15-я международная конференция ИОИ-2024
15th Int. Conf. Intelligent Data Processing 2024
Беларусь, Гродно, 23-27 сентября 2024

«Мастерская знаний»: большие языковые модели для поиска и систематизации научной информации

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН, руководитель лаборатории машинного обучения и семантического анализа

Институт искусственного интеллекта МГУ им. М.В. Ломоносова

Концепция «Мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, 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

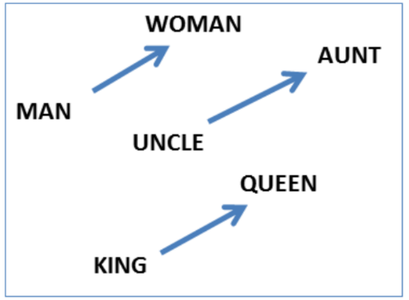
Эволюция подходов в обработке естественного языка

Декомпозиция задач по уровням «пирамиды NLP»

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

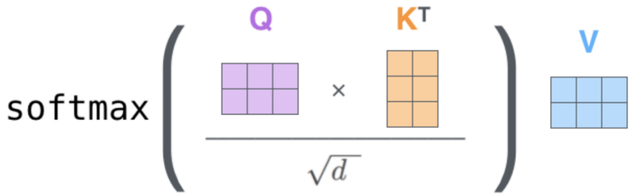
Модели векторных представлений слов (эмбедингов)

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



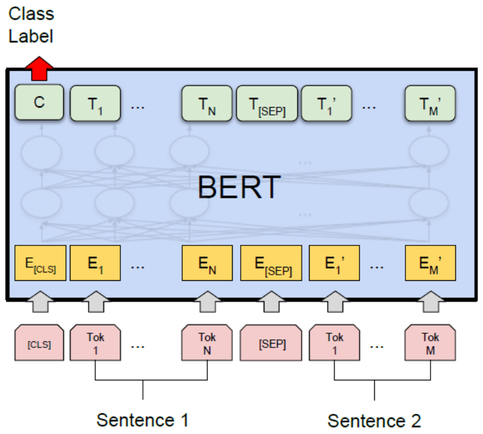
Нейросетевые векторные модели локальных контекстов

- рекуррентные нейронные сети: LSTM, GRU,…
- «end-to-end» модели внимания, трансформеры, LLM:
  машинный перевод, BERT [2018], GPT-3 [2020], GPT-4 [2023],…

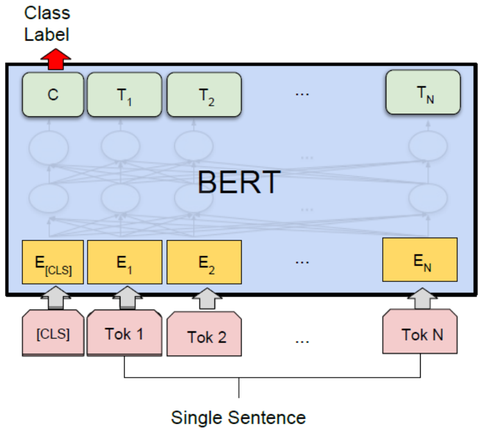


Трансформеры: нейросетевые модели языка

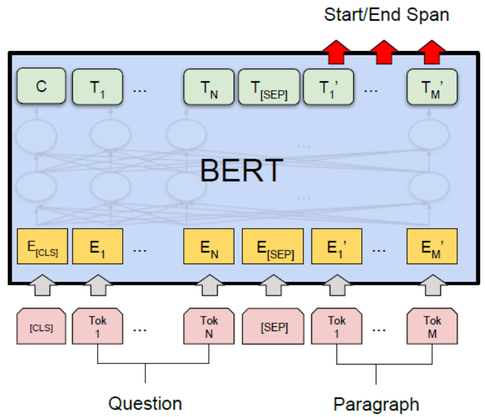
- обучаются векторизовать и предсказывать слова по контексту
- обучаются по терабайтам текстов, «они видели в языке всё»
- мультиязычны: обучаются на десятках языков
- мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



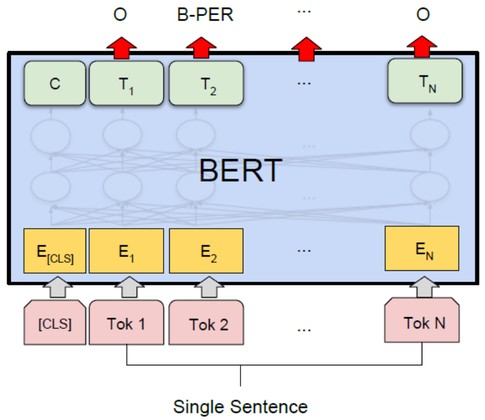
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



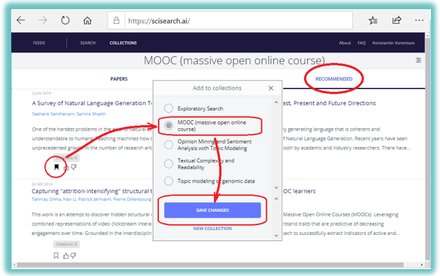
(c) Question Answering Tasks: SQuAD v1.1



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

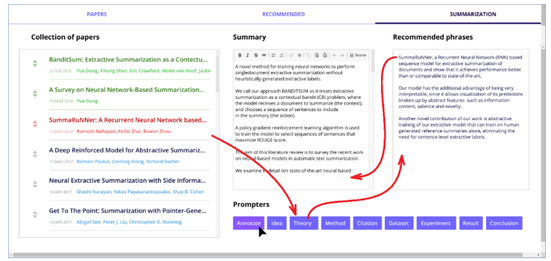
Функции «Мастерской знаний»

Подборка текстов – поисковый интерес и рабочее пространство пользователя/группы
Расширенная подборка — подборка + семантически близкие тексты



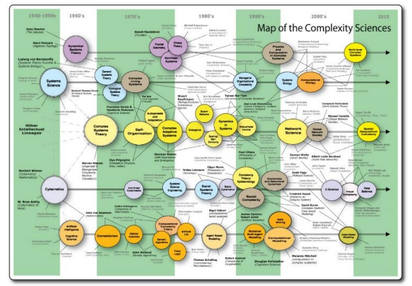
Поисково-рекомендательные сервисы:

- поиск семантически близких документов по подборке
- контекстный поиск по фрагменту документа из подборки
- мониторинг новых документов по тематике подборки



Аналитические сервисы:

- полуавтоматическое реферирование подборки
- тематизация, картирование, онтологизация подборки
- хронологизация, выявление трендов по тематике подборки
- контент-анализ, сбор и анализ фактов из документов подборки

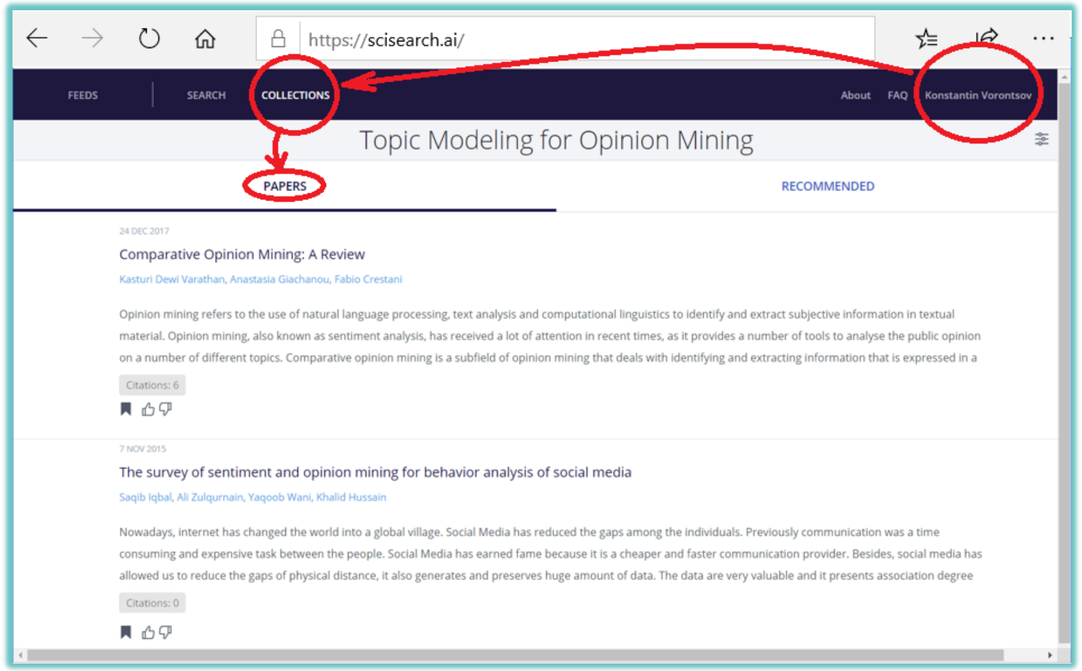


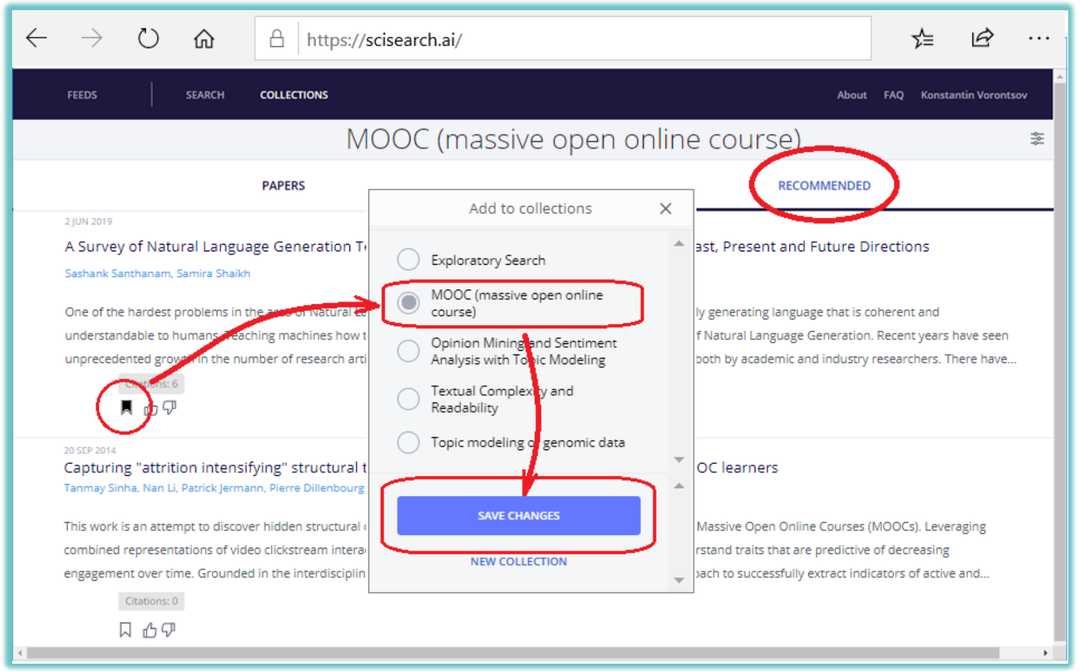
Коммуникативные сервисы:

- совместное составление, обсуждение, использование подборок
- инструментализация «коллективного разума» для подборки

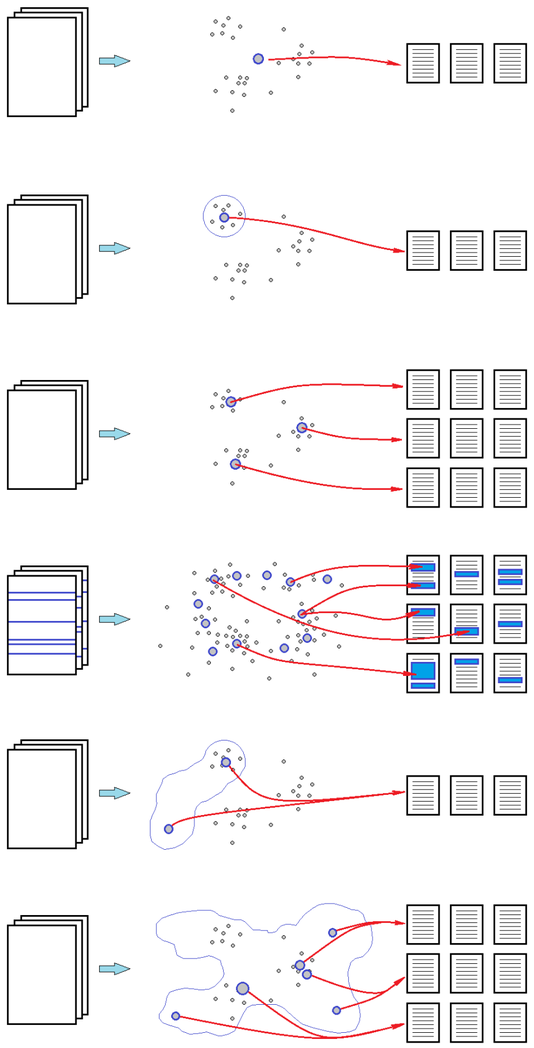
Поиск и рекомендации (прототип интерфейса)

Подборка играет роль поискового запроса и поисковой выдачи одновременно



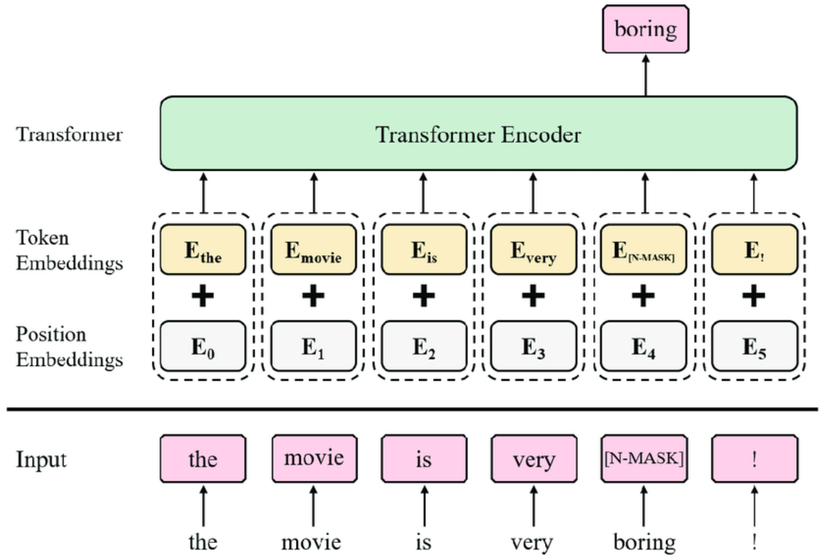


Стратегии векторного документного поиска



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 сегментам 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 (23M 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**
  205M публикаций, 121M авторов
  30M (12B токенов) отобрано для обучения модели,
  title+abstract, 85% на английском, 2% на русском

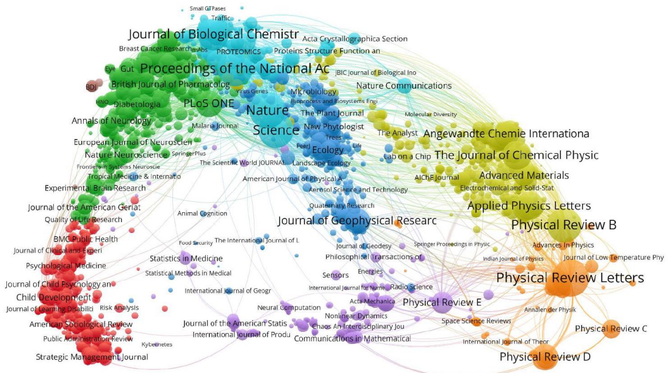
SEMANTIC SCHOLAR

- **eLibrary**, заголовки и аннотации (title+abstract):
  8.6M (2B токенов) на русском
  8.8M (1.2B токенов) на английском

eLIBRARY.RU

Данные для дообучения:

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

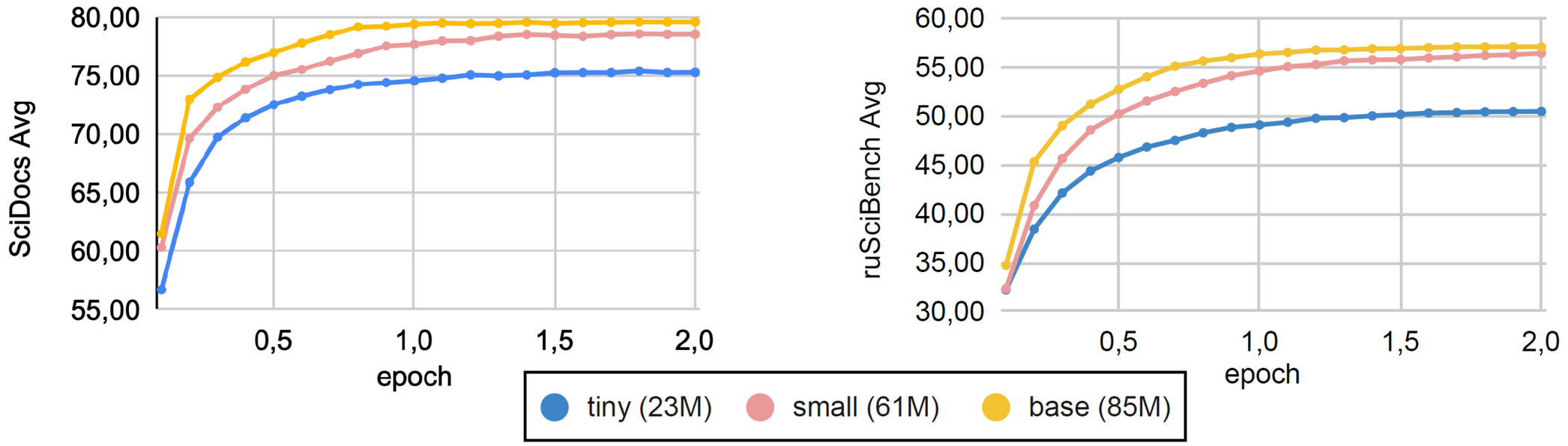
- классификация OECD/ГРНТИ по аннотации ru / en / ru+en
- поиск аннотации по её переводу ru→en / en→ru

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

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)

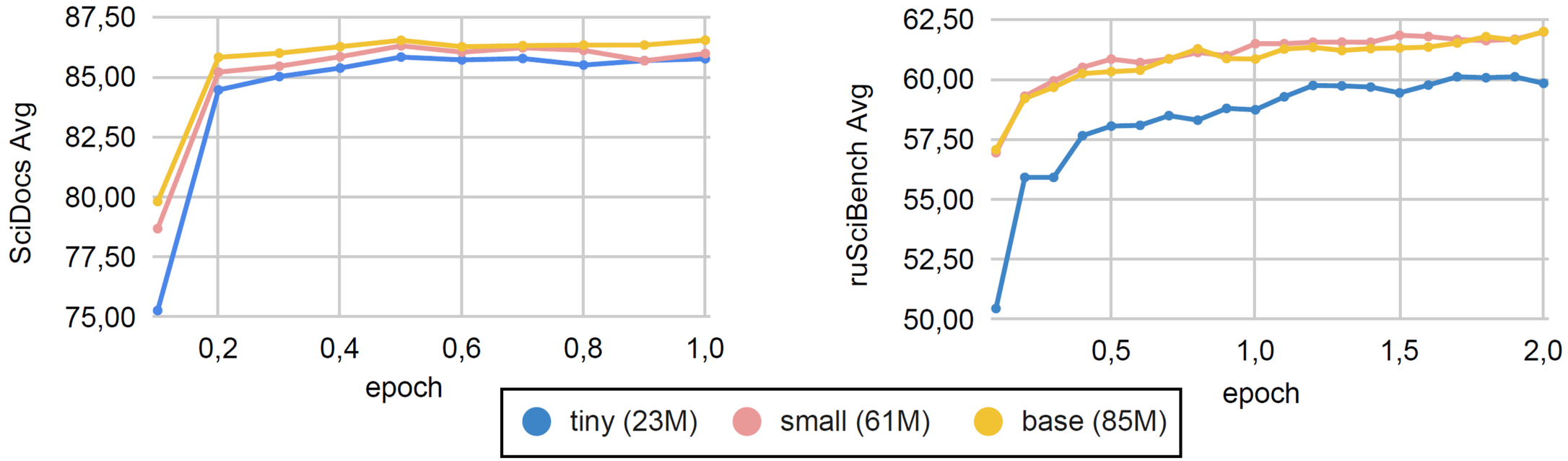
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



Этап 2: дообучение на парах title-abstract

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

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



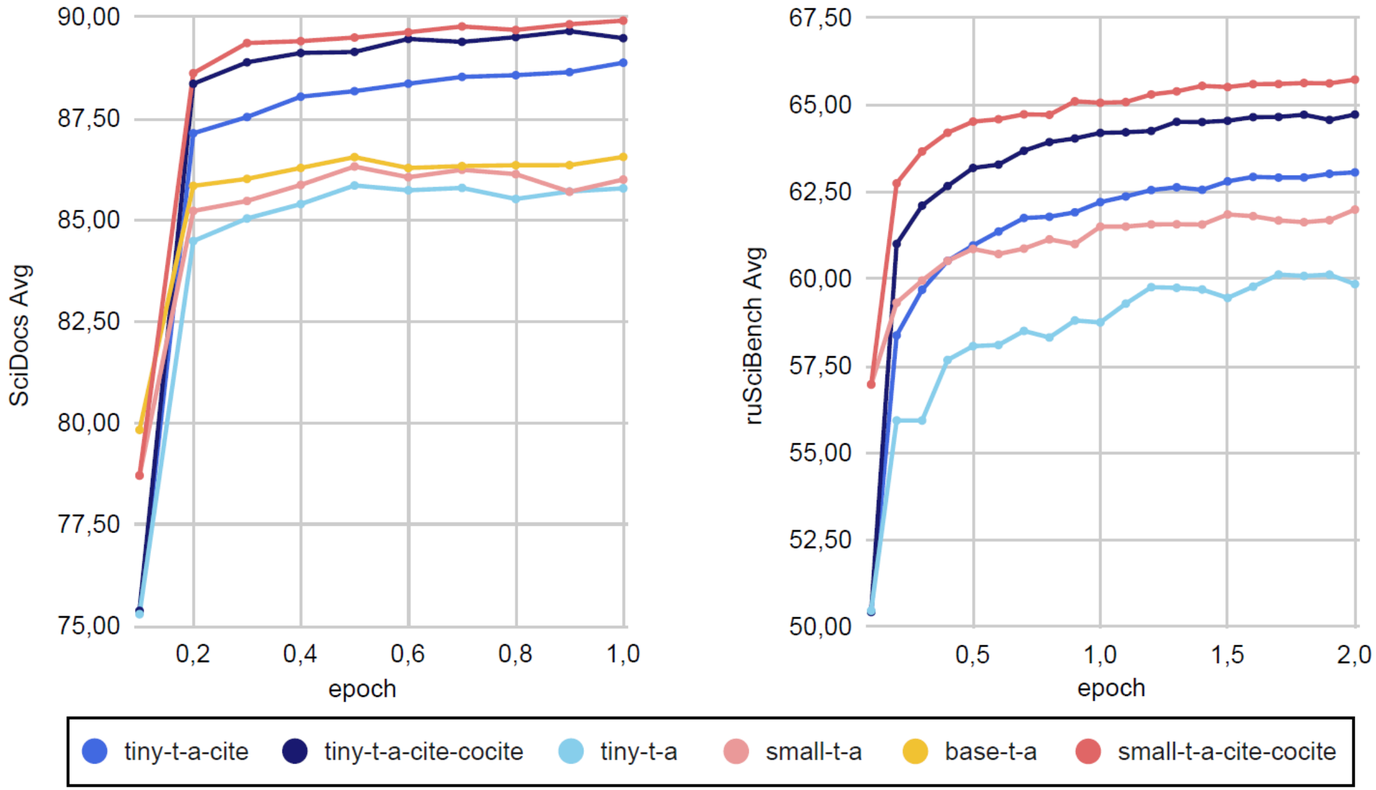
L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

Этап 3: дообучение на парах cite-cocite

Критерий: сближать эмбединги пары документов (A,B) при цитировании:
«cite» — статья A цитирует статью B
«co-cite» — третья статья C цитирует статьи A и B

S2AG:
- 13.3M пар cite
- 62M пар co-cite

eLibrary:
- 40M пар cite
- 33.7M 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	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

16

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: 23M против 110M
   — меньше размерность эмбедингов: 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

Н.Герасименко, А.Ватолин, А.Янина, К.Воронцов.
Маленькая большая языковая модель для обработки научных текстов. 2024. (на рецензировании)

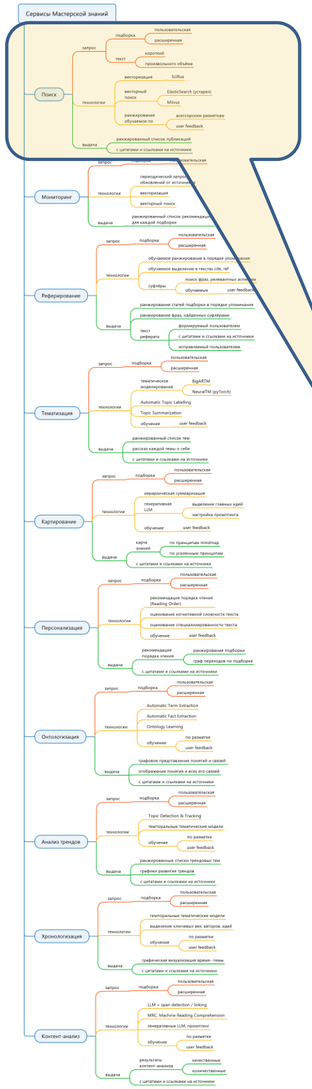
Первое внедрение

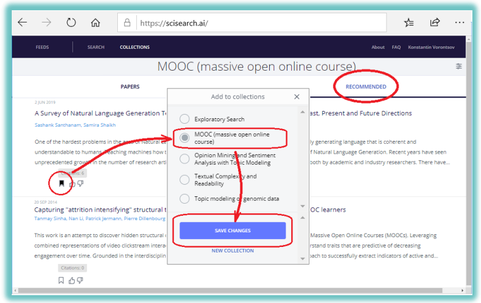
«Разработанная в рамках данного проекта модель уже широко используется в **Научной электронной библиотеке** для решения целого ряда задач, связанных с оценкой тематической близости научных документов. Уже протестирован специалистами полезный сервис для ученых, позволяющий *для заданной статьи или подборки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

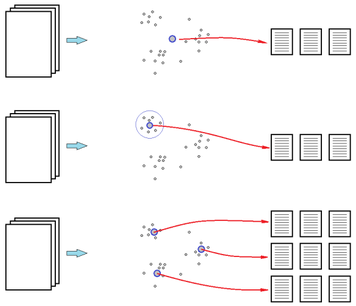
— Геннадий Еременко, генеральный директор НЭБ

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

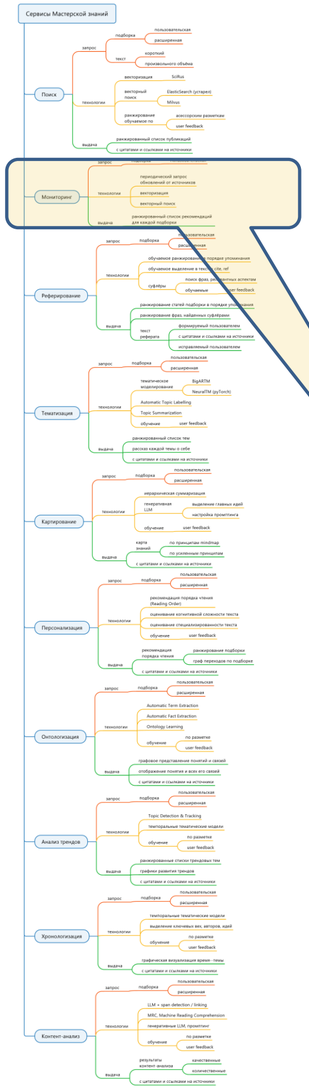
Концепция сервисов «Мастерской знаний»

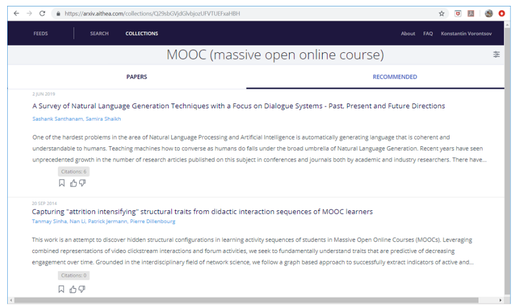






Концепция сервисов «Мастерской знаний»

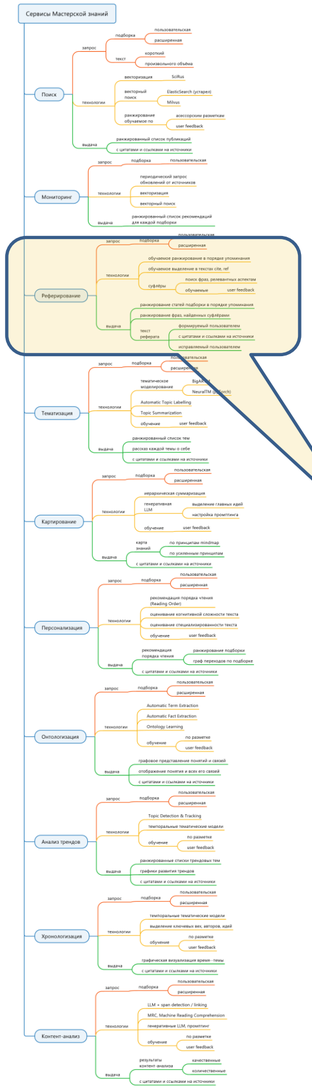


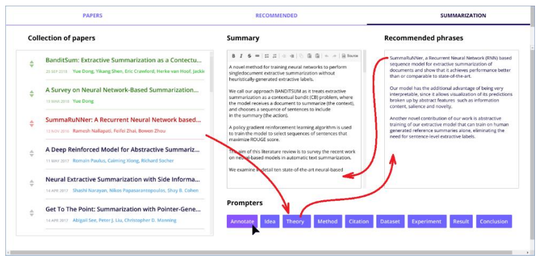


Мониторинг

- запрос
  - подборка
    - пользовательская
- технологии
  - периодический запрос обновлений от источников
  - векторизация
  - векторный поиск
- выдача
  - ранжированный список рекомендаций для каждой подборки

Концепция сервисов «Мастерской знаний»

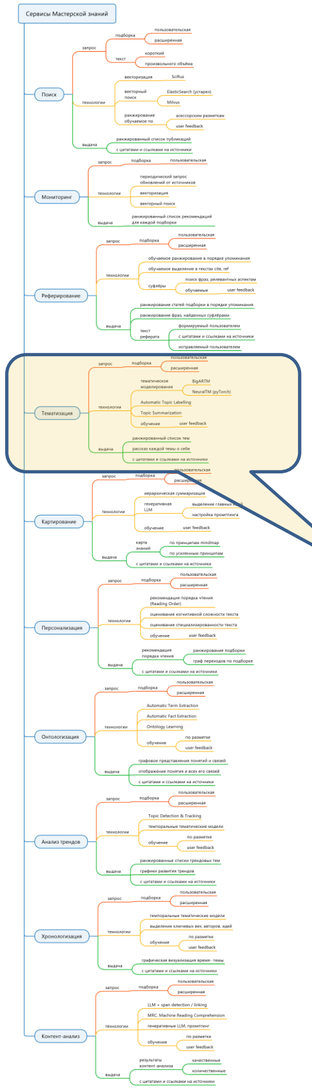




Реферирование

- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - обучаемое ранжирование в порядке упоминания
  - обучаемое выделение в текстах cite, ref
  - суфлёры
    - поиск фраз, релевантных аспектам
    - обучаемые — user feedback
- выдача
  - ранжирование статей подборки в порядке упоминания
  - ранжирование фраз, найденных суфлёрами
  - текст реферата
    - формируемый пользователем
    - с цитатами и ссылками на источники
    - исправляемый пользователем

Концепция сервисов «Мастерской знаний»

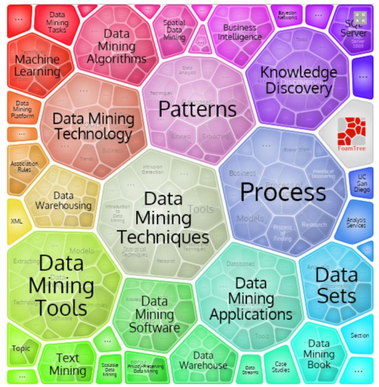


Тематизация

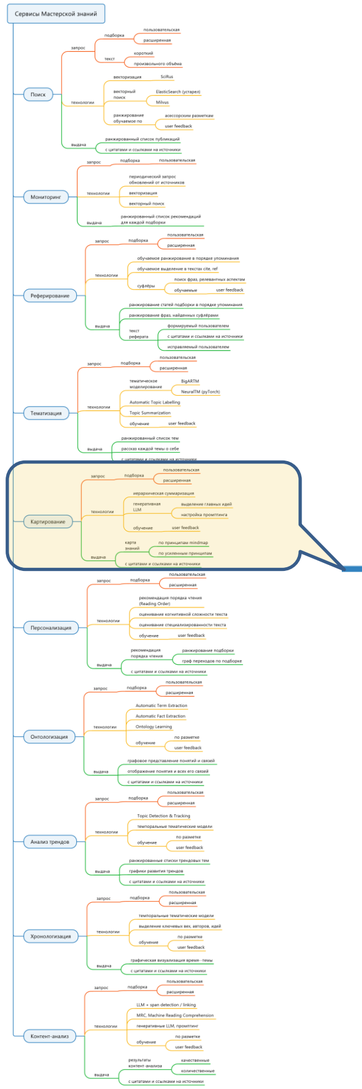
запрос
- подборка
  - пользовательская
  - расширенная

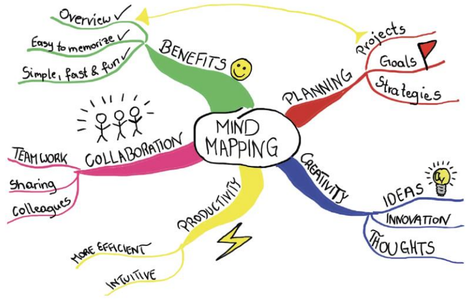
технологии
- тематическое моделирование
  - BigARTM
  - NeuralTM (pyTorch)
- Automatic Topic Labelling
- Topic Summarization
- обучение
  - user feedback

выдача
- ранжированный список тем
- рассказ каждой темы о себе
- с цитатами и ссылками на источники

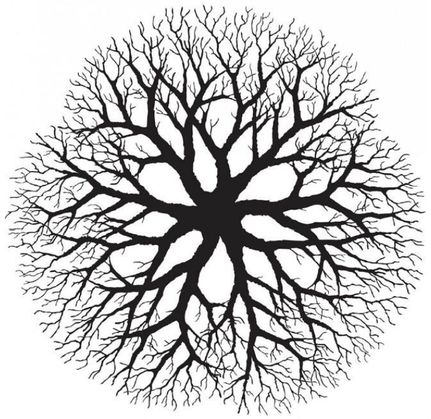


Концепция сервисов «Мастерской знаний»



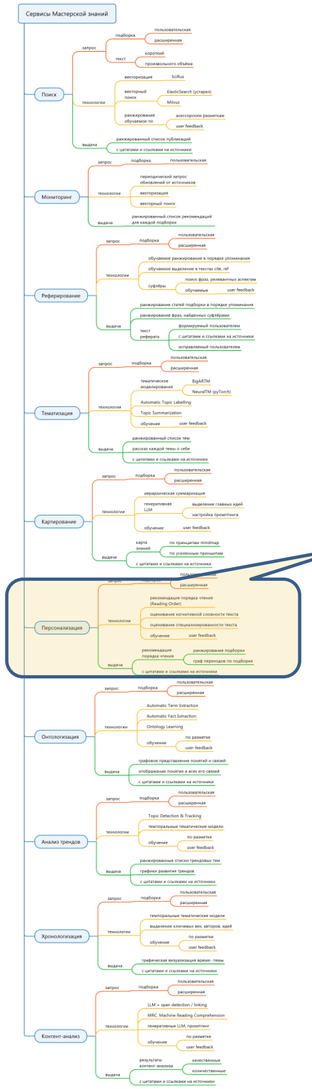


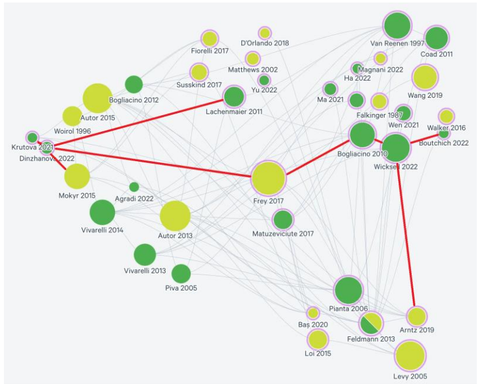
Картирование



- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - генеративная LLM
    - иерархическая суммаризация
    - выделение главных идей
    - настройка промптинга
  - обучение
    - user feedback
- выдача
  - карта знаний
    - по принципам mindmap
    - по усиленным принципам
  - с цитатами и ссылками на источники

Концепция сервисов «Мастерской знаний»

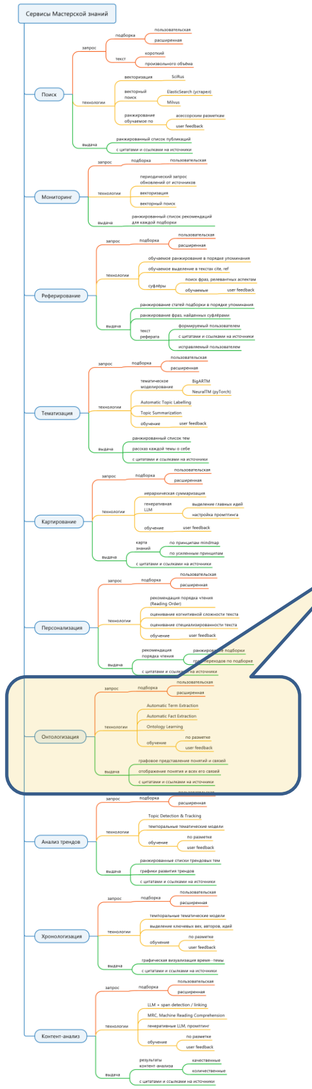


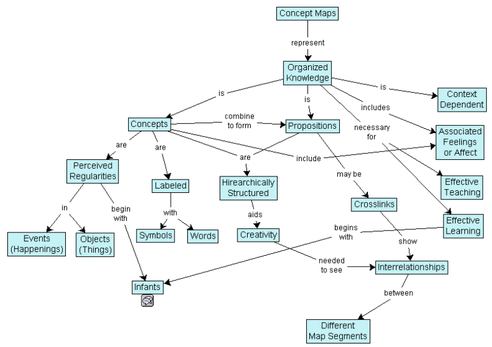


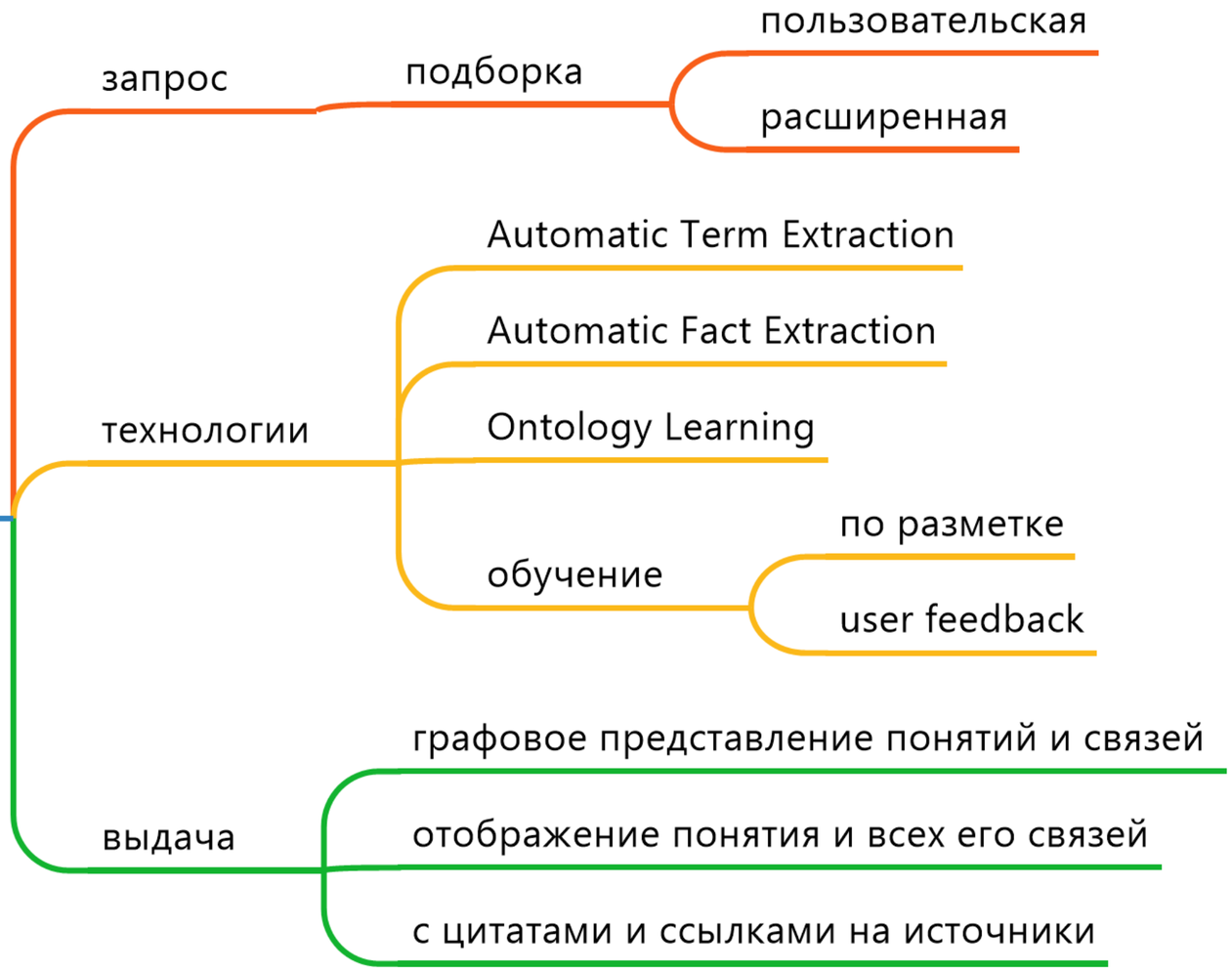
Персонализация

- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - рекомендация порядка чтения (Reading Order)
  - оценивание когнитивной сложности текста
  - оценивание специализированности текста
  - обучение — user feedback
- выдача
  - рекомендация порядка чтения
    - ранжирование подборки
    - граф переходов по подборке
  - с цитатами и ссылками на источники

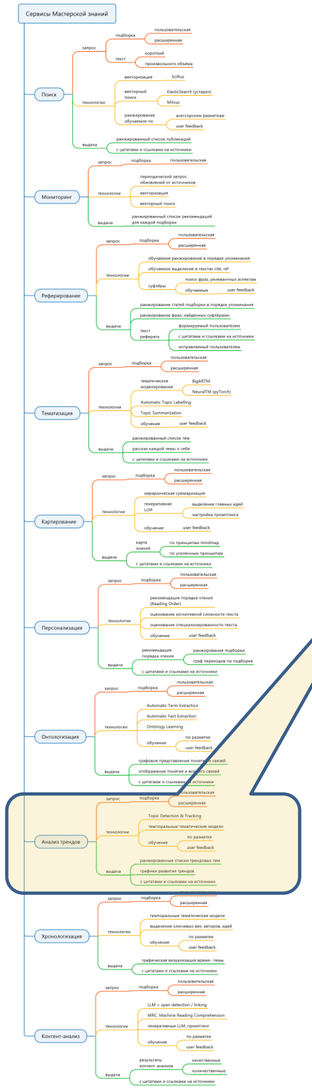
Концепция сервисов «Мастерской знаний»







Концепция сервисов «Мастерской знаний»

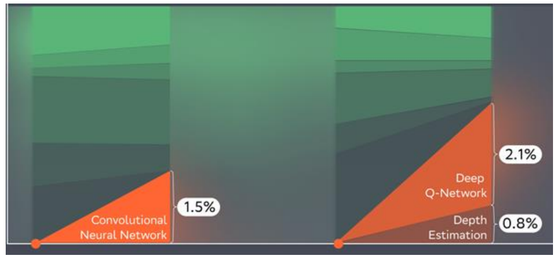


Анализ трендов

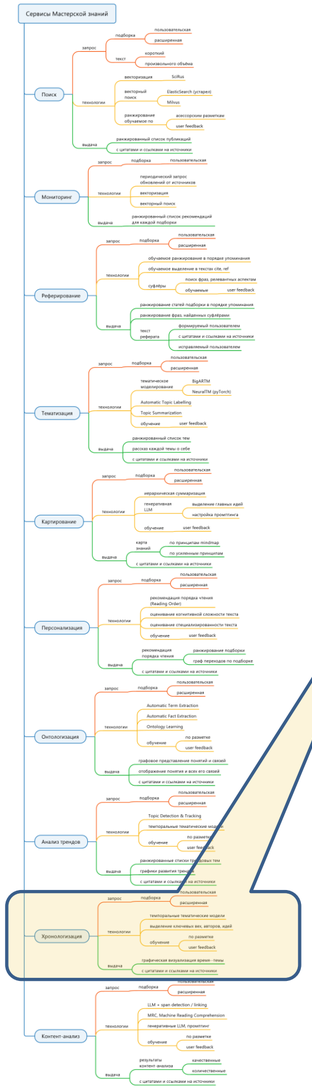
запрос
- подборка
  - пользовательская
  - расширенная

технологии
- Topic Detection & Tracking
- темпоральные тематические модели
- обучение
  - по разметке
  - user feedback

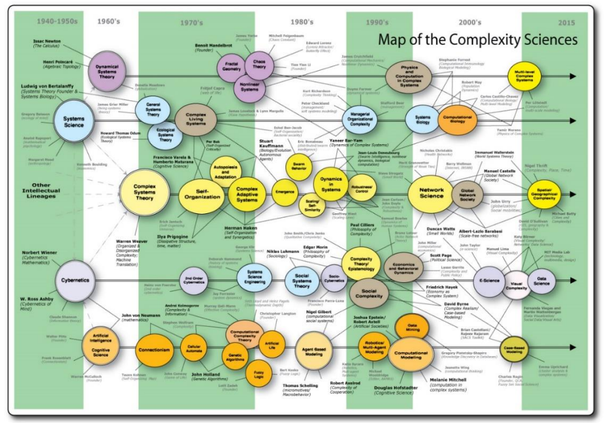
выдача
- ранжированные списки трендовых тем
- графики развития трендов
- с цитатами и ссылками на источники



Концепция сервисов «Мастерской знаний»

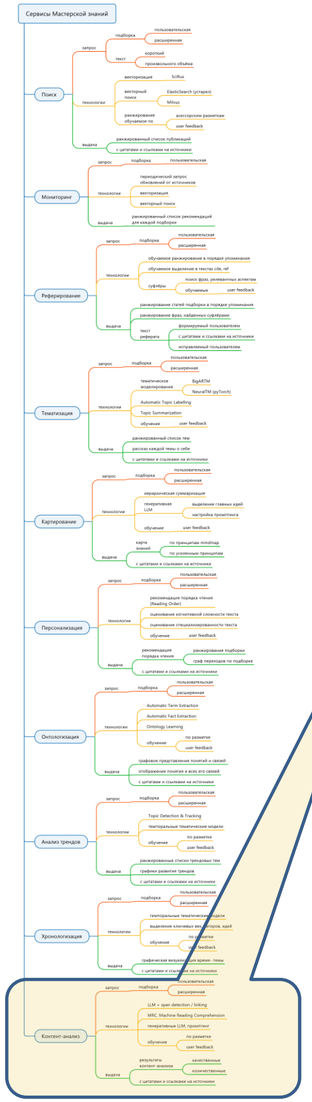


Хронологизация



- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - темпоральные тематические модели
  - выделение ключевых вех, авторов, идей
  - обучение
    - по разметке
    - user feedback
- выдача
  - графическая визуализация время--темы
  - с цитатами и ссылками на источники

Концепция сервисов «Мастерской знаний»



Контент-анализ

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.

- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - LLM + span detection / linking
  - MRC, Machine Reading Comprehension
  - генеративные LLM, промптинг
  - обучение
    - по разметке
    - user feedback
- выдача
  - результаты контент-анализа
    - качественные
    - количественные
  - с цитатами и ссылками на источники

Внимание, опрос!

Целевая аудитория — исследователи, преподаватели, студенты

- Отранжируйте предлагаемые сервисы «Мастерской знаний» по их полезности для работы с научными публикациями
- Какие из этих видов работы Вы уже делаете?
- Какие системы Вы для этого используете?
- Сколько времени это занимает?
- Какой ещё сервис Вам необходим, опишите его как BB (вход-выход)
- Укажите его позицию в ранжированном списке сервисов
- Если «Мастерская знаний» будет реализована в полном объёме и удобно, какую долю рабочего времени Вы бы в ней проводили?

Мастерская знаний

Миссия: устранять барьеры между человеком и знанием

Реализовано: кросс-языковой поиск текстов, схожих по смыслу

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

Планы:

- **развитие сервисов:** поиск, мониторинг, реферирование, тематизация, онтологизация, хронологизация, картирование, персонализация, анализ трендов, контент-анализ
- **источники:** проектная документация, патенты, новости
- **мультиязычность:** русский—английский—китайский—…

Спасибо за внимание!

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. лабораторией машинного обучения и семантического анализа Института ИИ МГУ,

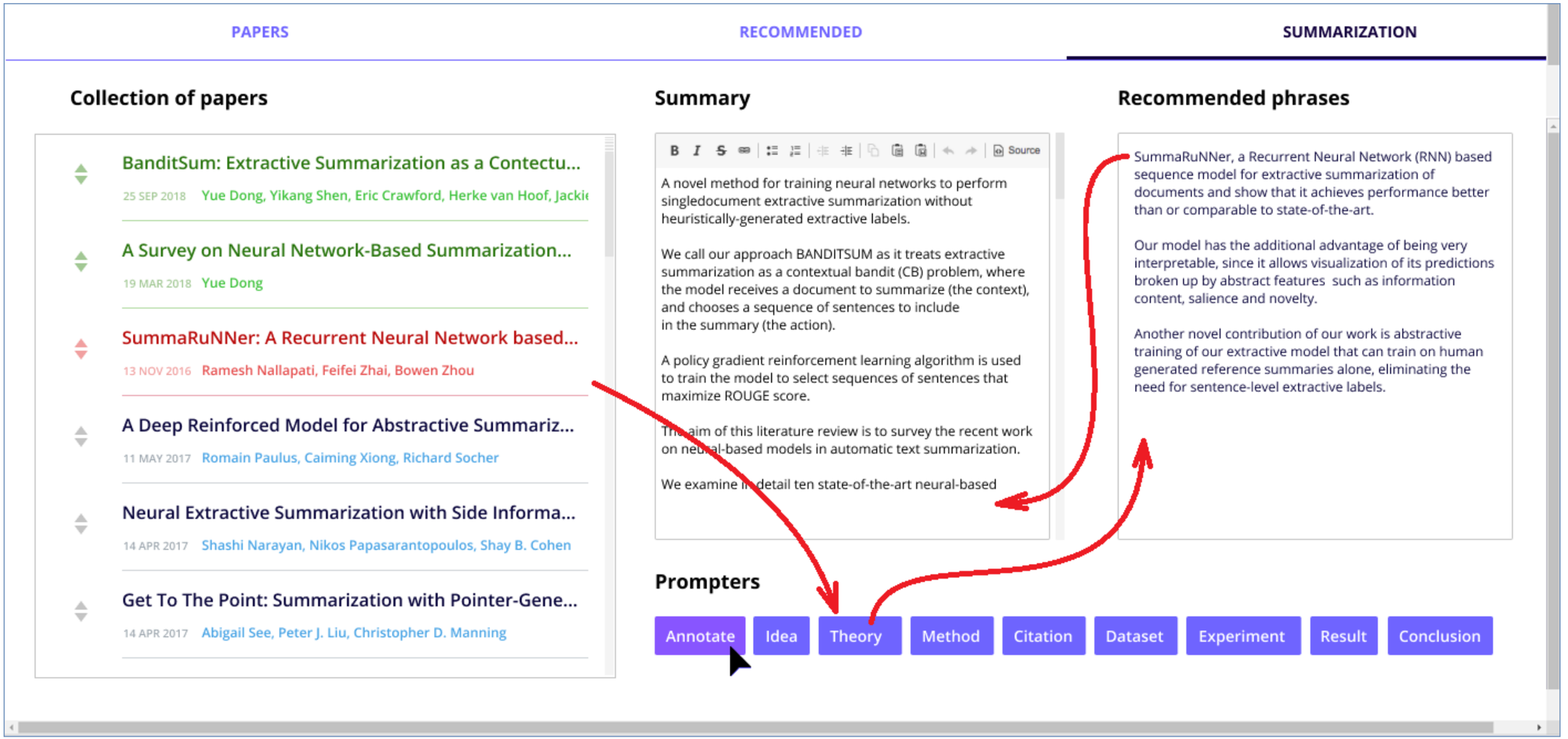
зав. кафедрой ММП ВМК МГУ

k.vorontsov@iai.msu.ru

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

Дополнения

Полуавтоматическое реферирование подборки



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

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

Полуавтоматическое реферирование подборки

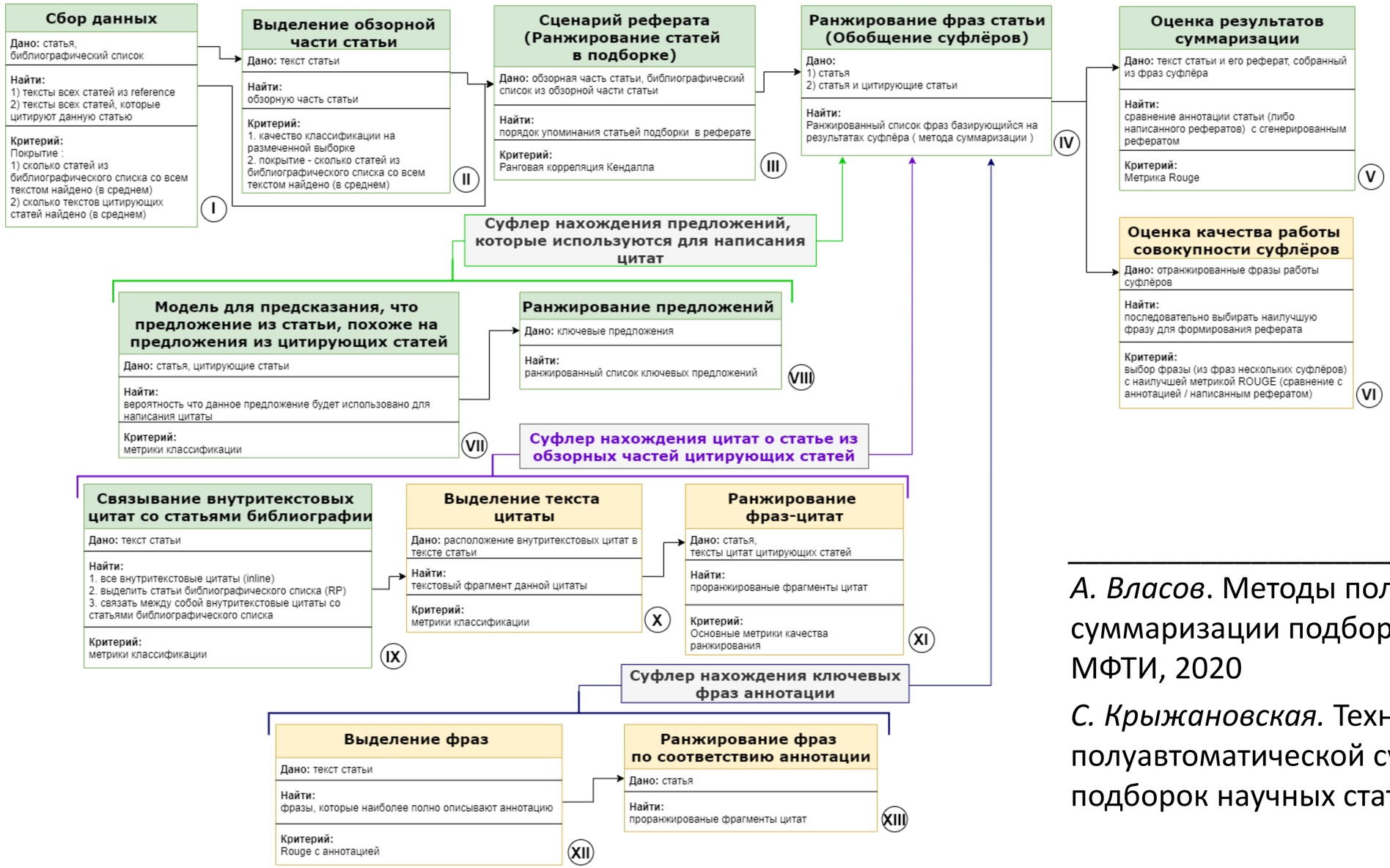
Основные задачи машинного обучения:

- 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.

M.Yasunaga, J.Kasai, R.Zhang, A.Fabbri, I.Li, D.Friedman, D.Radev. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. 2019.

Полуавтоматическое реферирование подборки



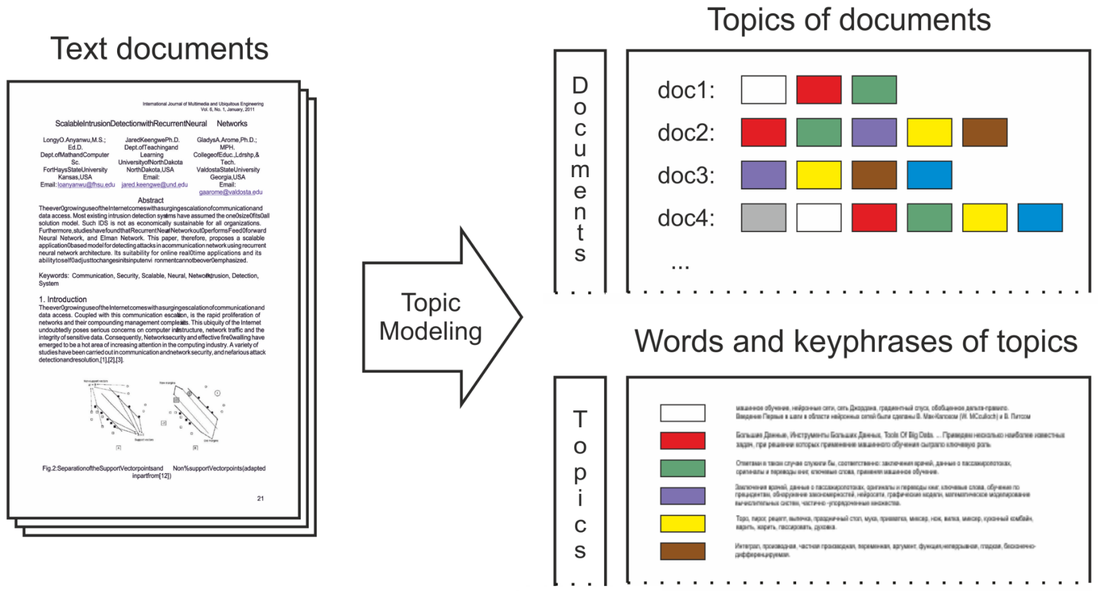
А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

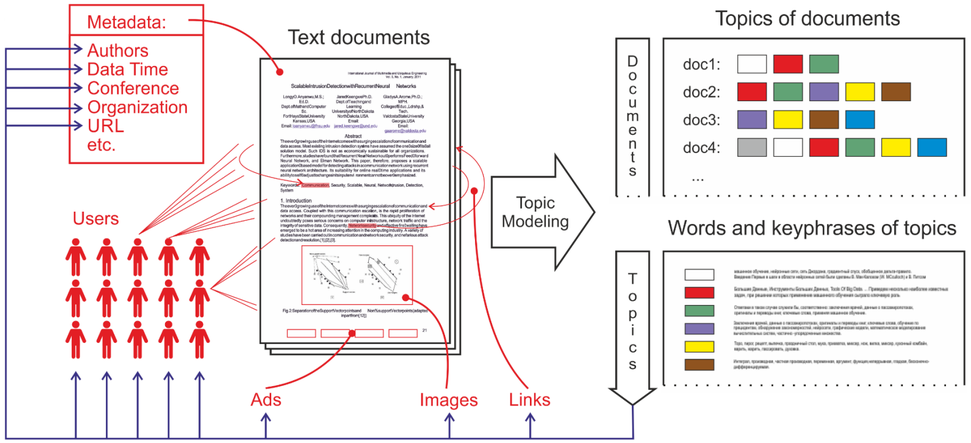
С. Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

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
- Онлайновый 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

3.7M статей Википедии, 100K 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)

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:

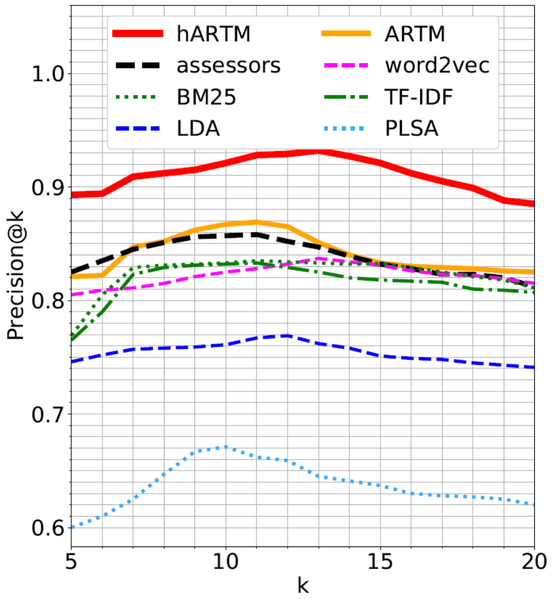
- длинные запросы (1 стр. А4)
- 100 запросов на коллекцию
- 3 асессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:

(170K Russian docs)

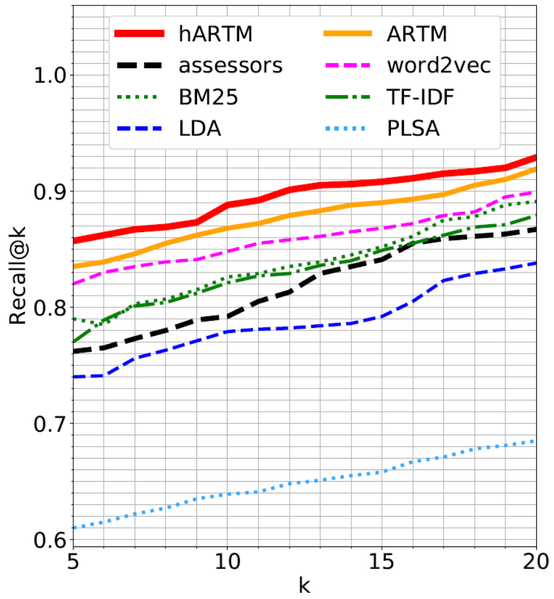
(750K English docs)

Оценки качества поиска:

точность (precision@k)



полнота (recall@k)



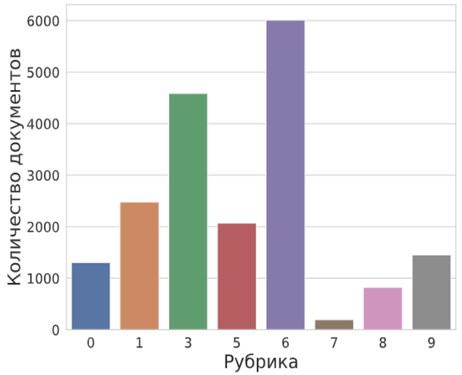
Ianina A., Golitsyn L., Vorontsov K. Multi-objective topic modeling for exploratory search in tech news. AINL 2017.

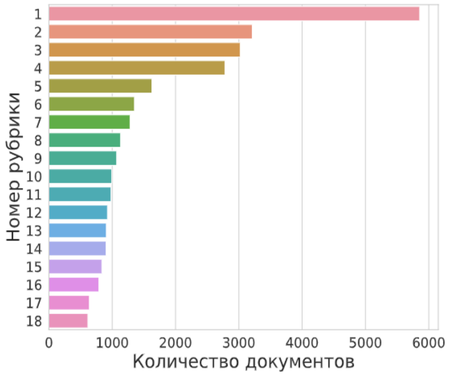
Ianina A., Vorontsov K. Regularized multimodal hierarchical topic model for document-by-document exploratory search. 2019.

Мультиязычный тематический поиск и категоризация

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

94%

Точность поиска

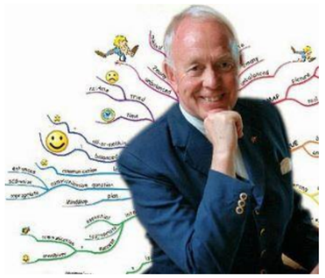
Особенности решения:

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

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

Карты знаний:
базовые принципы интеллект-карт (mind map)

предложены в 70-е годы британским **психологом** Тони Бьюзеном



способ визуализации того, как темы (мысли, идеи) разбиваются на подтемы иерархически

графическое оформление

активация зрительной памяти[42]

радиантность: линии расходятся из центра

размер шрифта отражает важность

цвет выделяет поддеревья

картинки усиливают образность

дополнительные элементы

ассоциативные связи между темами

комментарии, выноски, теги, (гипер)ссылки

техника запоминания

посмотреть, понять, обсудить, принять

самостоятельно воспроизвести через 10 минут → сутки → неделю → месяц

Карты знаний:
+11 известных принципов

позже были дополнены различными **принципами**

в зависимости от практических потребностей, целей и задач

ветвления

значимость: (под)темы отбираются и ранжируются по важности

однородность:
- подтемы образуют сюжет, нарратив
- либо отвечают на общий вопрос

полнота: подтемы охватывают все аспекты темы

точность: среди подтем невозможно выделить лишнюю

компактность: у темы 5±3 подтем (число Ингве-Миллера 7±2)

эргономичности

наглядность: слова подкрепляются изображениями

лаконичность: темы формулируются максимально кратко

обозримость: карту понимают и запоминают целиком

эстетичности

красота, живость: эмоциональные карты лучше запоминаются

гармоничность: впечатление целостности, складности карты

сбалансированность: ветви примерно равны и равноценны

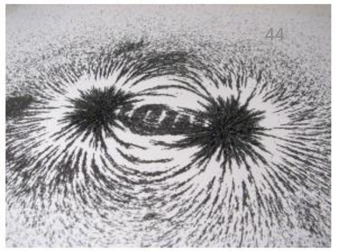
Карты знаний:
+6 новых принципов

(1) глобальная радиантная связность

компромисс с обозримостью

всех карт через ключевые понятия в единую **Систему Знаний**

в центре находится **смысловое ядро**



естественно-научное, цивилизационное

знания, которые важны всегда и для всех

критерии важности тем: *что в теме главное?* *для чего?* *для кого?*

метафора:
источник силовых линий, по которым ранжируется семантическое поле карты

(2) отторгаемость

компромисс с лаконичностью

комментарии автора не обязательны для понимания карты

карта способна «жить своей жизнью»

Карты знаний:
+6 новых принципов

(3) коллективность,
на всех этапах
жизненного цикла

компромиссы между авторами

создание
- рецензирование, согласование

развитие
- уточнение, реструктуризация
- детализация, разрастание

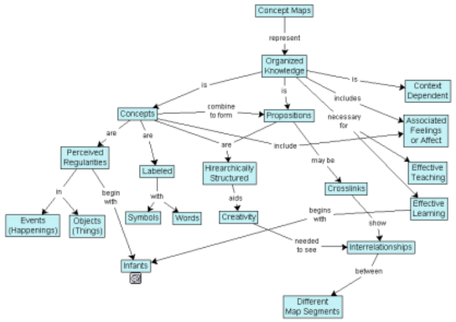
применение
- в практической деятельности
- с разграничением прав доступа

(4) читабельность

компромисс с лаконичностью
и обозримостью

любой фрагмент карты читается
как связный текст, нарратив
- легко и однозначно
- даже автоматически

в отличие
от других техник
представления
знаний



- онтологий
- фреймов и др.

Карты знаний:
+6 новых принципов

(5) сворачиваемость

компромисс с читабельностью

- любой темы без утраты
  - читабельности
  - сбалансированности
- позволяет «отложить на потом» любую детализацию
- способствует
  - выделению главного в каждой теме
  - пониманию и взаимопониманию

(6) возможность машинной обработки

компромисс с антропоцентричностью

- на этапе создания карт:
  - подбор
    - источников, ссылок
    - картинок по контексту
  - суммаризация текстов в виде карт знаний
- на этапе чтения: автоматическое
  - сворачивание карты по слайдам
  - преобразование карты в нарратив
- обучение по картам знаний больших языковых моделей
  - думающих как люди
  - безопасных для людей

Выводы про карты знаний

- **Универсальный инструмент мышления** для человека и машины.
- **Перспективный инструмент** «коллективного разума»

- **Важные навыки** для работы с научной информацией
  — во всём выделять главное (7±2),
  — делать это быстро, формулировать лаконично
- **Прежде чем обучать ИИ** по тексто-графическим представлениям,
  — необходимо освоить их самим,
  — в своей практической деятельности,
  — в том числе коллективной

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**

Стат < Син < Син+Стат < Тем < Стат+Тем, Син+Тем < 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

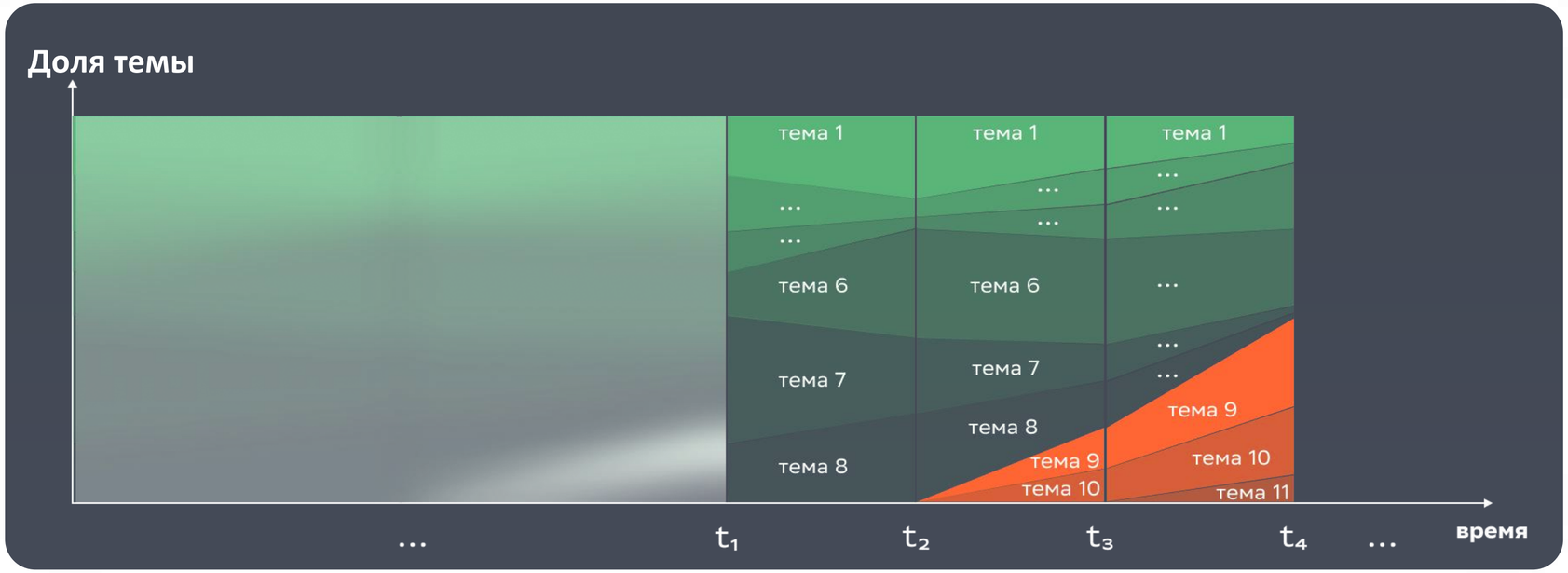
Особенности соревнования:

- 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

Конференция по компьютерной лингвистике Диалог https://www.dialog-21.ru/evaluation/
https://github.com/mlsa-iai-msu-lab/cl-ruterm3 Телеграм-канал: https://t.me/rutermeval

Поиск 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



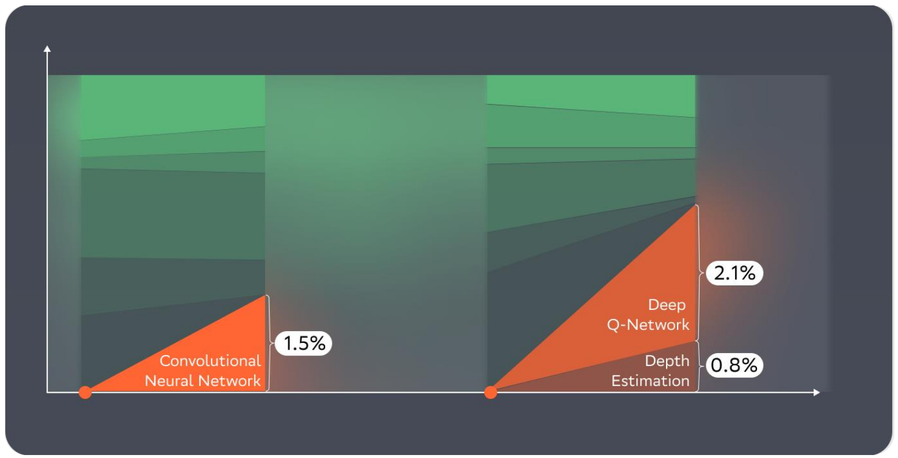
Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В.
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. Математика, информатика, процессы управления, 2022, том 508, С.106–108

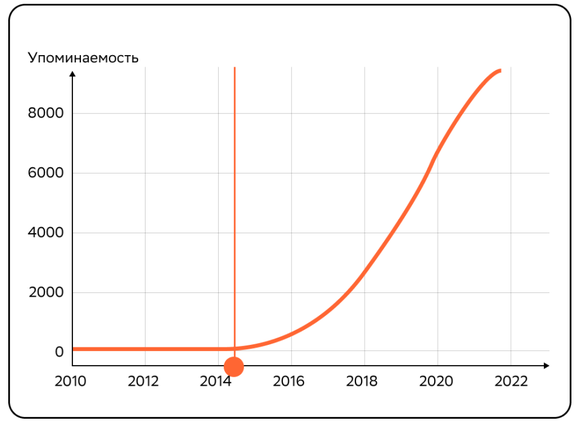
Поиск научных трендов

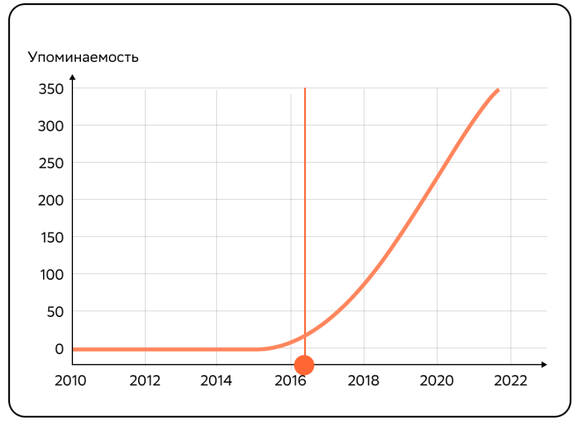
Трендовая тема:

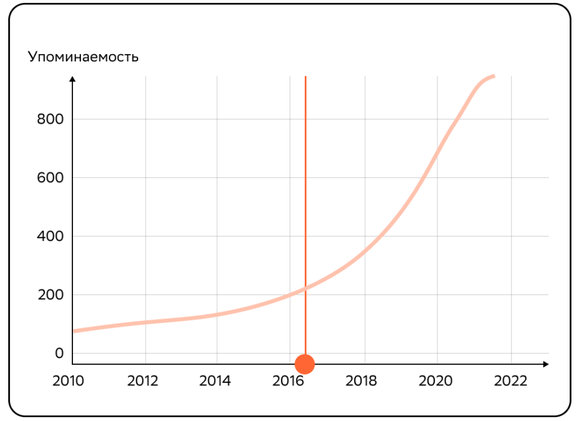
- наличие семантического ядра
- наличие быстрого (обычно экспоненциального) роста

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

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

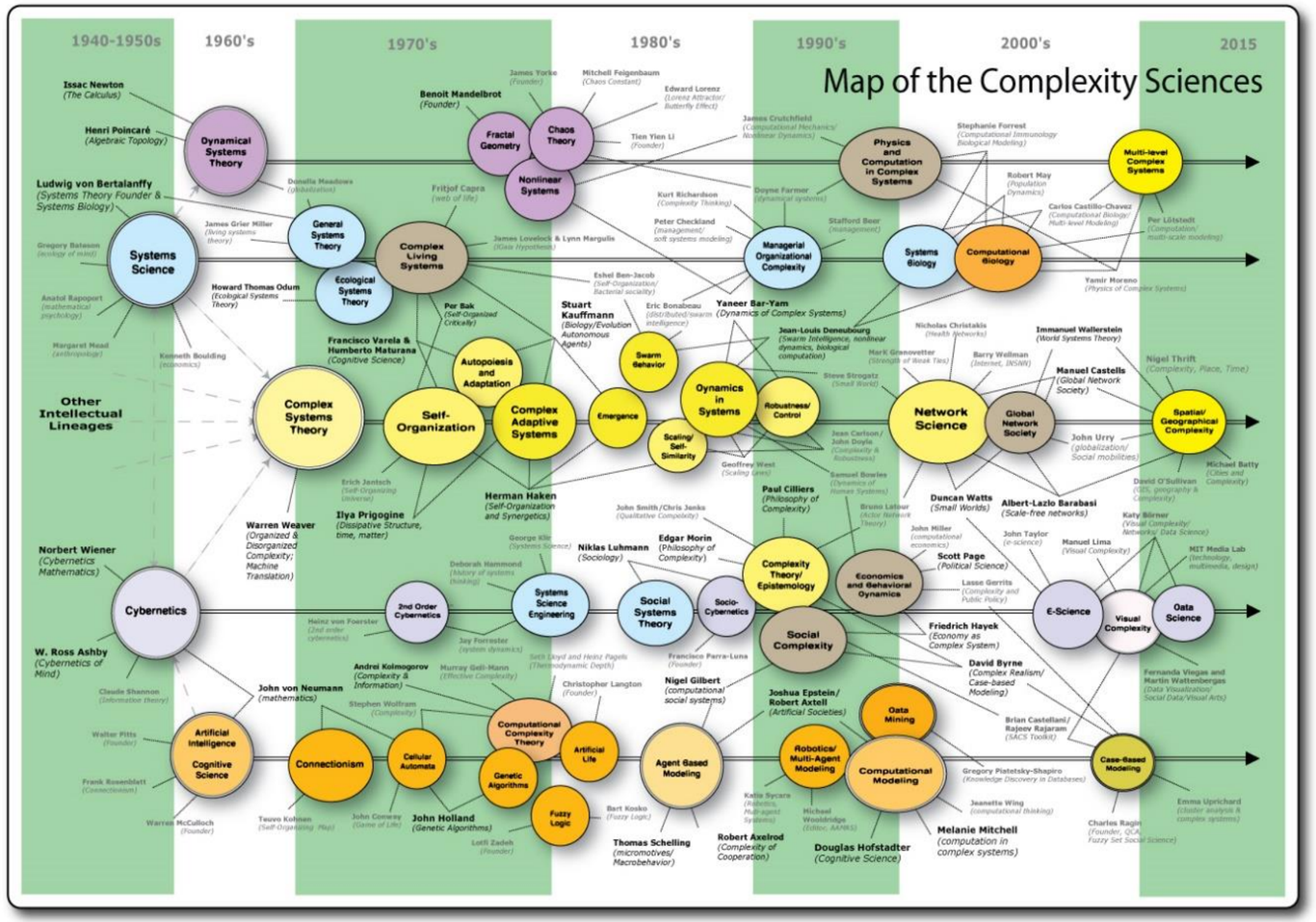
Поиск научных трендов: примеры тем

StyleGAN	Meta Learning	NERF
stylegan	meta model	neural radiance field
latent code	meta train	accurate depth estimation
mapping network	meta optimization	additional qualitative result
ablation study	meta update	novel loss function
text generation	meta testing	optical flow prediction
generation quality	training task	image reconstruction loss
generator architecture	continual learning	monocular depth prediction
mask	previous task	geometric consistency loss
encoder	catastrophic forgetting	depth estimation method
gan model	ablation study	optical flow network

Хронологизация: карта направлений и вех развития

Осями на карте могут быть:

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



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

1) выделить фрагмент
2) классифицировать (тегировать) фрагмент по рубрикатору
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 рубрикатора 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.

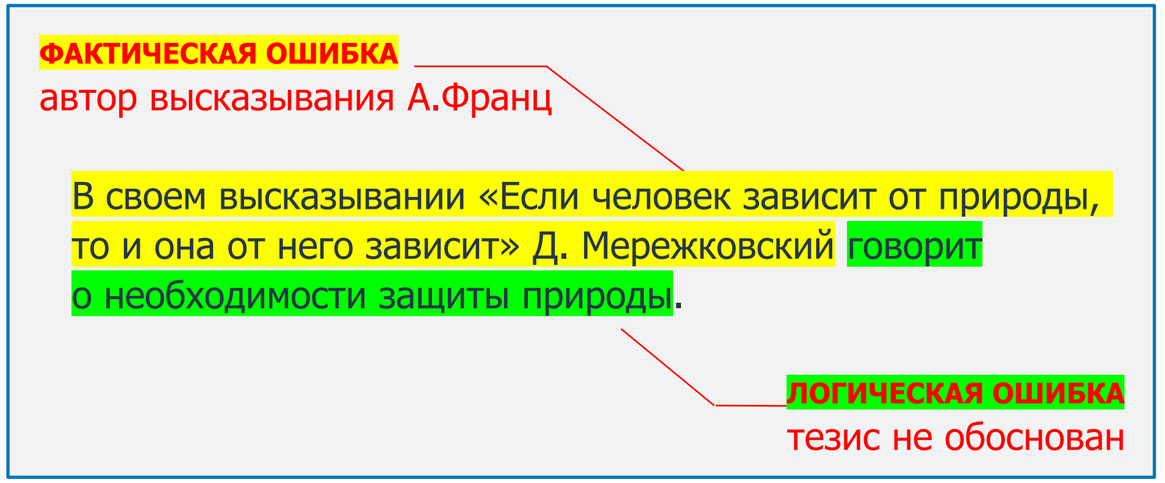
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

Призовой фонд: ₽100M русский язык + ₽100M 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.



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

Сравнение разметки, сгенерированной алгоритмом, с разметкой эксперта

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

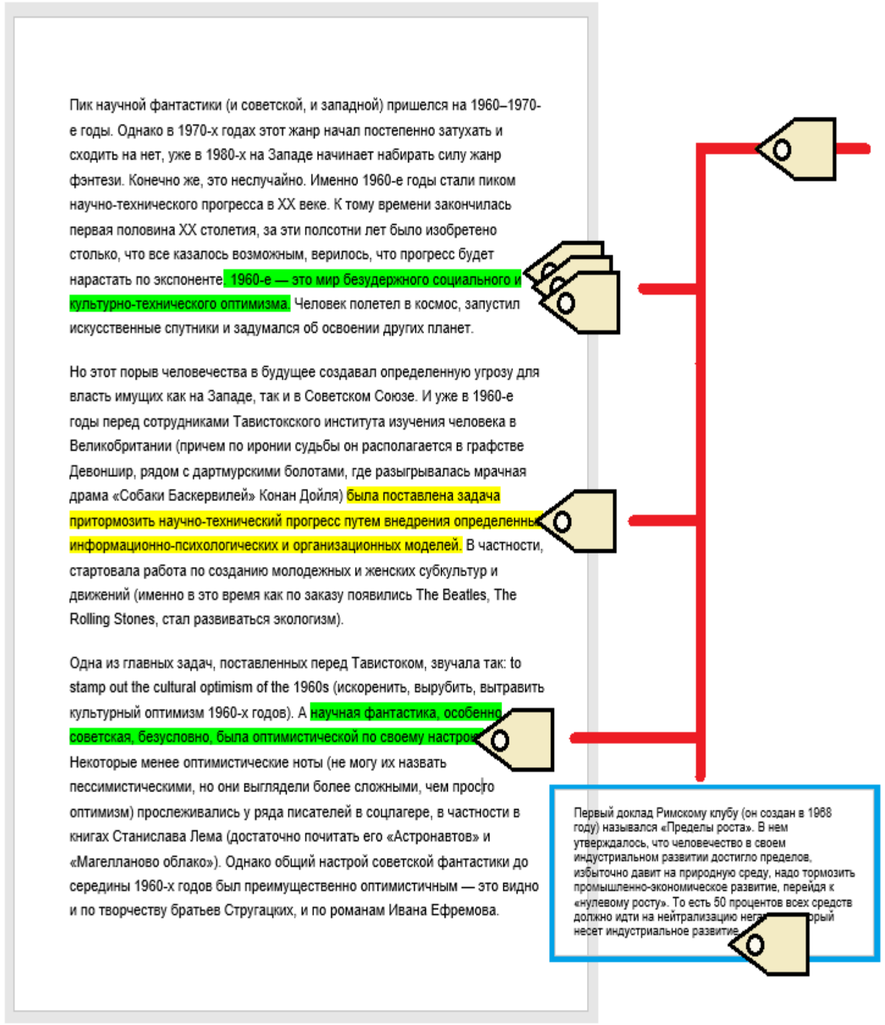
Commissio
PopulusQue
Europaea

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

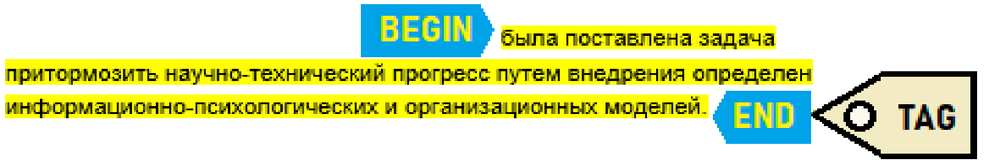


Разметка состоит из элементов

Элемент разметки — несколько взаимо-связанных фрагментов, затекстов и тегов

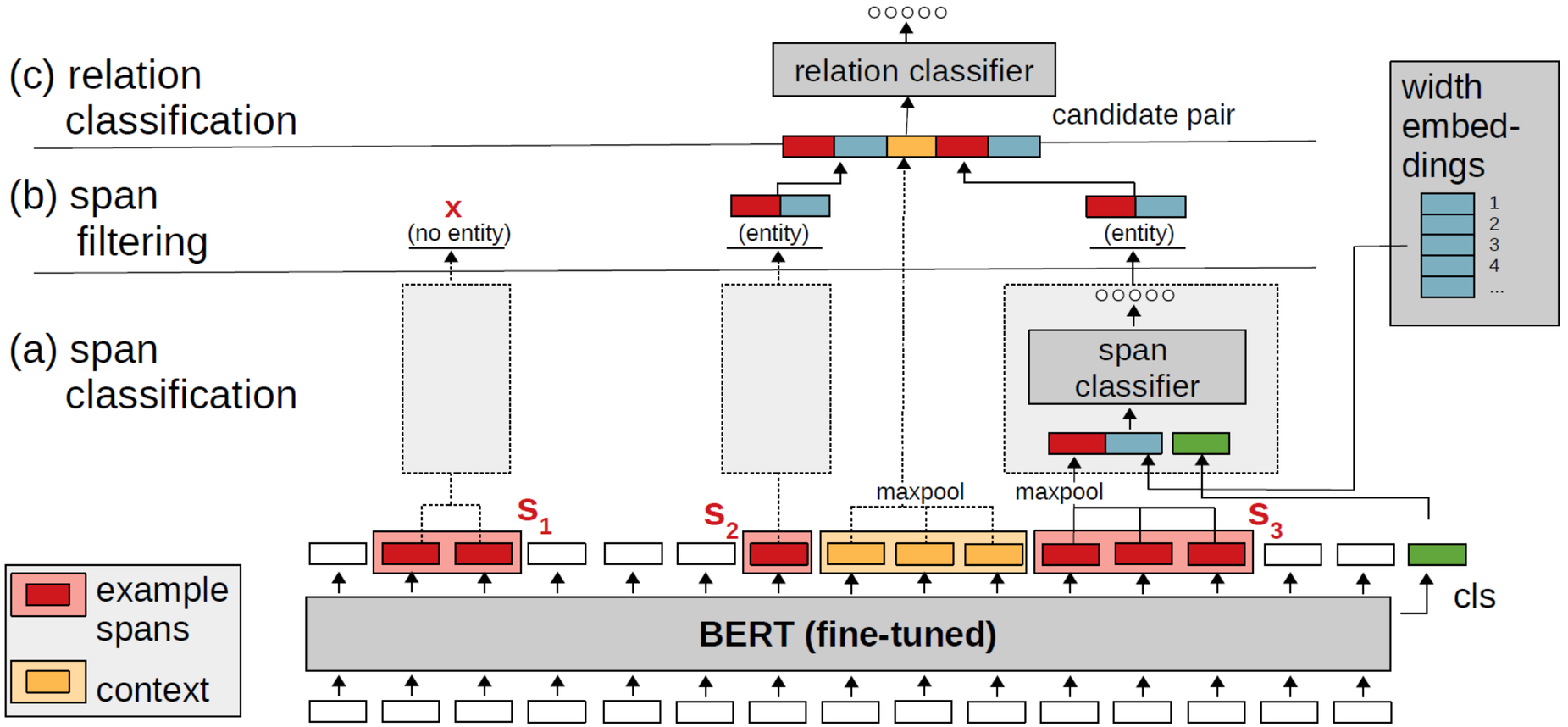
Теги (классы) выбираются из рубрикатора

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

Нейросетевые обучаемые модели разметки

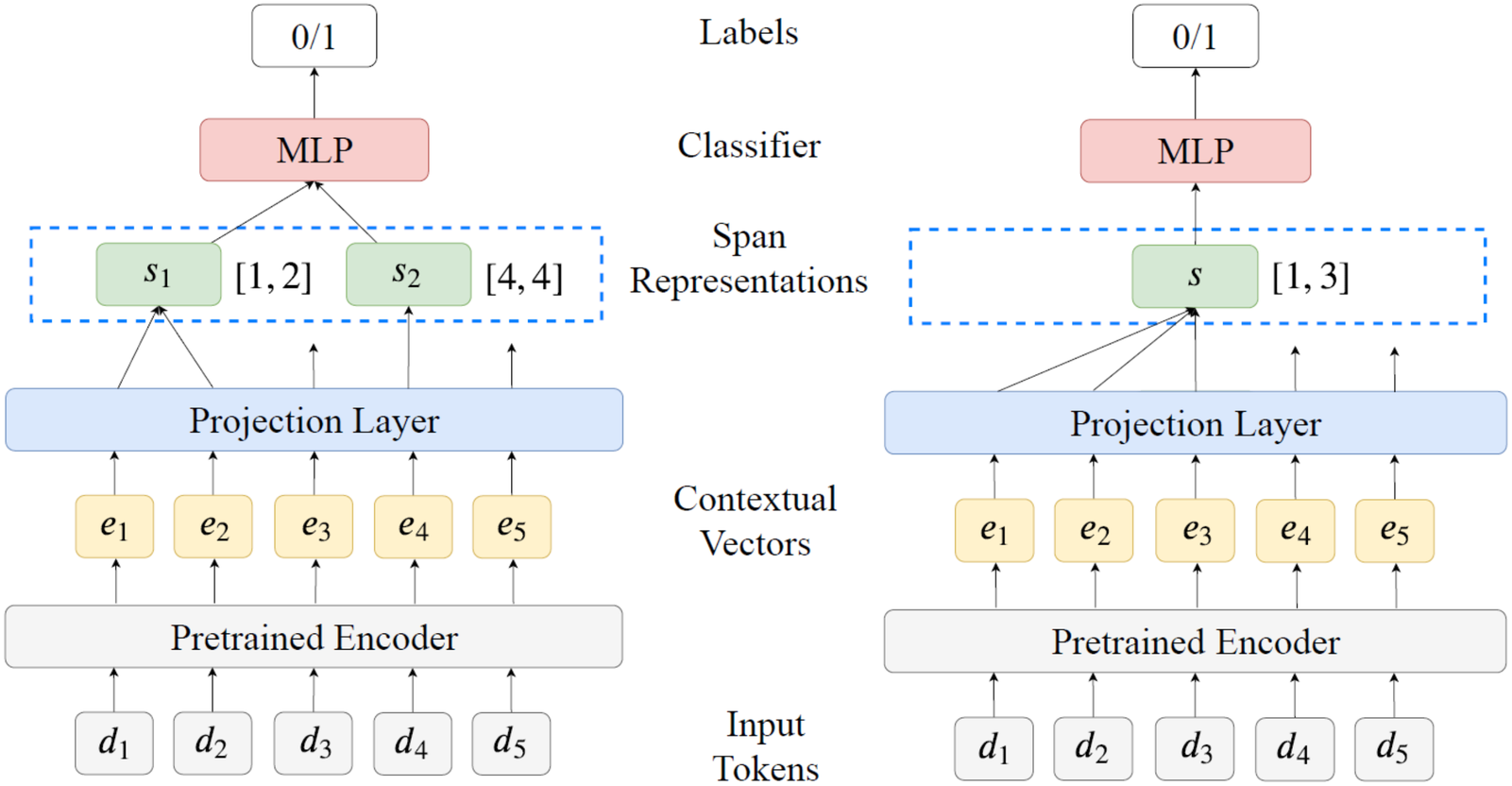


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

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.

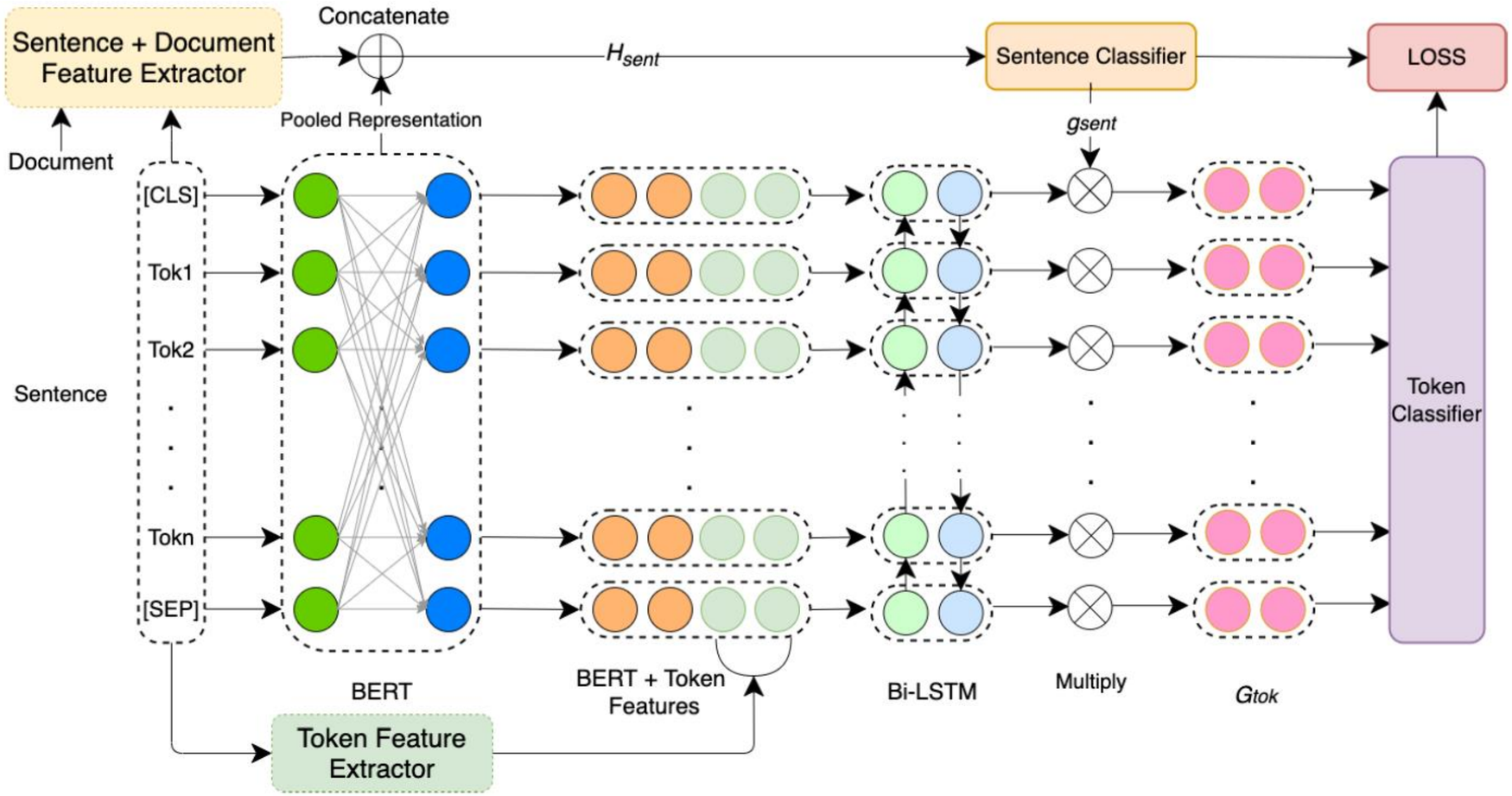
Нейросетевые обучаемые модели разметки



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



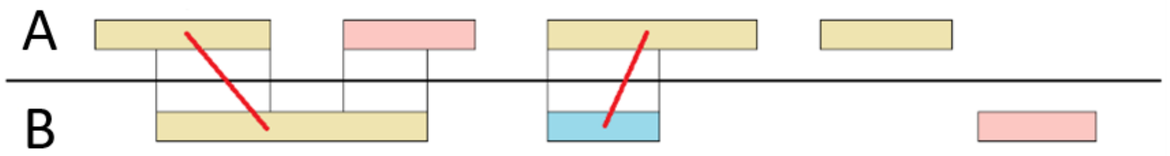
Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

Методика оценивания алгоритмической разметки

- В основе методики — парное сравнение разметок текста:
  «алгоритм ↔ эксперт», «эксперт-1 ↔ эксперт-2»
  на основе оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок Con1,…,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

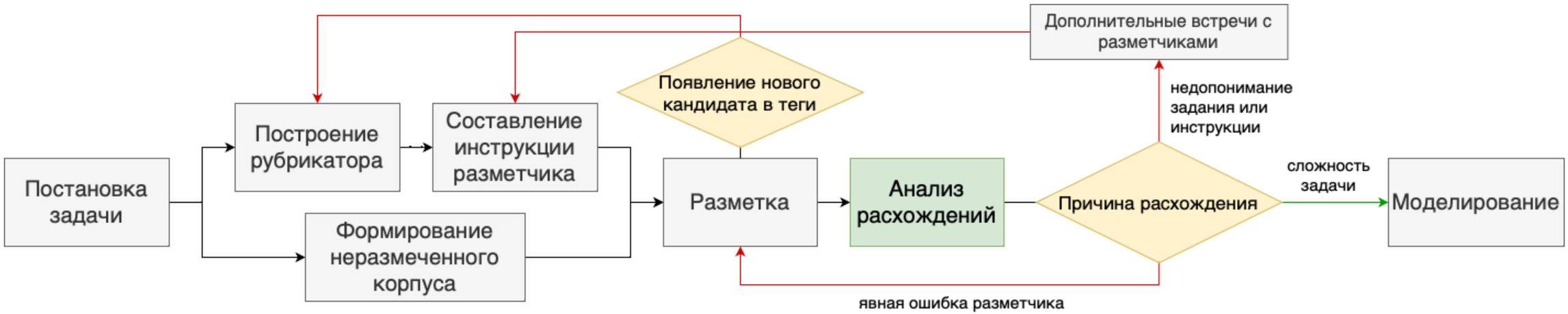
Con2 = точность наложения сопоставленных фрагментов

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