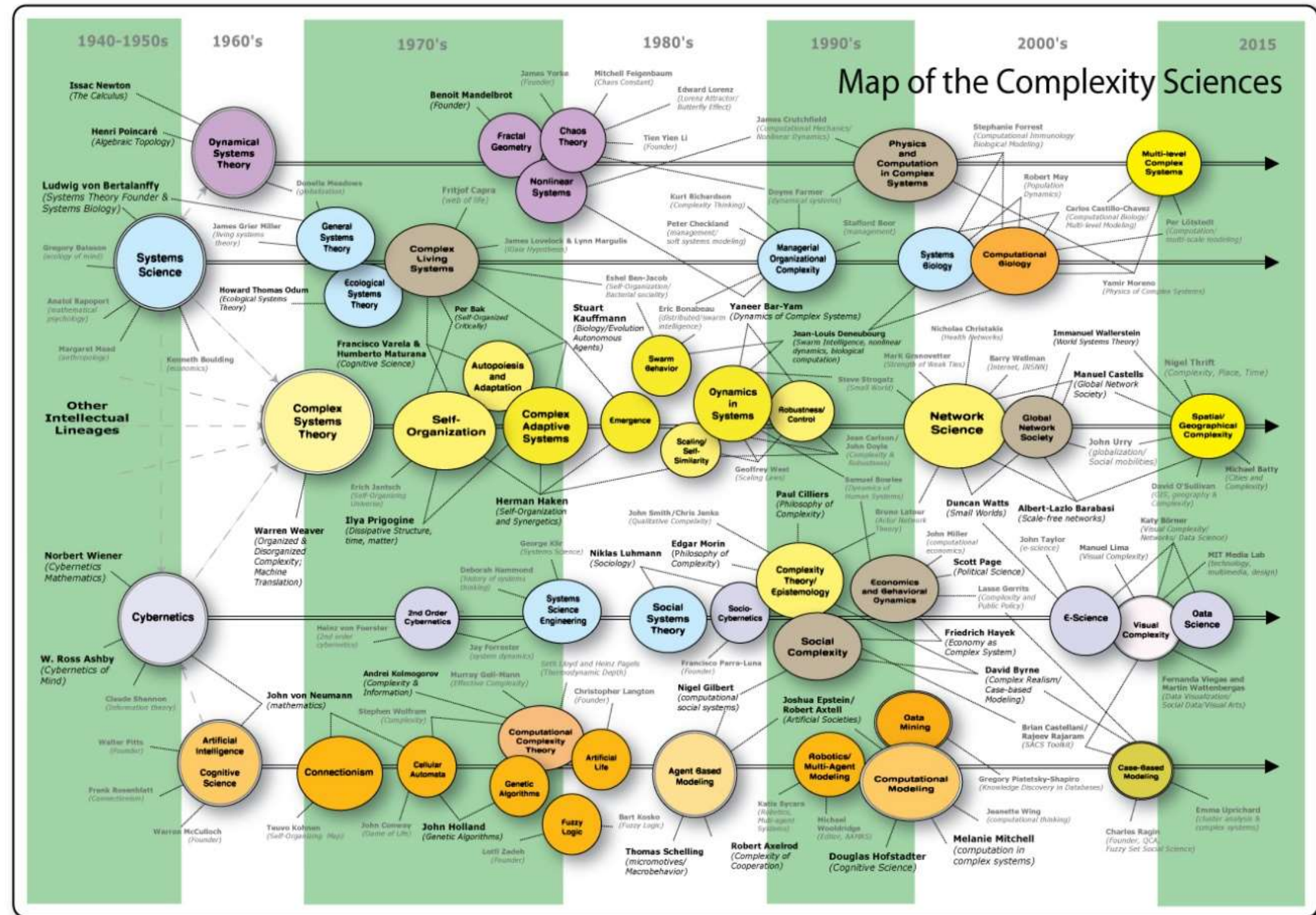
15. *C.Jacobi.* Quantitative analysis of large amounts of journalistic texts using topic modelling. 2015.

16. *H.Jelodar et al.* Latent Dirichlet allocation ({LDA}) and topic modeling: models, applications, a survey. 2019.

Визуализация: дистантное чтение (distant reading)

Осями на карте могут быть:

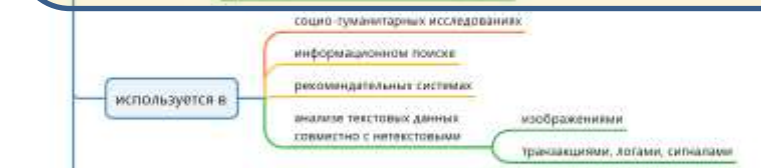
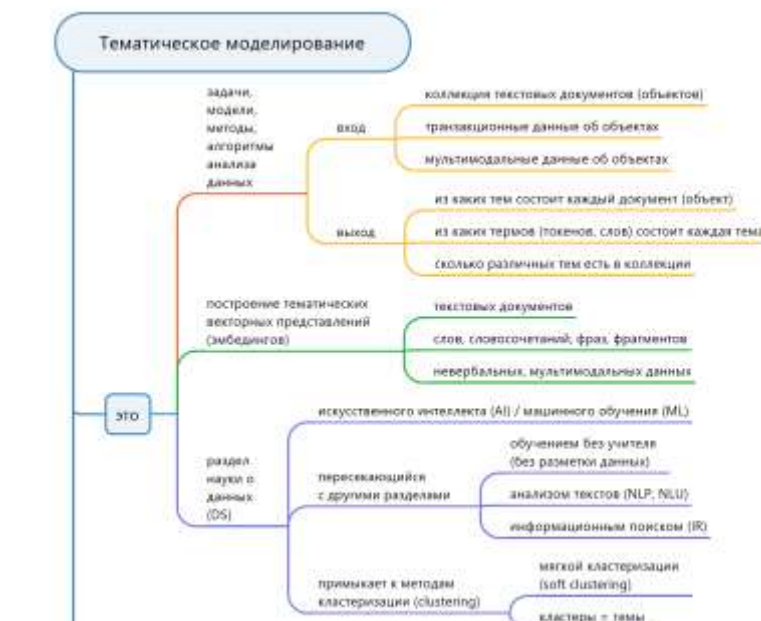
- время
- спектр тем
- сложность
- обзорность
- актуальность
- «хайповость»
- цитируемость



Визуализация: нелинейное чтение (non-linear reading)

Иерархическая суммаризация подборки документов:

- текстовое отображение — иерархический список
- графическая визуализация — mind-map



тематическое моделирование используется для



Выводы. Мастерская знаний

Миссия проекта: устранять барьеры между человеком и знанием

Реализовано: кросс-языковой поиск текстов, схожих по смыслу (не только научных статей: патенты, новости, тех.документация и т.д.)

Уверенность: большие языковые модели позволяют сегодня решать задачи, которые ещё 5 лет назад считались непреодолимо трудными

Планы:

- применение технологий ML/NLP/LLM к другим текстовым корпусам
- автоматическое структурирование, анализ, визуализация больших объёмов знаний, представленных в текстовом виде
- мульти-языковой поиск: китайский—русский—английский

Содержание

1. Технологии машинного обучения и искусственного интеллекта

- Обзор истории развития технологий ML/AI
- Обзор постановок задач машинного обучения
- Задачи анализа текстов и большие языковые модели

2. Проект «Мастерская знаний»

- Концепция мастерской знаний
- Большая языковая модель научных текстов
- Технологии анализа подборок текстовых документов

3. Языковые модели разметки текста

- Задачи автоматизации контент-анализа
- Методика оценивания моделей разметки
- Задача детекции ценностей социокультурного кода

Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

- 1) выделить фрагмент
- 2) классифицировать (тегировать) фрагмент по рубриктору
- 3) *связать несколько фрагментов*
- 4) *дать комментарий (затекст) к фрагменту или связи*

Цель — автоматизировать контент-анализ больших текстовых массивов по небольшим размеченным корпусам, в любой предметной области

Три задачи построения обучаемой модели разметки:

- 1) разработка рубриктора и инструкций разметчика
- 2) выбор большой языковой модели и её (до)обучение по разметке
- 3) оценивание качества разметки, сравнение и выбор моделей

Конкурс ПРО//ЧТЕНИЕ (<http://ai.upgreat.one>)

Задача: разметка смысловых ошибок в сочинениях ЕГЭ по русскому языку, литературе, истории, обществознанию и английскому языку.

Период: декабрь 2019 — июнь 2022, три цикла испытаний.

Призовой фонд: ₹100М русский язык + ₹100М английский язык

Типов ошибок: 152

(р:70 л:16 о:23 и:20 а:23)

Подтипов ошибок: 236

(р:112 л:19 о:29 и:26 а:50)

Помимо выделения ошибок, надо давать их объяснения.

ФАКТИЧЕСКАЯ ОШИБКА

автор высказывания А.Франц

В своем высказывании «Если человек зависит от природы, то и она от него зависит» Д. Мережковский **говорит** о необходимости защиты природы.

ЛОГИЧЕСКАЯ ОШИБКА

тезис не обоснован

Конкурс ПРО//ЧТЕНИЕ (<http://ai.upgreat.one>)

Сравнение разметки, сгенерированной алгоритмом, с разметкой эксперта

Алгоритмическая разметка

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безнравственный поступок? Именно эту проблему В. Ф. Тендряков поднимает в своем тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытав при этом чувство стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безнравственные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести - это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безнравственные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безнравственные поступки.

СВЯЗЬ РПОВТОР
РПОВТОР РЛИШН ПРОБЛЕМА
РПОВТОР РПОВТОР РПОВТОР
РЛИШН
РПОВТОР
РПОВТОР
РПОВТОР
РПОВТОР ГЛДНОР ГЛДНОР ГЛДНОР
ГВИДОВР РПОВТОР
РПОВТОР РПОВТОР
РПОВТОР РПОВТОР
РПОВТОР ГВИДОВР РПОВТОР
РПОВТОР
РПОВТОР

Экспертная разметка 2

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безнравственный поступок? Именно эту проблему В. Ф. Тендряков поднимает в своем тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытав при этом чувство стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безнравственные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести - это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безнравственные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безнравственные поступки.

РПОВТОР Т1
РПОВТОР Т1
РПОВТОР Т2 РПОВТОР Т1
ПРОБЛЕМА РПОВТОР Т2
ПРИМЕР РПОВТОР Т3
РТАВТ Т4 РПОВТОР Т1 РЛ
РПОВТОР Т1
РТАВТ Т4
РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
ПРОБЛЕМА
РПОВТОР Т1
РПОВТОР Т1

Конкурсы SemEval по детекции пропаганды

Базовая разметка: «фрагмент, метка класса»



Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum **omnium fortissimi** sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos **animos pertinent important**, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa **Helvetii quoque reliquos Gallos virtute praecedunt**, quod fere **cotidianis proeliis cum Germanis contendunt**, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values



Commissio
PopulusQue
Europaea

Упрощённая разметка: «предложение, метка класса»

Продвинутая разметка: «фрагмент, мишень, метка класса»

- SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup. <https://propaganda.math.unipd.it/semEval2023task3>
- *G.Martino, P.Nakov et al.* A survey on computational propaganda detection. 2020.
- *F.Alam, P.Nakov et al.* Overview of the WANLP 2022 shared task on propaganda detection in Arabic. 2022.

Разметка текста: обобщённый контент-анализ

Пик научной фантастики (и советской, и западной) пришелся на 1960–1970-е годы. Однако в 1970-х годах этот жанр начал постепенно затухать и сходить на нет, уже в 1980-х на Западе начинает набирать силу жанр фэнтези. Конечно же, это неслучайно. Именно 1960-е годы стали пиком научно-технического прогресса в XX веке. К тому времени закончилась первая половина XX столетия, за эти полсотни лет было изобретено столько, что все казалось возможным, верилось, что прогресс будет нарастать по экспоненте. **1960-е — это мир безудержного социального и культурно-технического оптимизма.** Человек полетел в космос, запустил искусственные спутники и задумался об освоении других планет.

Но этот порыв человечества в будущее создавал определенную угрозу для власти имущих как на Западе, так и в Советском Союзе. И уже в 1960-е годы перед сотрудниками Тавистокского института изучения человека в Великобритании (причем по иронии судьбы он располагается в графстве Девоншир, рядом с дартмурскими болотами, где разыгрывалась мрачная драма «Собаки Баскервилей» Конан Дойля) **была поставлена задача притормозить научно-технический прогресс путем внедрения определенных информационно-психологических и организационных моделей.** В частности, стартовала работа по созданию молодежных и женских субкультур и движений (именно в это время как по заказу появились The Beatles, The Rolling Stones, стал развиваться экологизм).

Одна из главных задач, поставленных перед Тавистокком, звучала так: to stamp out the cultural optimism of the 1960s (искоренить, вырубить, вытравить культурный оптимизм 1960-х годов). А **научная фантастика, особенно советская, безусловно, была оптимистической по своему настрою.**

Некоторые менее оптимистические ноты (не могу их назвать пессимистическими, но они выглядели более сложными, чем просто оптимизм) прослеживались у ряда писателей в соцлагере, в частности в книгах Станислава Лема (достаточно почитать его «Астронавтов» и «Магелланово облако»). Однако общий настрой советской фантастики до середины 1960-х годов был преимущественно оптимистичным — это видно и по творчеству братьев Стругацких, и по романам Ивана Ефремова.

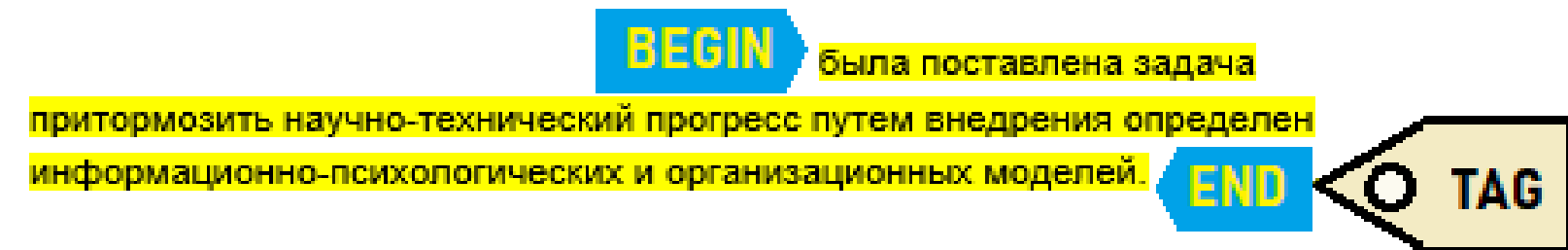
Первый доклад Римскому клубу (он создан в 1968 году) назывался «Пределы роста». В нем утверждалось, что человечество в своем индустриальном развитии достигло пределов, избыточно давит на природную среду, надо тормозить промышленно-экономическое развитие, перейдя к «нулевому росту». То есть 50 процентов всех средств должно идти на нейтрализацию негативных последствий, которые несет индустриальное развитие.

Разметка состоит из элементов

Элемент разметки — несколько взаимосвязанных фрагментов, затекстов и тегов

Теги (классы) выбираются из рубрикатора

Фрагмент задаётся началом и концом, может иметь один или несколько тегов:



Затекст — комментарий, объяснение, дополнительная информация и т.п., может иметь один или несколько тегов

Инструмент разметки

Весы войны Весь этот этический разлом дискуссии о военных действиях в Украине о том: стоит ли защита государства детской слезы? Стоит ли спасение некой организационной формы человеческого существования реально человеческой жертвы? Когда мои знакомые, скажу больше родственники, берут оружие и защищают свой город Львов - я понимаю их... Для них нет многосложных расчётов рисков России, для них нет геополитических вопросов и политики вообще. Есть город их детства, есть шум взрывов и солдаты другой страны... Когда Президент моей страны говорит о том, что перейдены все границы дозволенного, и мы вынуждены делать то, что делаем; а я смотрю на многолетние убийства и издевательства в ДНР; я вижу и знаю сколько денег и сил вложено в уничтожение России и подрыв наших ценностей - я понимаю это геополитическое решение. Проблема в том, что в пацифистском сознании любого современного образованного человека война - очень сложная и абстрактная химера. Когда сталкиваешься с ее реальностью становится страшно и хочется под лавку. Но России и нам россиянам под лавку больше нельзя. Мы сидели там долго и растеряли все. А на этой лавке сидели лорды и владельцы денег всех мастей. Никто не одобряет войну, об этом мы говорим все эти дни, но одуревший сосед, который ломится в НАТО и...

← Вернуться

Сохранить и выйти

Есть город их детства, есть шум взрывов и солдаты другой страны X

Есть город их детства, есть шум взрывов и солдаты другой страны X

+ добавить затекст

+ добавить затекст элемента разметки

добавить теги к элементу разметки

Тональность: отрицательная

Жилище

Жизнь

Чувство принадлежности (единство народов)

Этничность

сколько денег и сил вложено в уничтожение России X

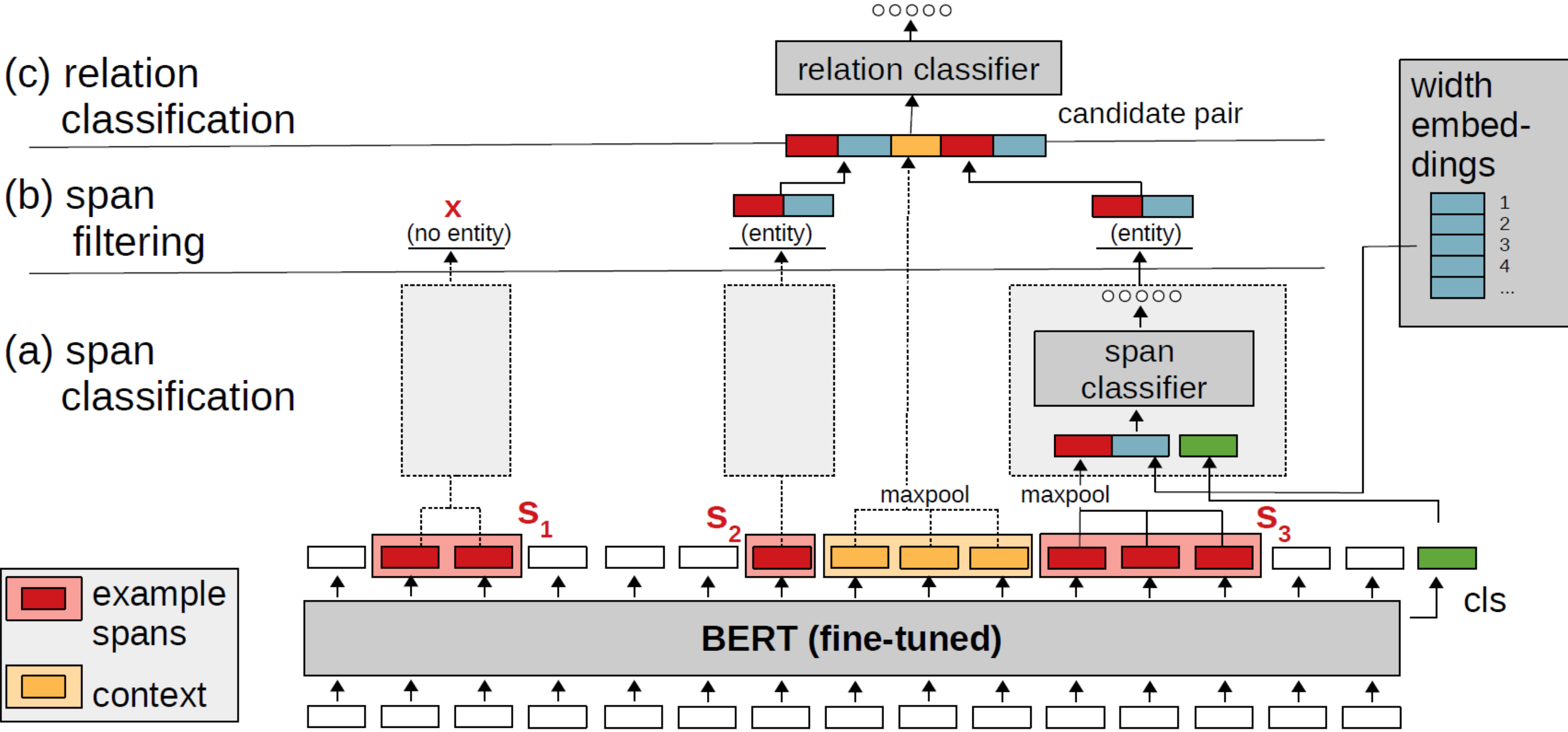
на этой лавке сидели лорды и владельцы денег всех мастей X

+ добавить элемент разметки

Введите имя тега

- ▶ Группа витальных ценностей
- ▶ Группа морально-нравственных ...
- ▶ Группа политических ценностей
- ▶ Группа религиозных ценностей
- ▶ Группа социальных ценностей
- ▶ Группа экзист. и познавательн...
- ▶ Группа эстетич. и гедонистичес...
- ▶ Пустое
- ▶ Служебные метки
- ▶ Тональность

Нейросетевые обучаемые модели разметки

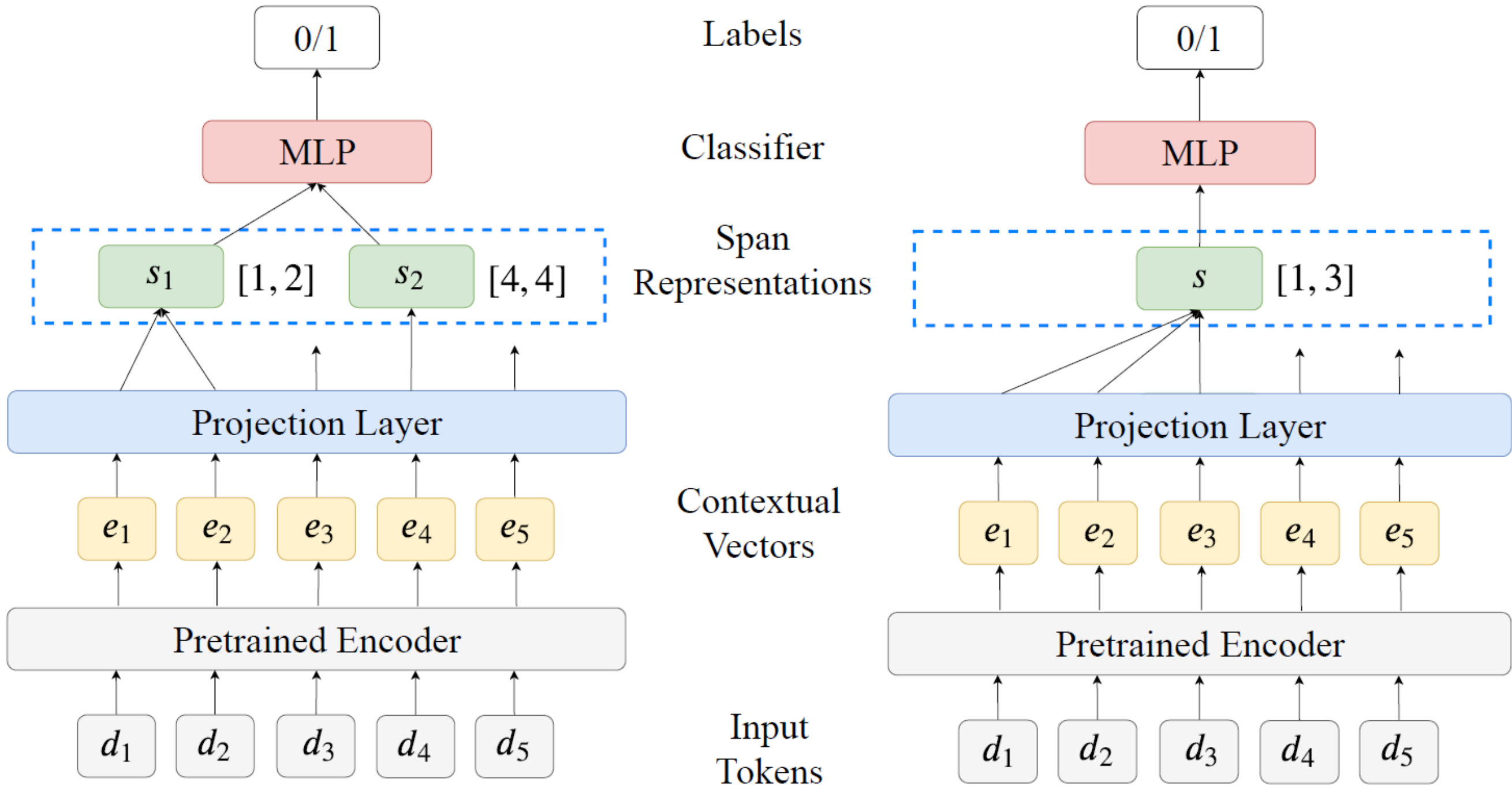


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

L.Anisiutin, T.Batura, N.Shvarts. Information extraction from news texts using a joint deep learning model. 2021.

Wayne Xin Zhao et al. A Survey of Large Language Models. ArXiv, 29 Jun 2023.

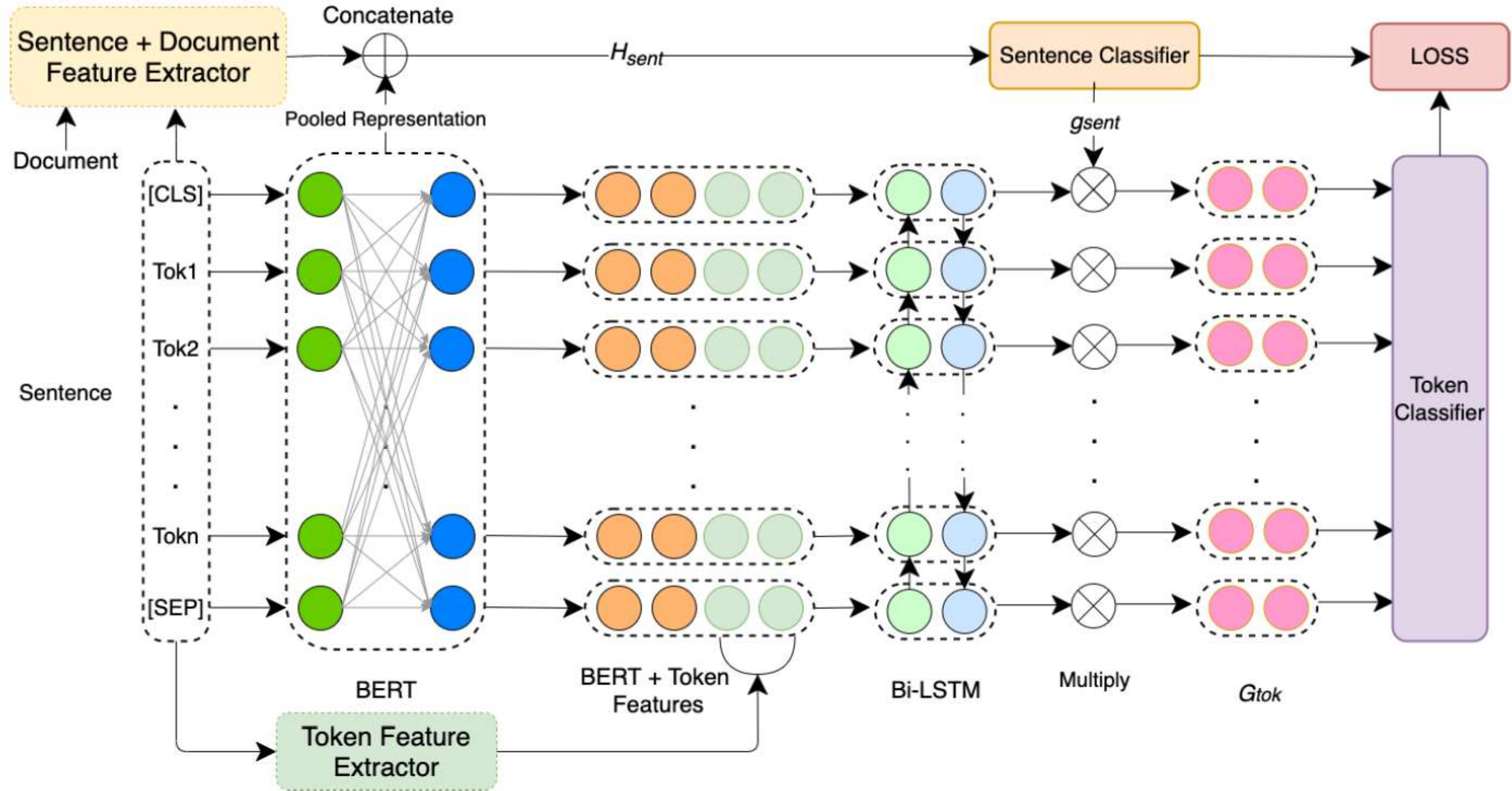
Нейросетевые обучаемые модели разметки



Xiaoya Li et al. A Unified MRC Framework for Named Entity Recognition. 2022.

S.Toshniwal et al. A Cross-Task Analysis of Text Span Representations. 2020.

Нейросетевые обучаемые модели разметки



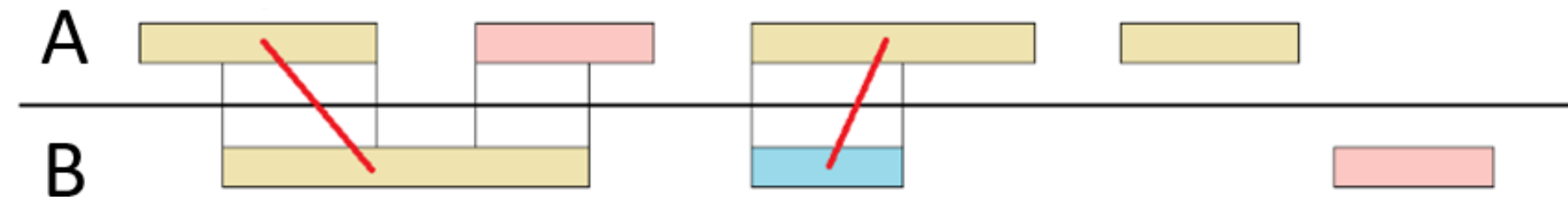
Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

Методика оценивания алгоритмической разметки

- В основе методики — парное сравнение разметок текста:
«алгоритм \leftrightarrow эксперт», «эксперт-1 \leftrightarrow эксперт-2»
на основе оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $Con_{1,...,5}(A,B)$
- Вводится их средневзвешенная согласованность $Con(A,B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по выборке $Con(A,E)$ разметок алгоритма A и эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по выборке $Con(E1,E2)$ разметок двух экспертов, E1 и E2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

Критерии согласованности разметок

Оптимальное сопоставление элементов разметок A и B



Критерии (числовые величины от 0 до 1; чем выше, тем лучше):

Con1 = доля фрагментов, для которых найдено сопоставление

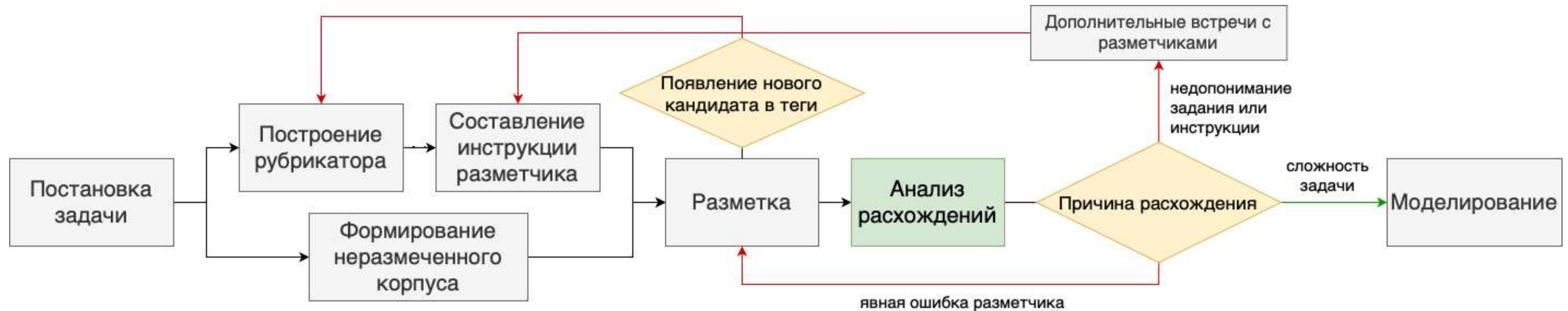
Con2 = точность наложения сопоставленных фрагментов

Con3 = точность совпадения тегов сопоставленных фрагментов

Con4 = точность совпадения связей сопоставленных фрагментов

Con5 = точность совпадения затекстов сопоставленных фрагментов

Организация процесса разметки



- каждый документ размечается несколькими экспертами (min 3)
- документы ранжируются по согласованности экспертов $Con(E, E')$
- наибольшие расхождения обсуждаются, вырабатывается консенсус
- происходит доработка инструкции и/или переразметка документов

Детекция ценностей социокультурного кода

Создание рубрикатора на основе кросс-дисциплинарного подхода

Исследователи	Предмет исследований
Милтон Рокич	Ценностные ориентации людей - психология, социология
Герт Хоффстеде	Культурные характеристики народов - социология
Шолом Шварц	Теория базовых человеческих ценностей - социальная психология
Рональд Инглхарт	Исследование мировых ценностей - политология, социология
Сэмюэл Хантингтон	Этнокультурное описание цивилизаций - политология, социология
Юрий Сергеевич Степанов	Концепты русской культуры - лингвистика
Александр Александрович Аузан	Культурные коды экономики - экономика

Указ президента Российской Федерации № 400 от 2-07-2021 «О стратегии нацбезопасности»

Указ президента Российской Федерации № 809 от 9-11-2022 «Об утверждении основ госполитики по сохранению и укреплению традиционных российских духовно-нравственных ценностей»

Какие ценности брать для рубрикатора: аксиоматический подход

- **Общественная значимость**
ценность — это то, что оказывает влияние на социальную жизнь
- **Индивидуальная значимость**
то, что влияет на принятие решений отдельными людьми
- **Субъективная измеримость**
то, что человек может принимать, отвергать или быть безразличным
- **Коммуникативность**
то, на отношение к чему можно повлиять в процессе коммуникации
- **Текстуальность**
то, что возможно описать, выразить текстом, фразой, историей
- **Атомарность**
то, что не сводится к набору других ценностей

Рубрикатор ценностей (1 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

Группа социальных ценностей

<ul style="list-style-type: none"> Социальные ценности Авторитет Альтруизм Благородство происхождения / аристократизм Важность общественного мнения Воспитание Гендерное разнообразие Дети Долгосрочная ориентация Дружба Избегание неопределённости Индивидуализм Интеллигентность Коллективизм Культура (нормы) поведения 	<ul style="list-style-type: none"> Личные границы Материальные ценности Патриархальность Патриотизм Пацифизм / мир во всём мире Полезность (созидательный труд) Профессиональный успех Репутация Семья Социальное признание Суеверия Трудолюбие / продуктивность Чувство принадлежности / единство народов Этничность Язык
--	---

Группа витальных ценностей

- Витальные (необходимые) ценности
- Безопасность (личная)
- Время
- Еда
- Жизнь
- Жилище
- Здоровье
- Природа

Группа политических ценностей

- Политические ценности
- Власть
- Выборность власти (демократия)
- Институциональное доверие
- Историческая память и преемственность поколений
- Либерализм
- Национальная безопасность
- Права и свободы
- Правосознание (гражданская активность, гражданственность)
- Справедливость

Рубрикатор ценностей (2 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

Группа религиозных ценностей

- Религиозные ценности
- Благочестивость
- Бог
- Религиозность
- Эзотерика

Группа эстетических и гедонистических ценностей

- Эстетические и гедонистические ценности
- Жизнь, полная впечатлений
- Красота
- Культура и искусство
- Наслаждение жизнью
- Потворство желаниям
- Творчество
- Эстетика

Группа экзистенциальных и познавательных ценностей

- Экзистенциальные и познавательные ценности
- Интеллект
- Критическое мышление
- Любовь
- Любознательность
- Мудрость
- Образование
- Перфекционизм
- Познание
- Принятие жизни
- Развитие
- Самостоятельность (выбор собственных целей)
- Смелость
- Смысл жизни
- Спокойствие (внутренняя гармония)
- Счастье
- Талант
- Твёрдая воля
- Целеустремлённость
- Широта взглядов

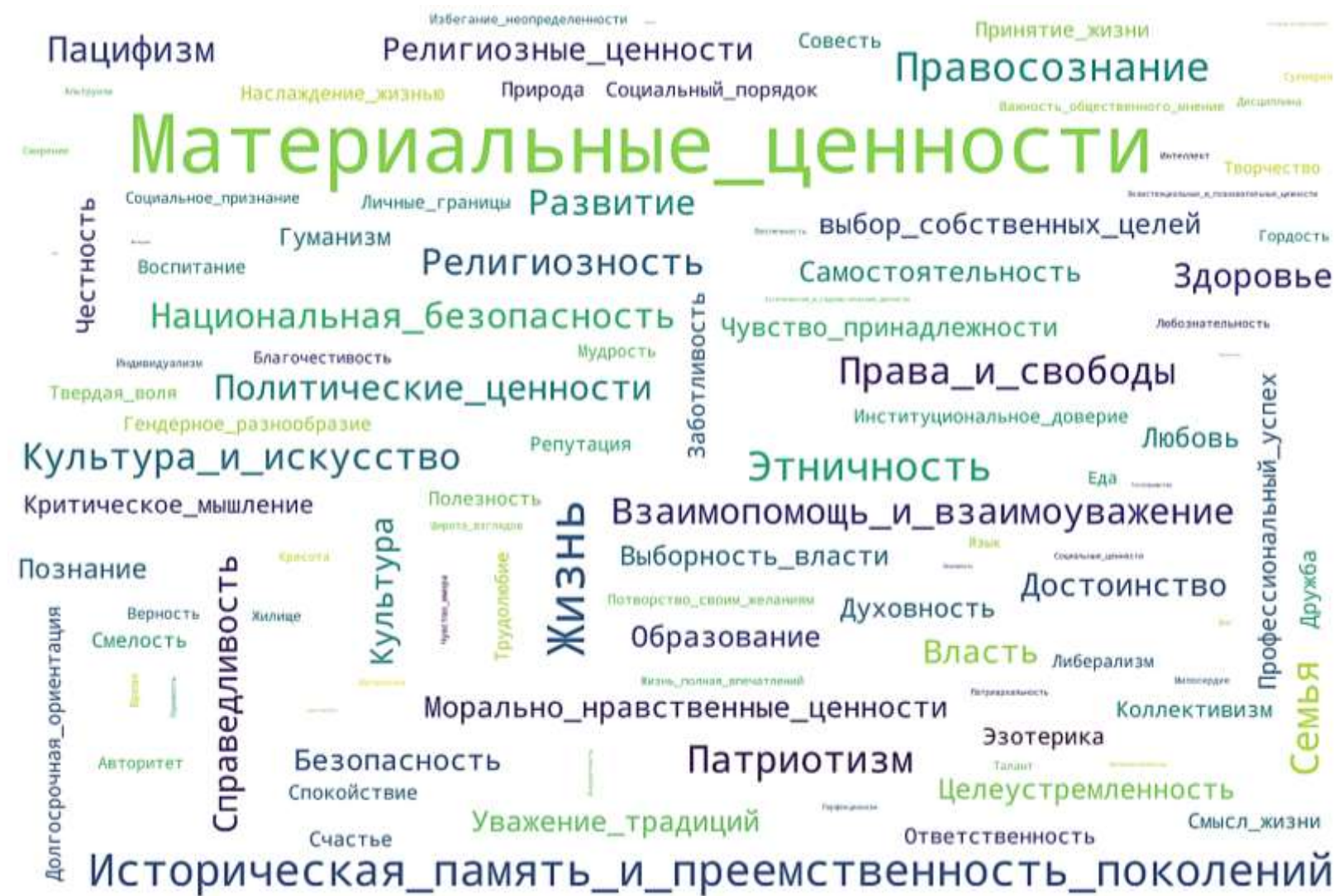
Группа морально-нравственных ценностей

- Морально-нравственные ценности
- Аккуратность
- Беспечность
- Вежливость
- Верность
- Взаимопомощь и взаимоуважение
- Гордость
- Гостеприимство
- Гуманизм
- Дисциплина
- Достоинство (самоуважение, самооценочность)
- Духовность (приоритет духовного над материальным)
- Заботливость
- Искренность
- Милосердие
- Ответственность
- Скромность
- Смирение (послушание, кротость)
- Совесь (нравственный закон, мораль)
- Терпение
- Уважение традиций
- Целомудрие
- Честность
- Чувство юмора

Ценностный ландшафт

Самые частотные теги
(по количеству элементов)

40%



1	Материальные ценности	1569
2	Жизнь	478
3	Историческая память и преемственность поколений	329
4	Правосознание (гражданская активность, гражданственность)	267
5	Политические ценности	246
6	Семья	243
7	Этничность	237
8	Культура и искусство	236
9	Права и свободы	235
10	Патриотизм	233

1. Rink Olga, Lobachev Viktor, Vorontsov Konstantin. Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCI 2024. Lecture Notes in Computer Science series (in print). Cham: Springer.
2. Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K. Methodology for detecting human values in large text collections // Bulletin of St. Petersburg University. International relations. In Russian (in print)

Рекомендуемые материалы

- *Визильтер Ю. В.* От слабого ИИ к общему универсальному интеллекту (обзор тенденций 2020-2023). Семинар РАИИ и ФИЦ ИУ РАН «Проблемы искусственного интеллекта» 31.01.2024
<https://rutube.ru/video/2aad53ec833f19918c1593398a2a1b88/>
- Не пропустите открытие тысячелетия! // Vital Math, 13 января 2024,
<https://www.youtube.com/watch?v=JZjH0it9Jyg>
- Report: AI Decrypted: A Guide for Navigating AI Developments in 2024, January 24, 2024
(Навигатор по ИИ-ландшафту от DENTONS GLOBAL ADVISORS)
<https://www.albrightstonebridge.com/news/report-ai-decrypted-guide-navigating-ai-developments-2024>
- 5 идей применения ИИ в вашем бизнесе прямо сейчас, 5 октября 2023.
<https://dzen.ru/a/ZR6ZeK5B3IL6OxXv>
- *Воронцов К. В.* Лекции по машинному обучению. www.MachineLearning.ru, 2004-2024.
- *Кевин Мэрфи.* Вероятностное машинное обучение. Введение. 2022.
- *Шумский С. А.* Машинный интеллект. РИОР ИНФРА-М, 2020.

Дискуссия

1. Предобученные модели внимания / трансформеры позволяют решать всё более трудные задачи NLP / NLU
2. Разметка текстовых данных — один из магистральных путей формализации гуманитарных знаний?
3. Какие задачи анализа исторической информации могли бы быть автоматизированы путём обучения моделей ИИ по разметкам?
4. Выявление реинтерпретаций исторических событий в СМК?

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

k.vorontsov@iai.msu.ru

<http://www.MachineLearning.ru/wiki?title=User:Vokov>