22-я конференция «Математические методы распознавания образов» (ММРО-2025),
посвященная 90-й годовщине со дня рождения академика Юрия Ивановича Журавлева (1935-2022)
Муромский институт ВлГУ в г. Муром, 22-26 сентября 2025 г.

Мастерская знаний:
как большие языковые модели меняют подходы к поиску 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

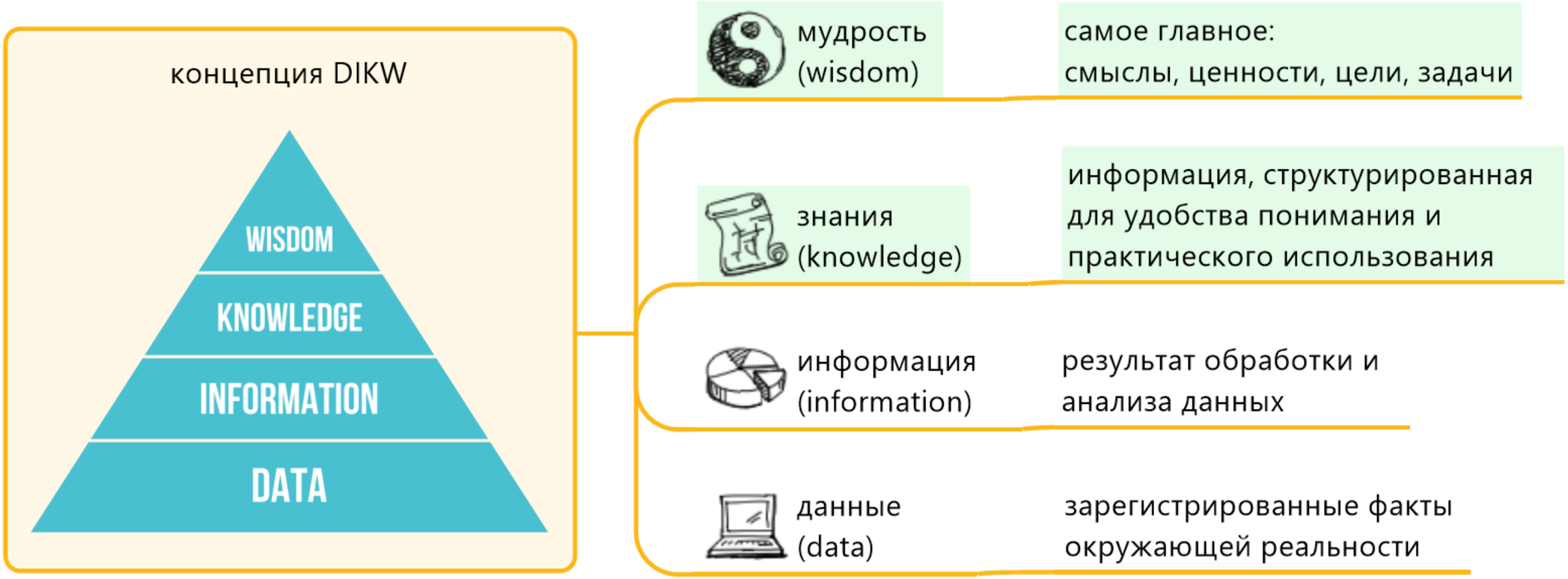
Концепция «Мастерской знаний»

«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*)

Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

Что такое «знания»



Технологии больших языковых моделей (Large Language Model, LLM) позволяют выделять знания и идеи из текста и систематизировать их

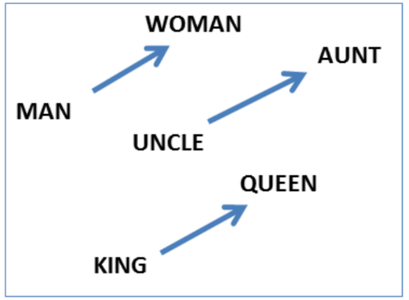
Эволюция подходов в обработке естественного языка

Декомпозиция задач по уровням «пирамиды NLP»

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

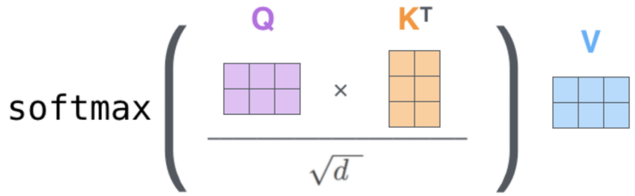
Модели векторных представлений слов (эмбедингов)

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



Нейросетевые большие языковые модели (БЯМ, LLM)

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



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 текстовые документы – чтобы их сохранять и накапливать
- накапливаю – чтобы их перечитывать, анализировать, понимать
- понимаю – чтобы получать, обрабатывать, систематизировать *знания*
- систематизирую – чтобы применять и передавать *знания и мудрость*

Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

Научный поиск на основе LLM и ИИ-агентов

perplexity

Consensus

NotebookLM

SEMANTIC SCHOLAR

Elicit

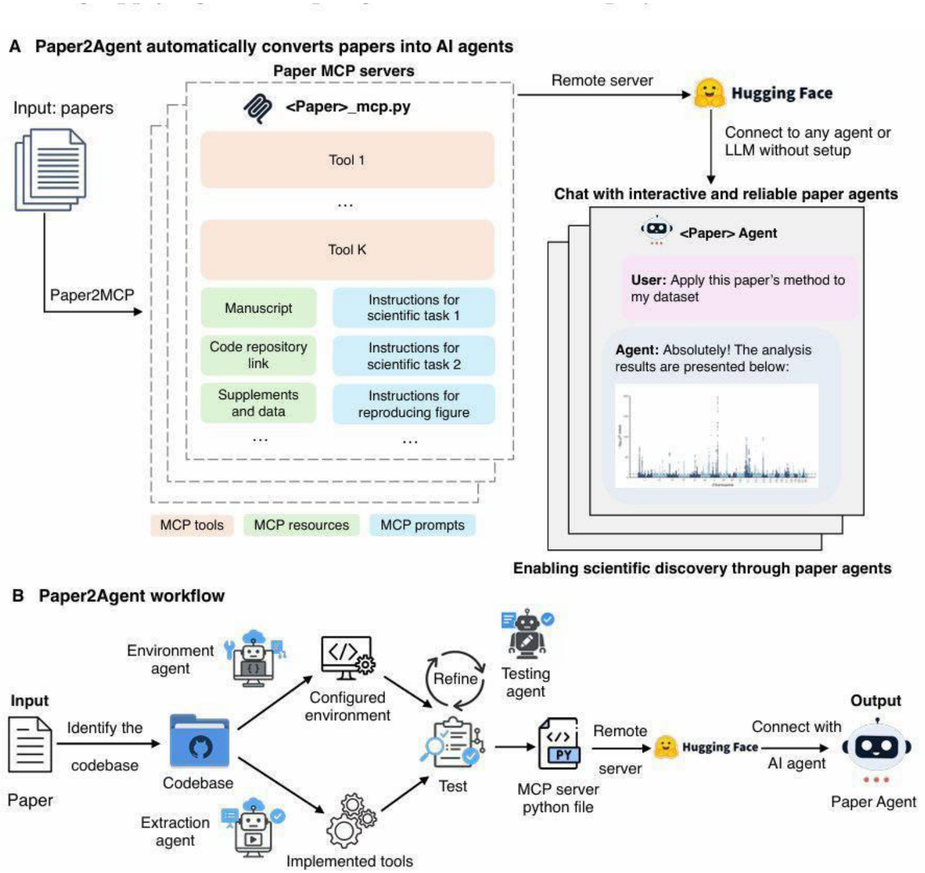
ChatPDF

ResearchRabbit

Недостатки ИИ-систем:

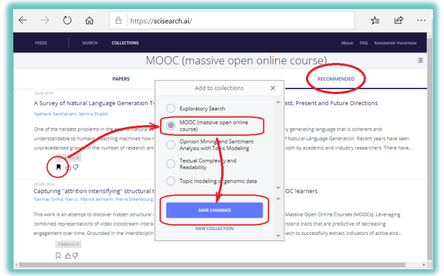
- как зафиксировать долгосрочный тематический поисковый интерес?
- как актуализировать знания?
- как обеспечить ясность представления знаний — «посмотрел и всё понял»?
- как включить «коллективный разум»?

Paper2Agent — интерактивный ИИ-агент
https://github.com/jmiao24/Paper2Agent



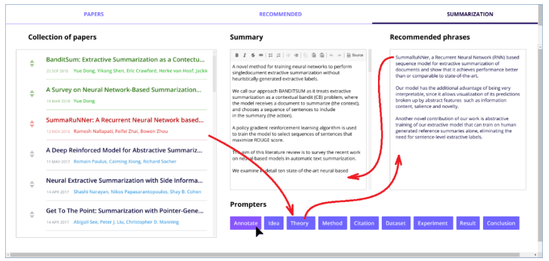
Концепция сервисов «Мастерской знаний»

Подборка текстов фиксирует тематику поискового интереса пользователя или группы
Расширенная подборка — подборка + семантически близкие тексты



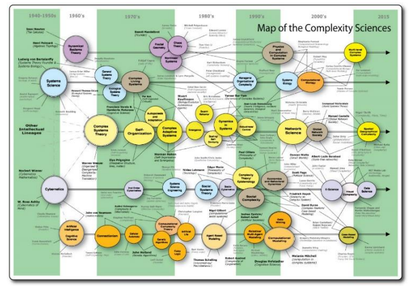
Поисково-рекомендательные сервисы:

- 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



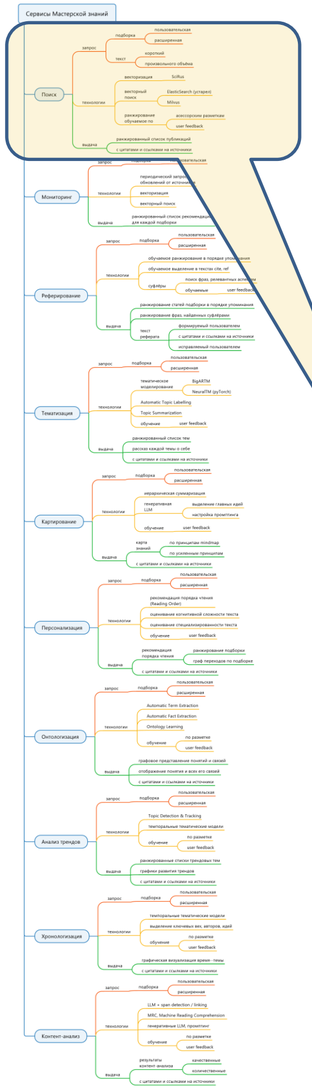
Коммуникативные сервисы:

- совместное составление, анализ, использование подборок
- инструментализация коллективного анализа подборки

Декларация принципов «Мастерской знаний»

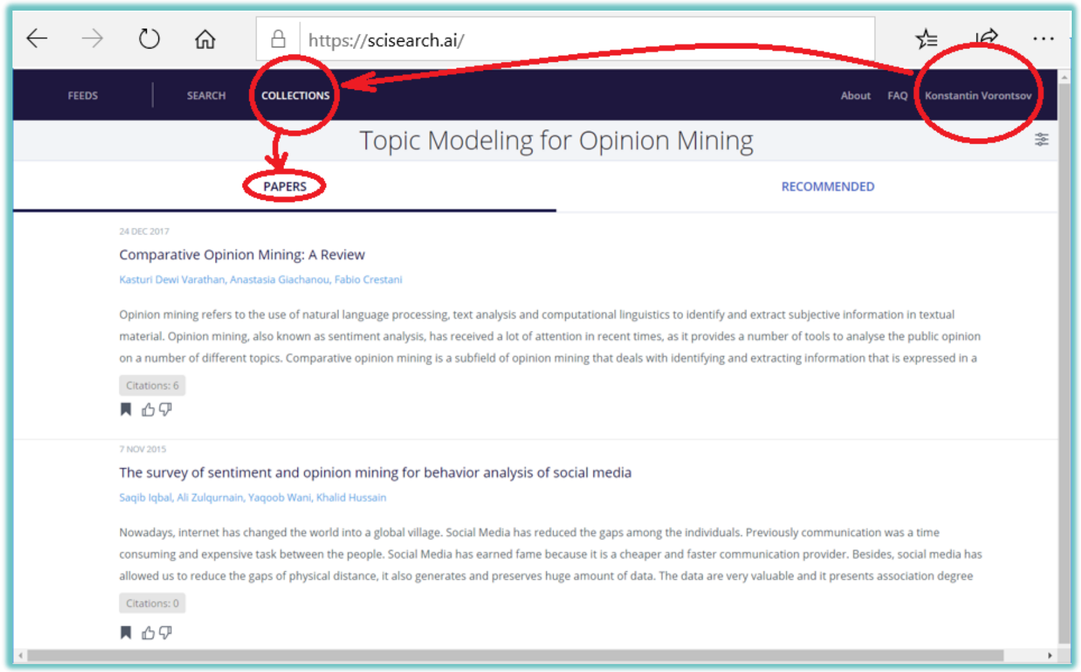
1. **Рабочее пространство** пользователя образуется тематическими подборками публикаций
2. **Текстуальность.** Знания представляются в виде массива текстов на естественном языке
3. **Коллегиальность.** Представление знаний служит взаимопониманию в рабочей группе
4. **Антропоцентричность.** Технологии не заменяют человека, а автоматизируют рутину
5. **Когнитивность.** Представление знаний с учётом восприятия, памяти, уровня образования
6. **Экстрактивность.** Меньше генерации, больше цитирования источников и ссылок
7. **Мультиязычность.** Автоматический перевод с языков источника на язык пользователя
8. **Расширяемость.** Платформа поддерживает возможность добавления сервисов
9. **Экономичность.** Сделать мир умнее труднее, чем сделать монетизируемый сервис

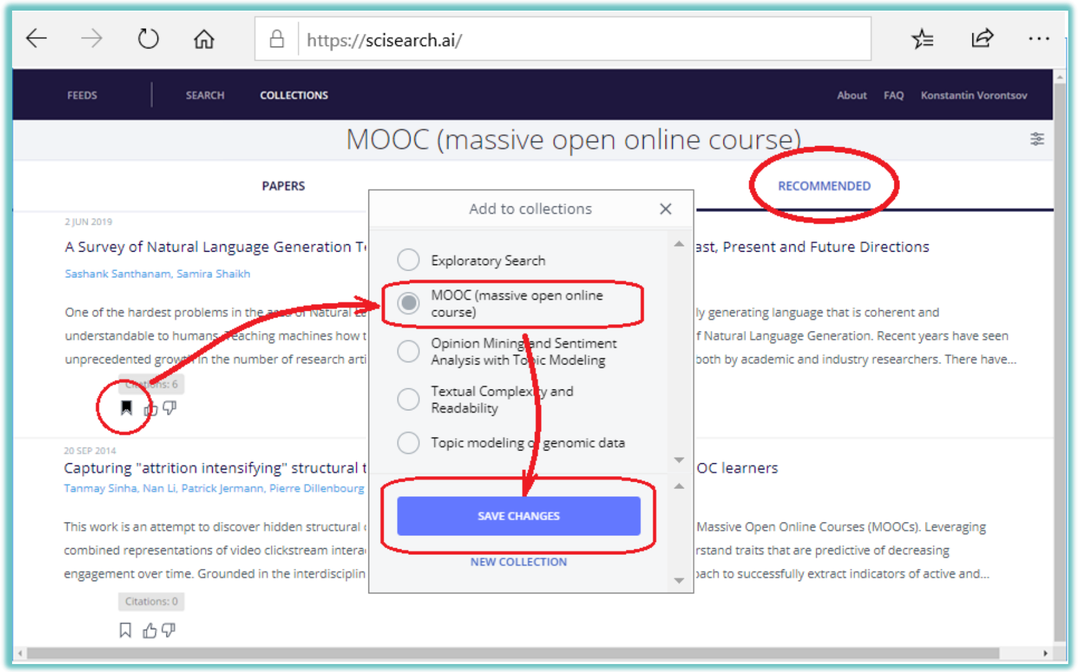
Мастерская знаний: сервис «Поиск»



Сервис поиска и ранжирования рекомендаций

Подборка играет роль поискового запроса и поисковой выдачи одновременно

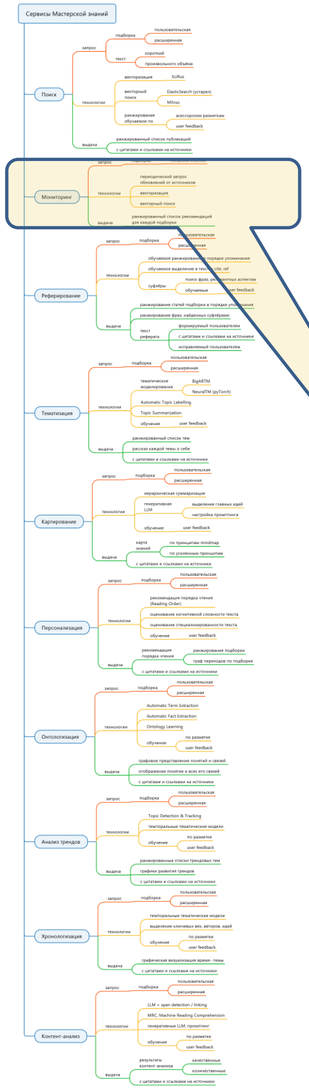


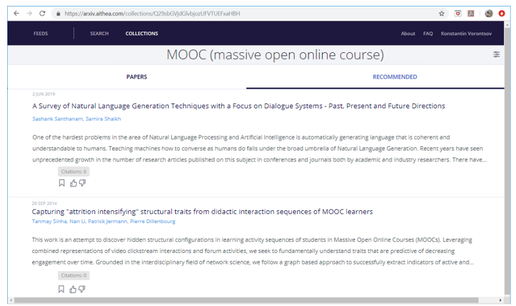


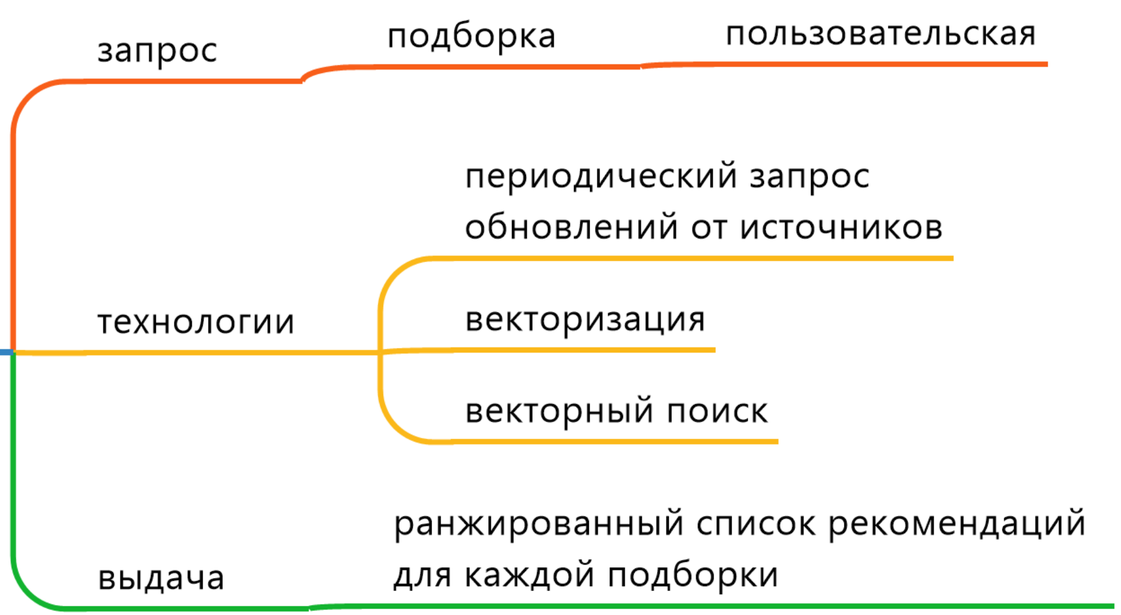
Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады РАН, 2024, том 520

Ватолин А.С., Герасименко Н.А., Янина А.О., Воронцов К.В. RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках // Доклады РАН, 2024, том 520

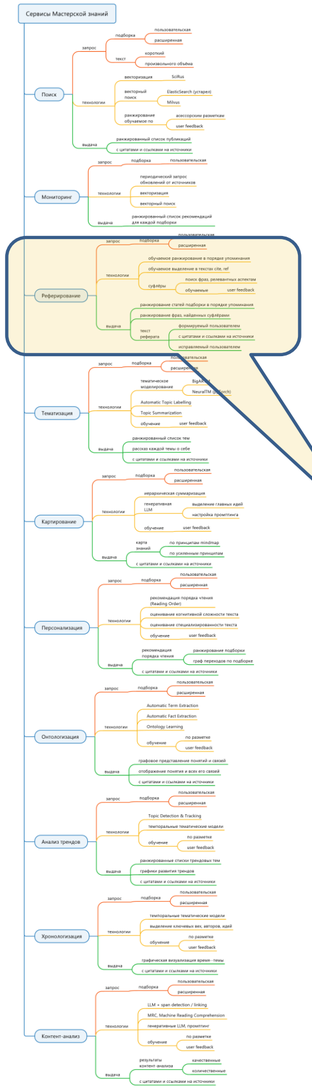
Мастерская знаний: сервис «Мониторинг»

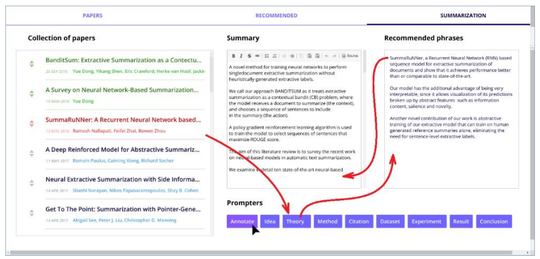


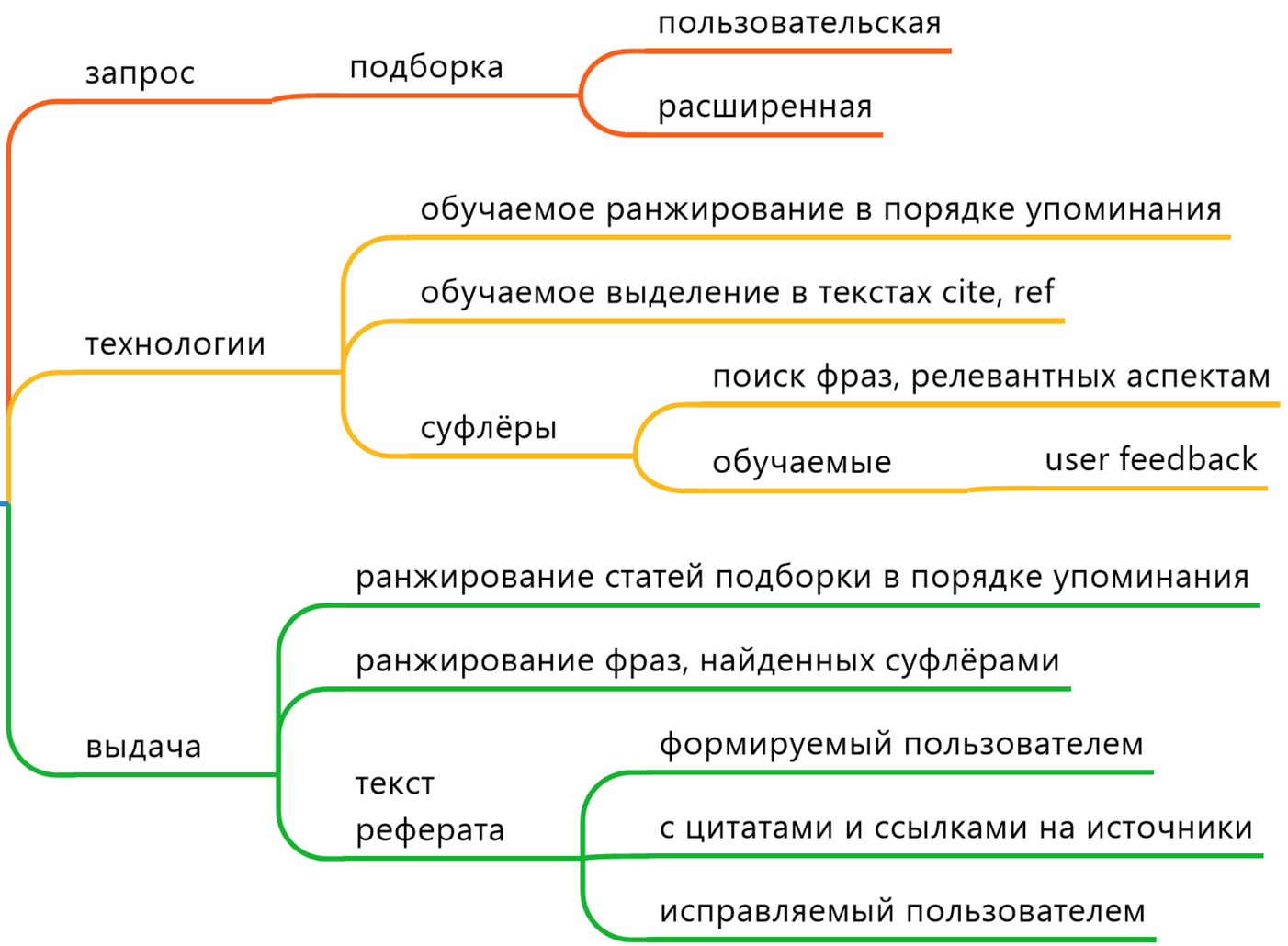




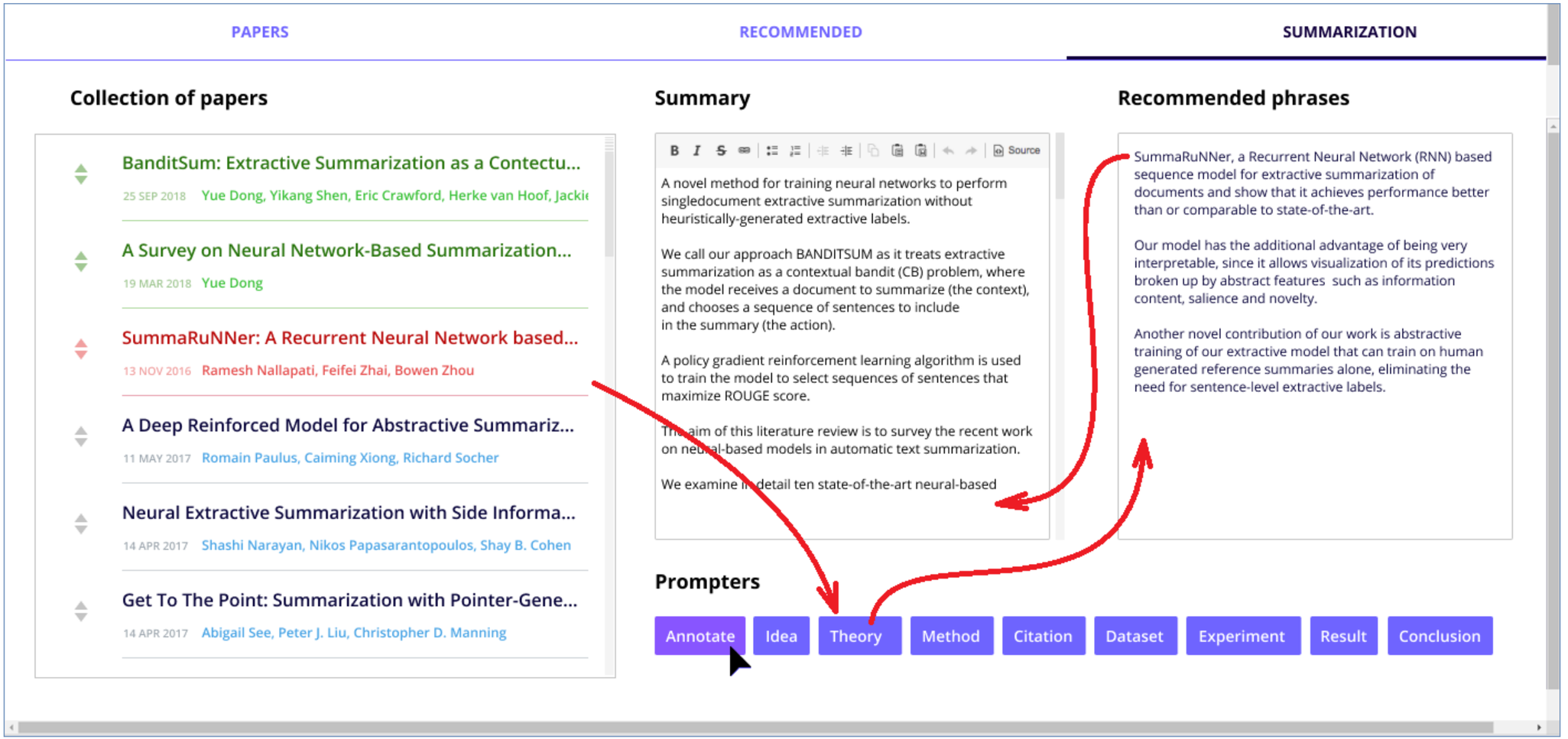
Мастерская знаний: «Реферирование»







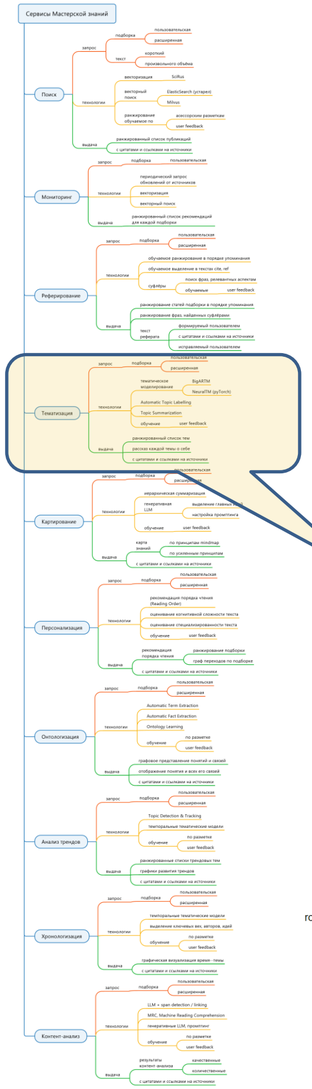
Сервис полуавтоматического реферирования

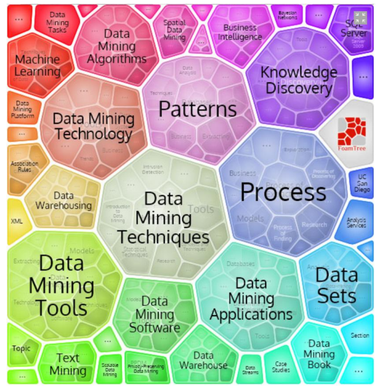


А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

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

Мастерская знаний: «Тематизация»





Тематизация

запрос
- подборка
  - пользовательская
  - расширенная

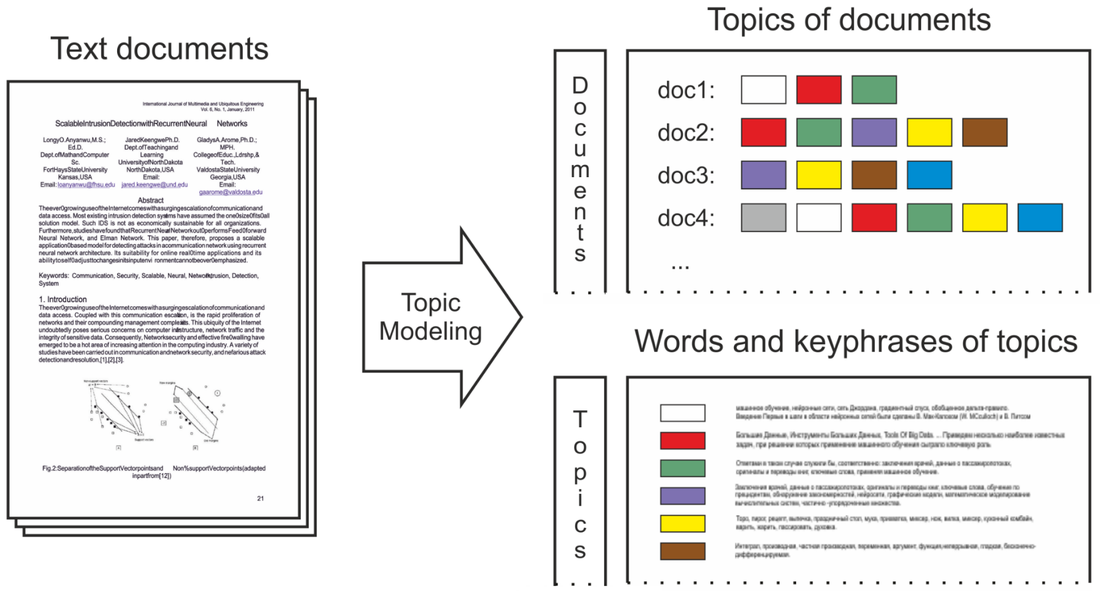
технологии
- тематическое моделирование
  - BigARTM
  - NeuralTM (pyTorch)
- Automatic Topic Labelling
- Topic Summarization
- обучение
  - user feedback

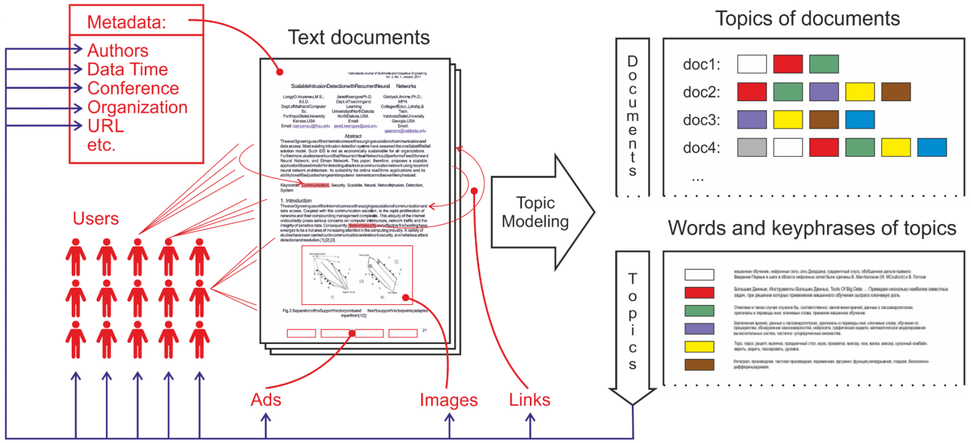
выдача
- ранжированный список тем
- рассказ каждой темы о себе
- с цитатами и ссылками на источники

Сервис тематизации подборки документов

Тематическая модель (ТМ) коллекции текстовых документов определяет

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





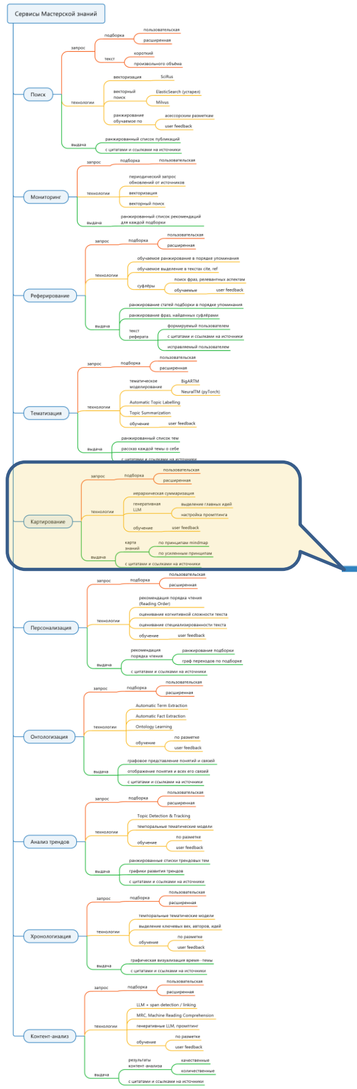
Мультимодальная ТМ определяет также,

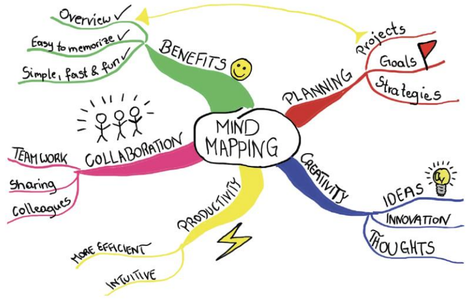
- какие ещё нетекстовые токены содержатся в каждой теме

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

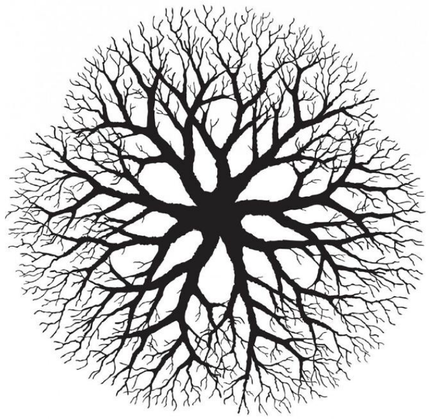
Vorontsov K. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023. http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Мастерская знаний: «Картирование»



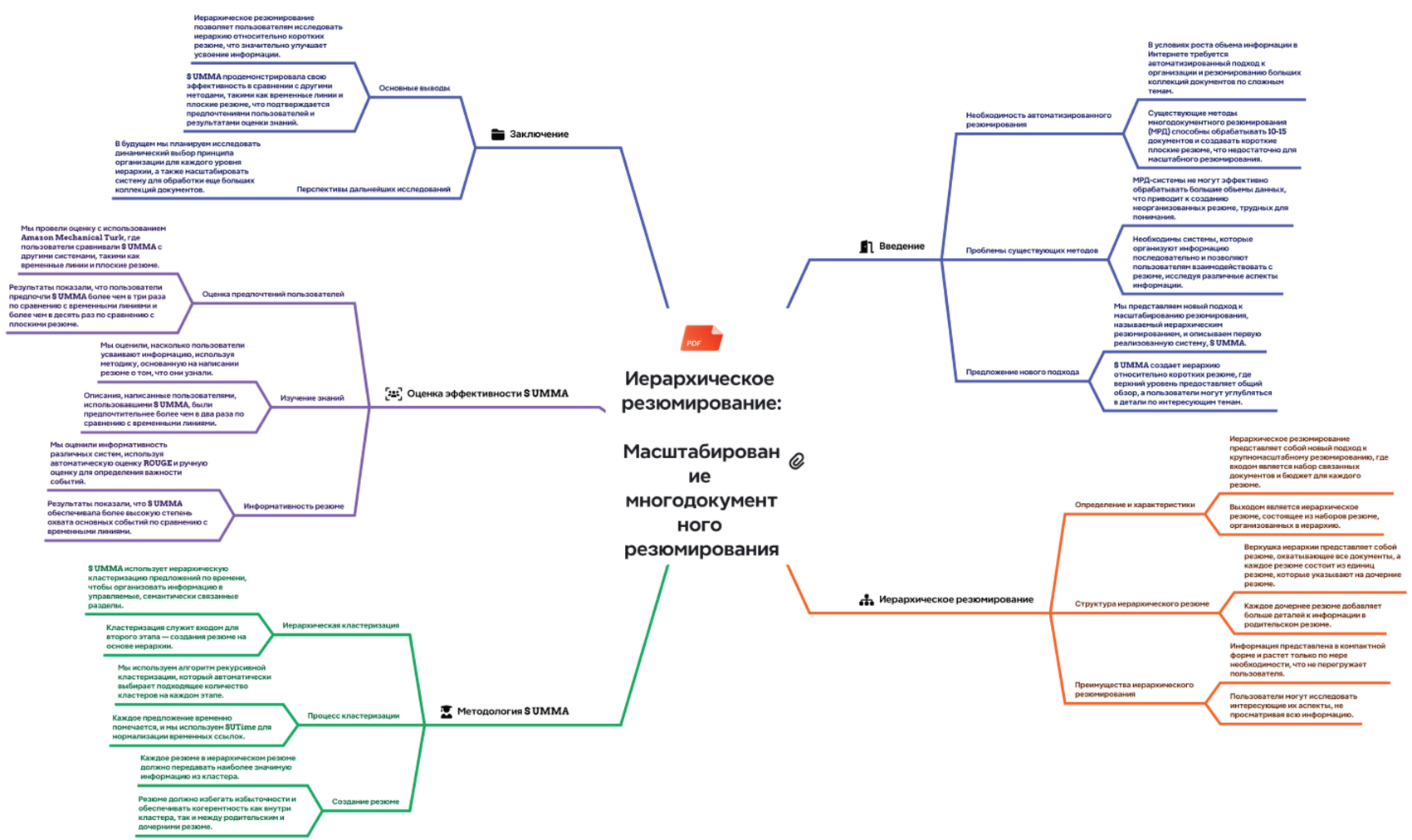


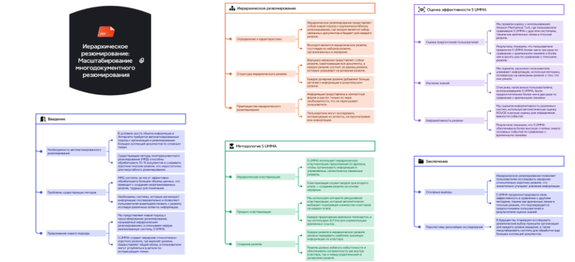
Картирование

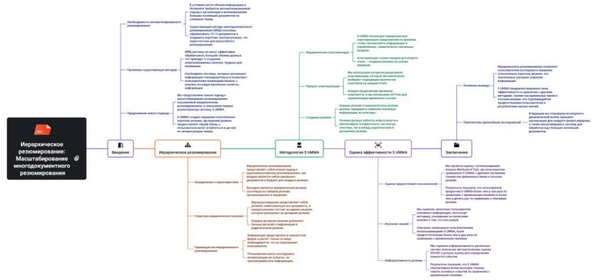


- запрос
  - подборка
    - пользовательская
    - расширенная
- технологии
  - иерархическая суммаризация
  - генеративная LLM
    - выделение главных идей
    - настройка промптинга
  - обучение
    - user feedback
- выдача
  - карта знаний
    - по принципам mindmap
    - по усиленным принципам
 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

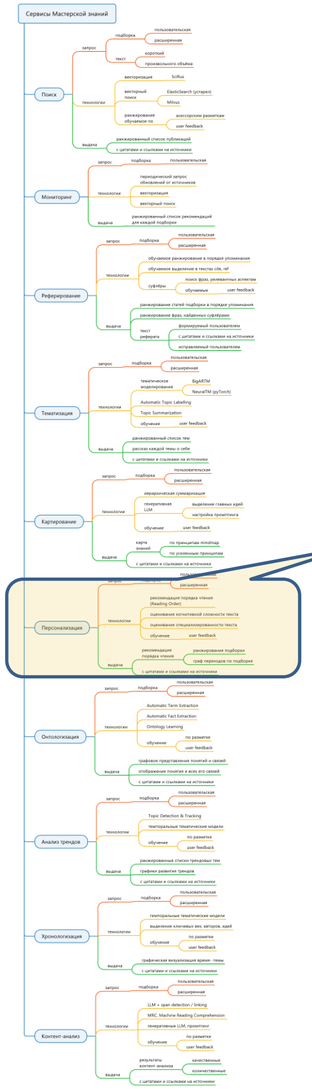
Пример: проект Mapify.so (AI MindMap Summarizer)

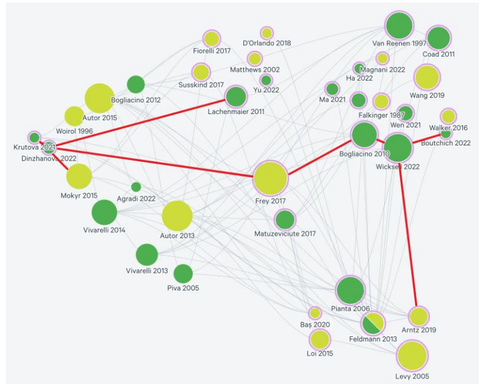






Мастерская знаний: «Персонализация»

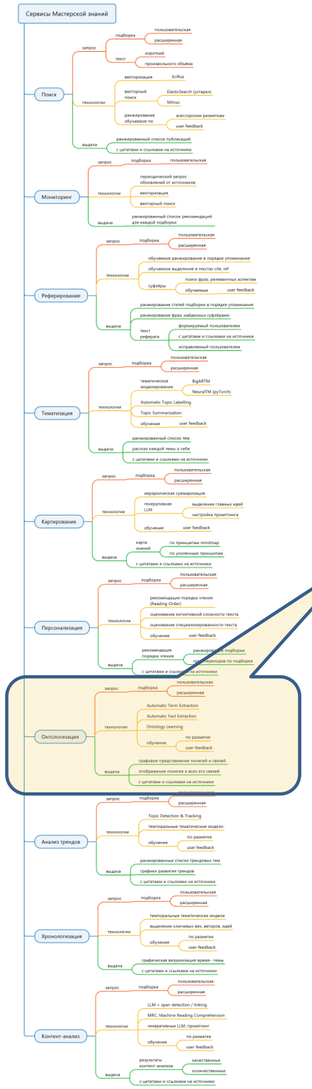


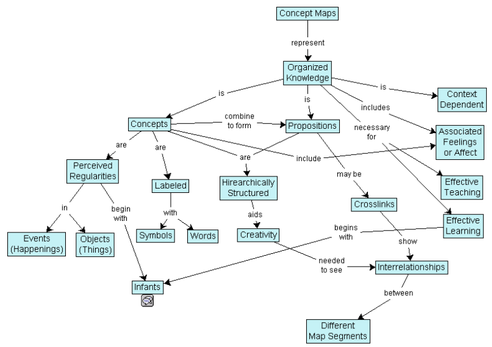


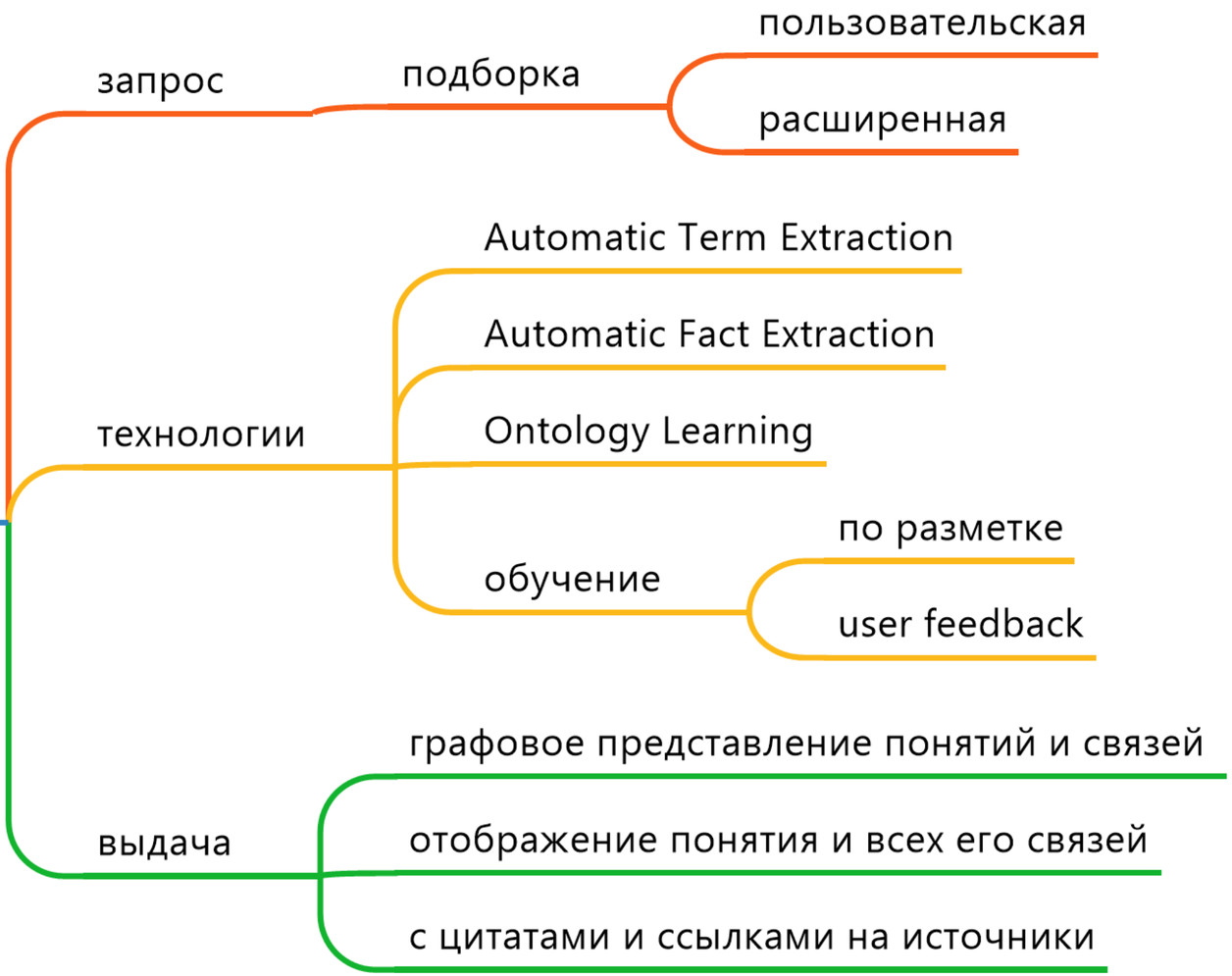
Персонализация

- 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

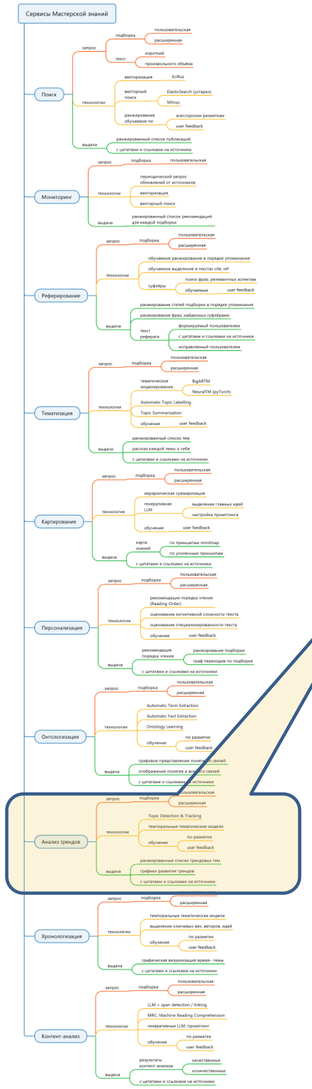
Мастерская знаний: «Онтологизация»





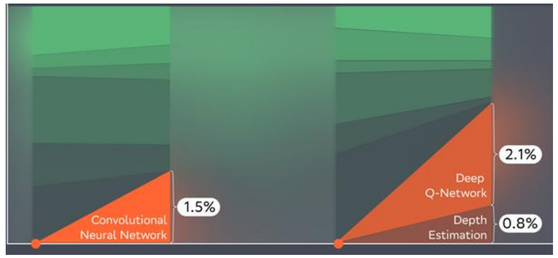


Мастерская знаний: «Анализ трендов»



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

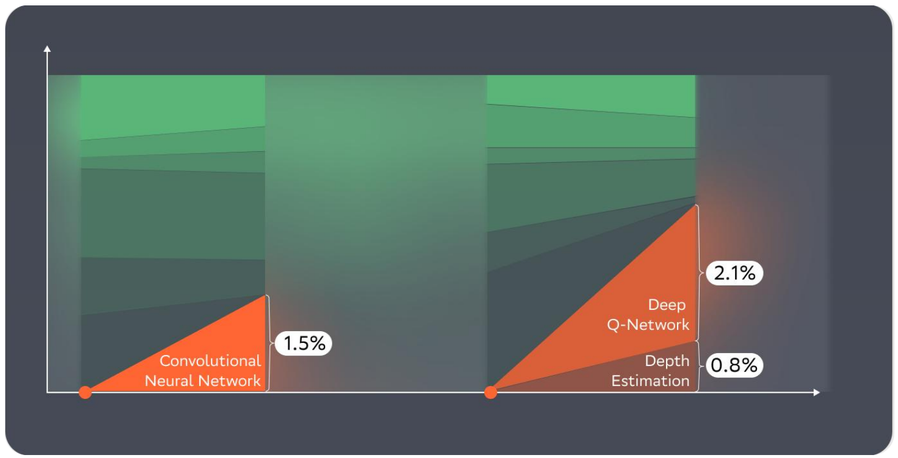


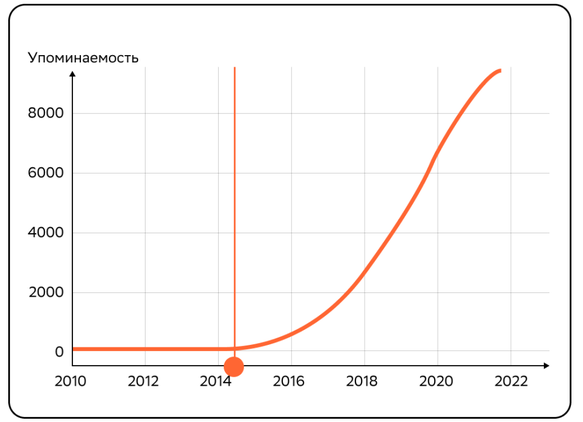
Сервис поиска и анализа трендов

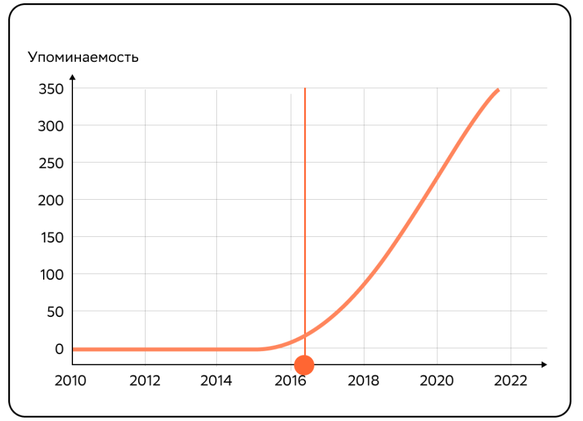
Определение трендовой темы:

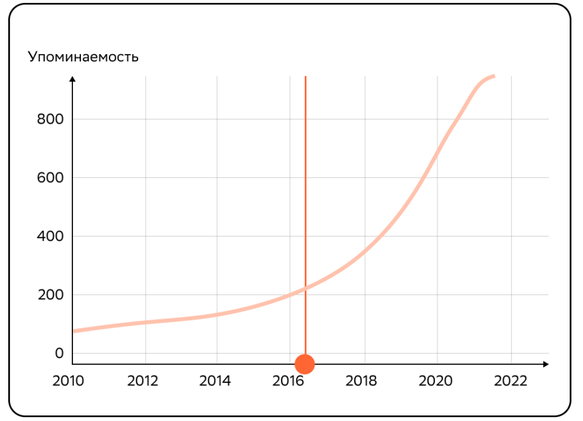
- наличие семантического ядра
- наличие быстрого (обычно экспоненциального) роста

Примеры динамики упоминаний трендовых тем



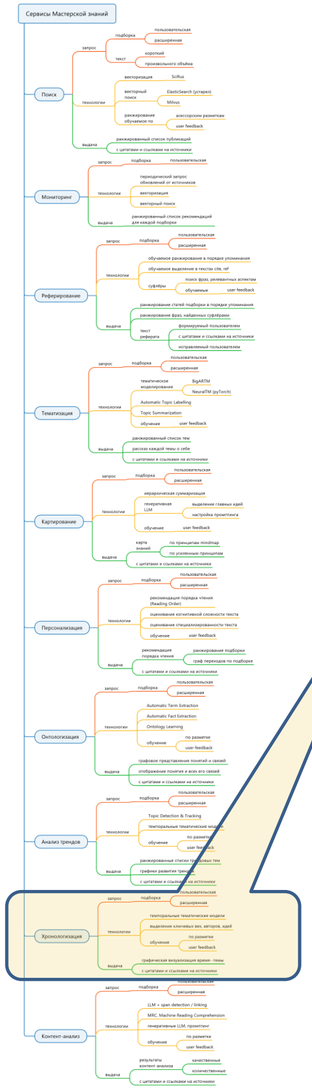


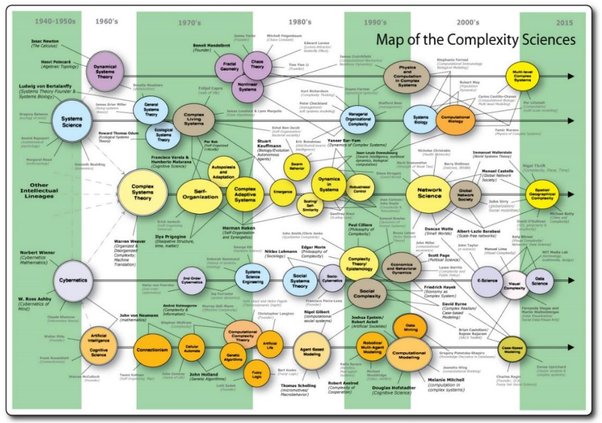


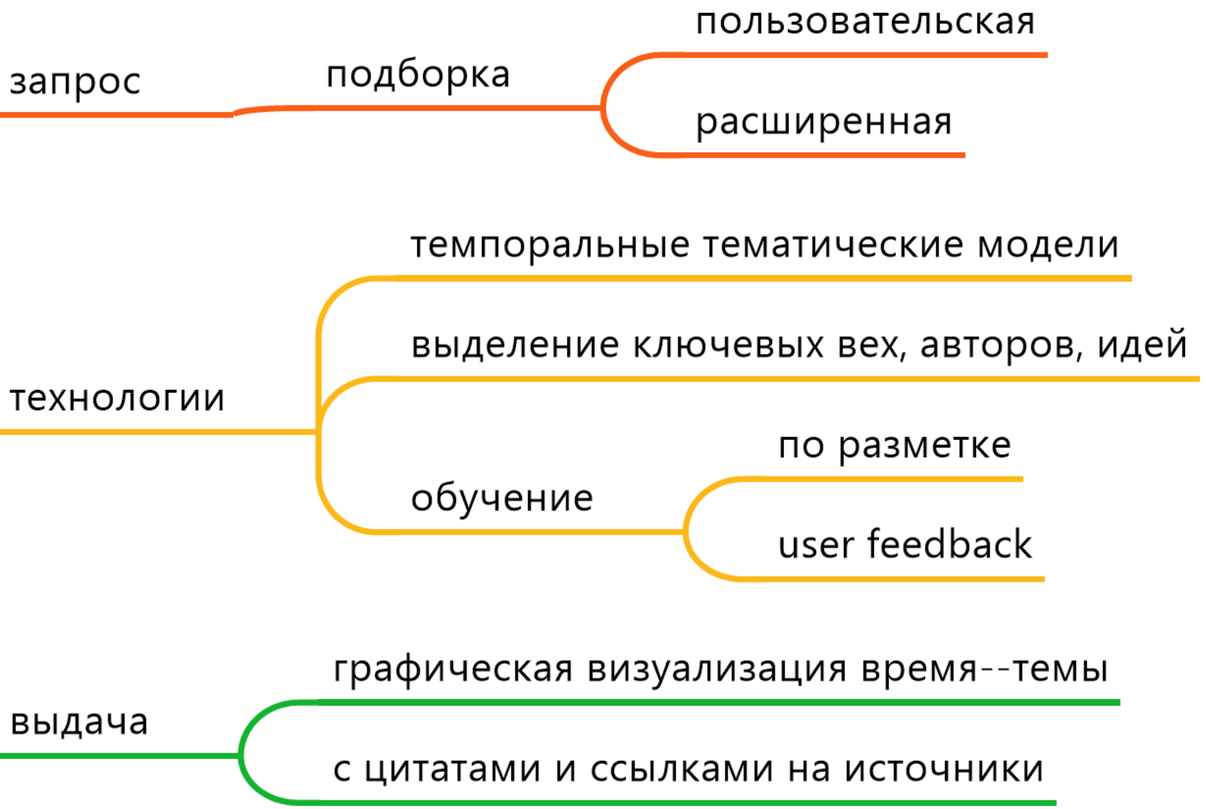


Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В. Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. Математика, информатика, процессы управления, 2022, том 508, С.106–108

Мастерская знаний: «Хронологизация»



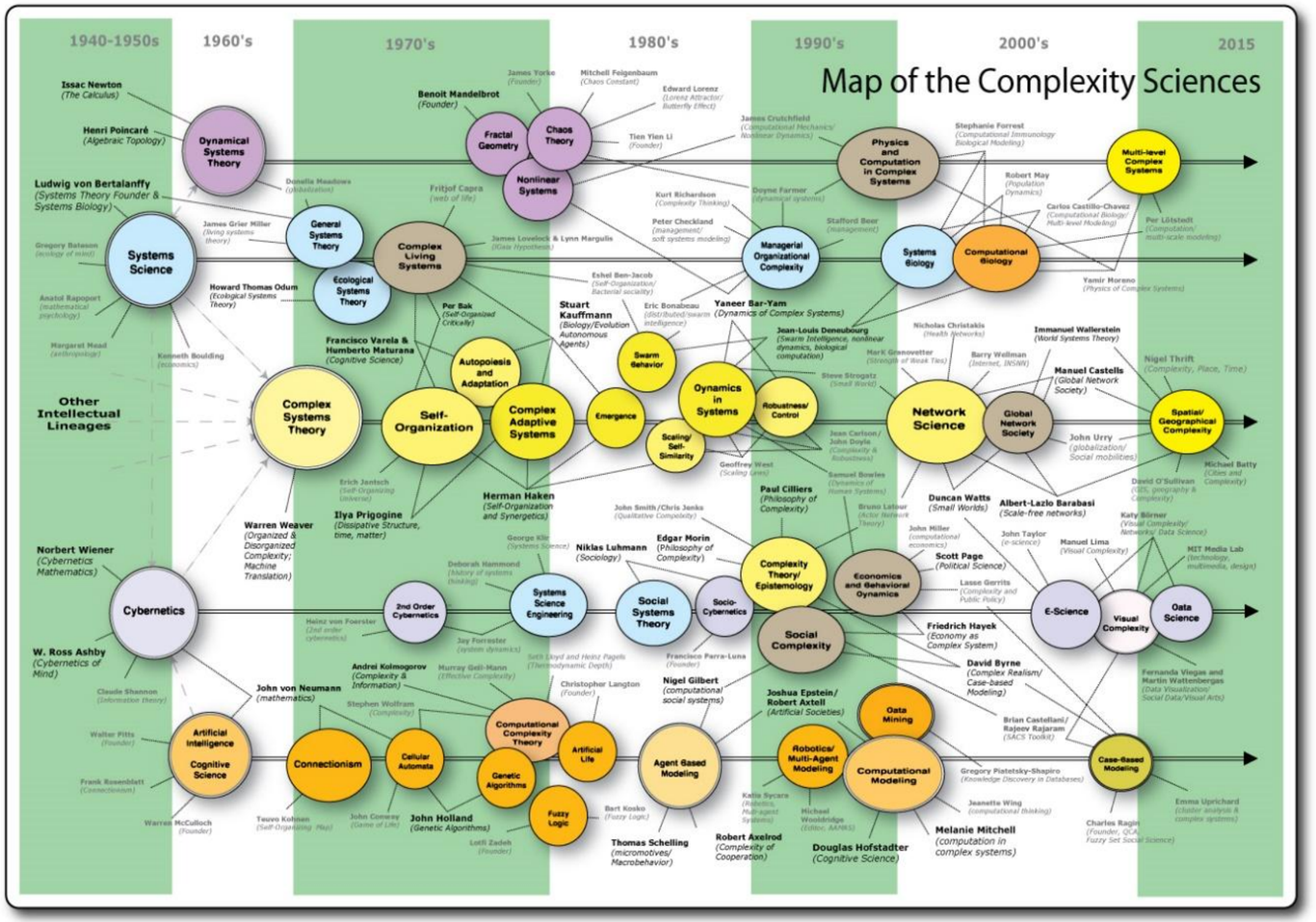




Сервис хронологических карт предметной области

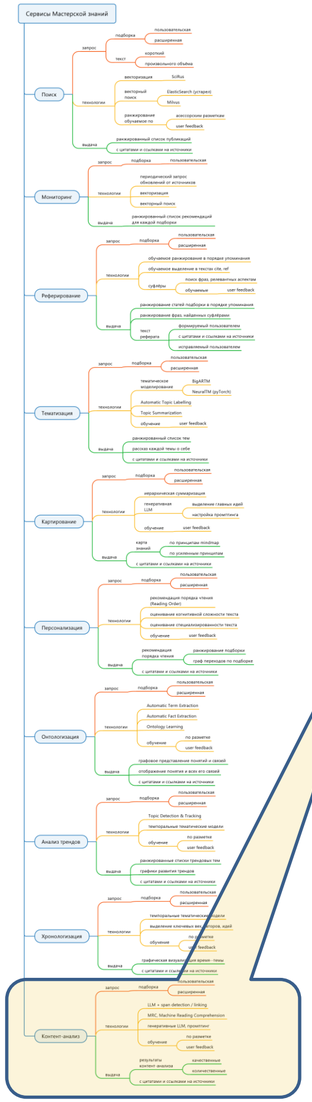
Осями на карте могут быть:

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



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

Мастерская знаний: «Контент-анализ»

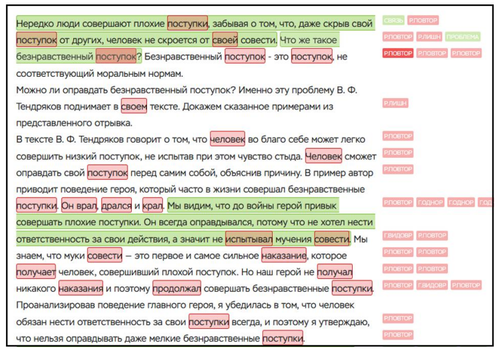


Контент-анализ

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



Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

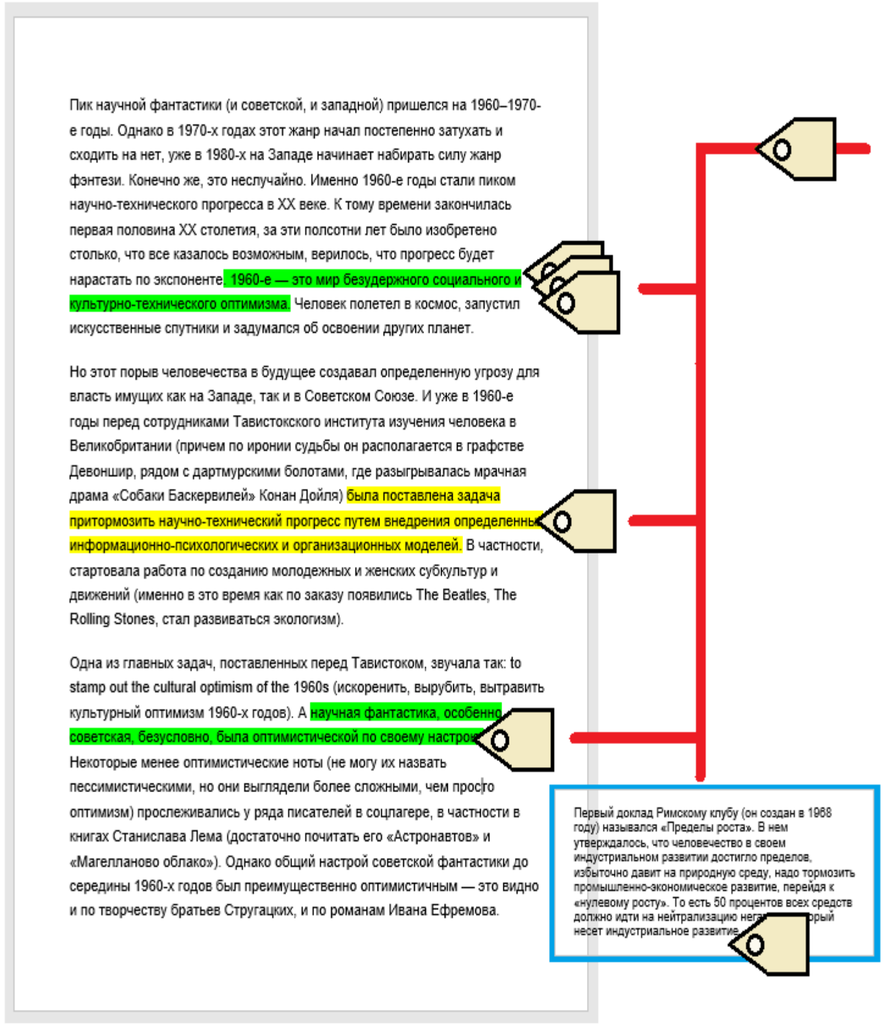
1) выделить фрагмент
2) классифицировать (тегировать) фрагмент по рубрикатору
3) связать несколько фрагментов
4) дать комментарий (затекст) к фрагменту или связи

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

Три подзадачи 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

Сервис автоматизации контент-анализа

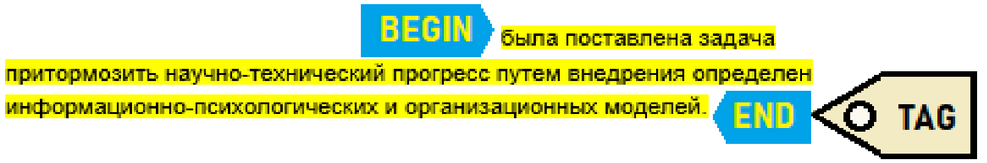


Разметка состоит из элементов

Элемент разметки — несколько взаимосвязанных фрагментов, затекстов и тегов

Теги (классы) выбираются из рубрикатора

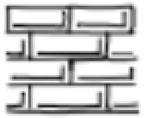
Фрагмент задаётся началом и концом, может иметь один или несколько тегов:



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

Миссия Мастерской Знаний

— устранять *барьеры* между человеком и знанием



технологические — из-за избыточности, неструктурированности, ненадёжности информации

когнитивные — из-за ограниченности наших возможностей запоминания, понимания, анализа

коммуникативные — из-за различий в мотивациях, уровне компетенций, социальном и служебном положении

Антропоцентричное определение ИИ

Искусственный интеллект —

вычислительные технологии, создаваемые для повышения производительности созидательного интеллектуального труда людей

не замена человека

не «загадочный новый тип разума»

не повод уподобиться Богу, чтобы «творить по образу и подобию Своему»

Спасибо за внимание!

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. кафедрой ММП ВМК МГУ,
зав. лаб. МОСА Института ИИ МГУ,
зав. кафедрой ИС и кафедрой МОЦГ МФТИ,
г.н.с. ФИЦ ИУ РАН

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

Научный семинар ИПУ РАН
«Проблемы управления знаниями»
руководители:
академик РАН Д.А.Новиков,
проф. РАН К.В.Воронцов

Дополнения и технические подробности

От интеллект-карт (mind-maps) к картам знаний

Интеллект-карты (mind maps)

текстографическое отображение того, как темы (мысли, идеи) разбиваются на подтемы иерархически

максимально близкое к тому, как мы храним знания у себя в головах

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

От интеллект-карт (mind-maps) к картам знаний

нацелены на повышение эффективности

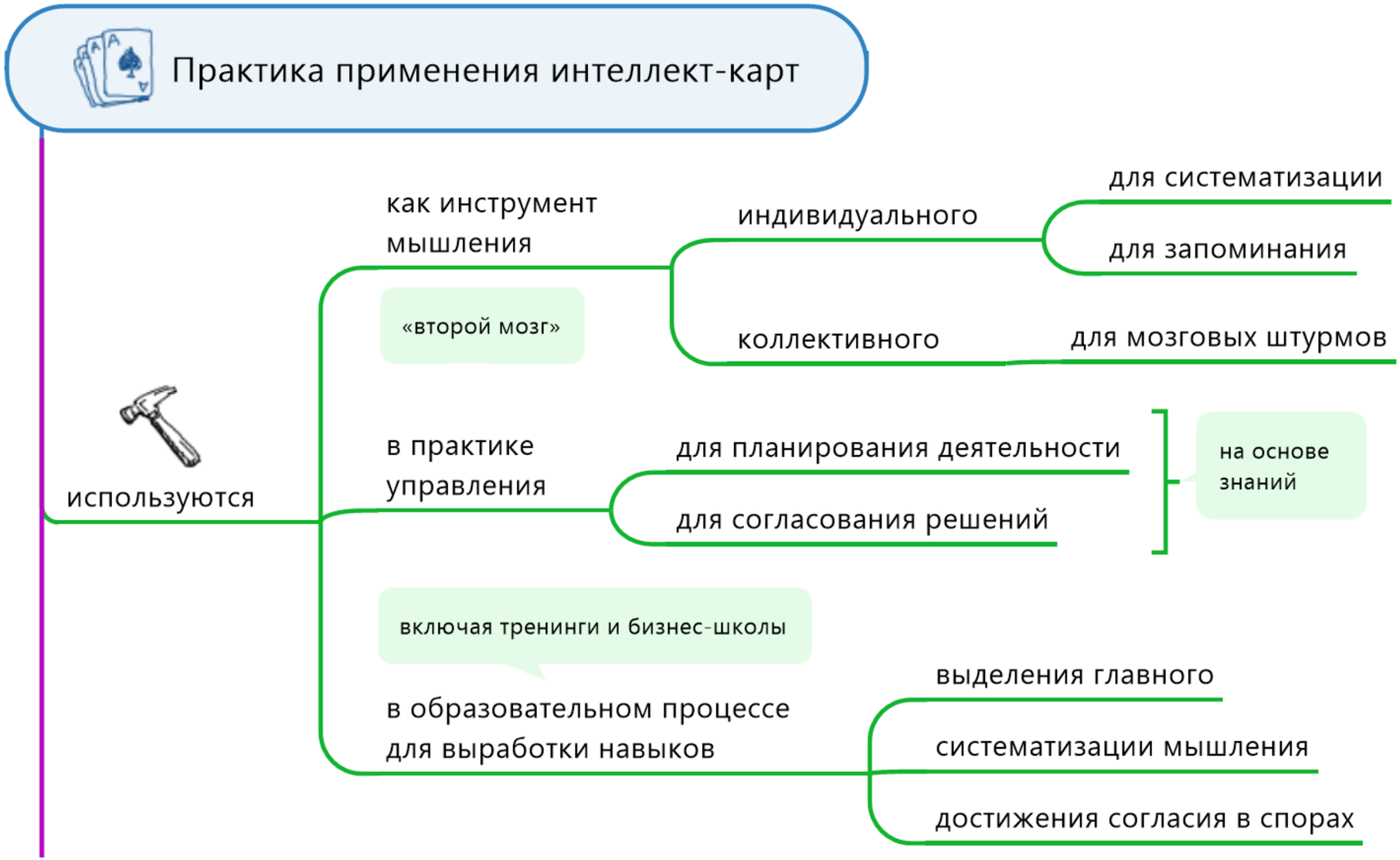
- конспектирования
- понимания
- запоминания
- систематизации
- поиска консенсуса

благодаря активизации обоих полушарий мозга, учёта особенностей восприятия, мышления, памяти

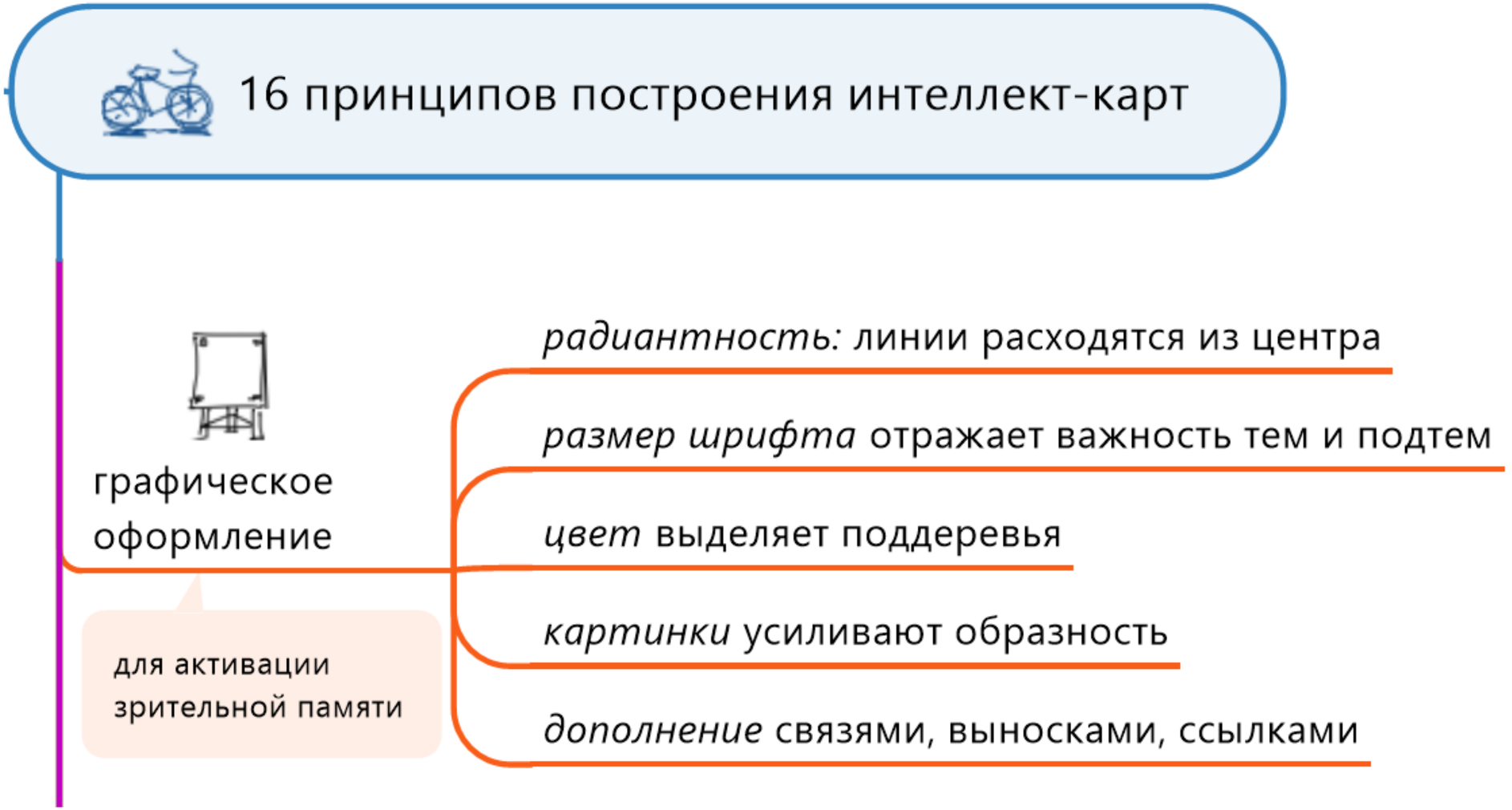
техника запоминания

- посмотреть, понять, обсудить, договориться, принять
- самостоятельно воспроизвести через 10 минут → сутки → неделю → месяц

От интеллект-карт (mind-maps) к картам знаний

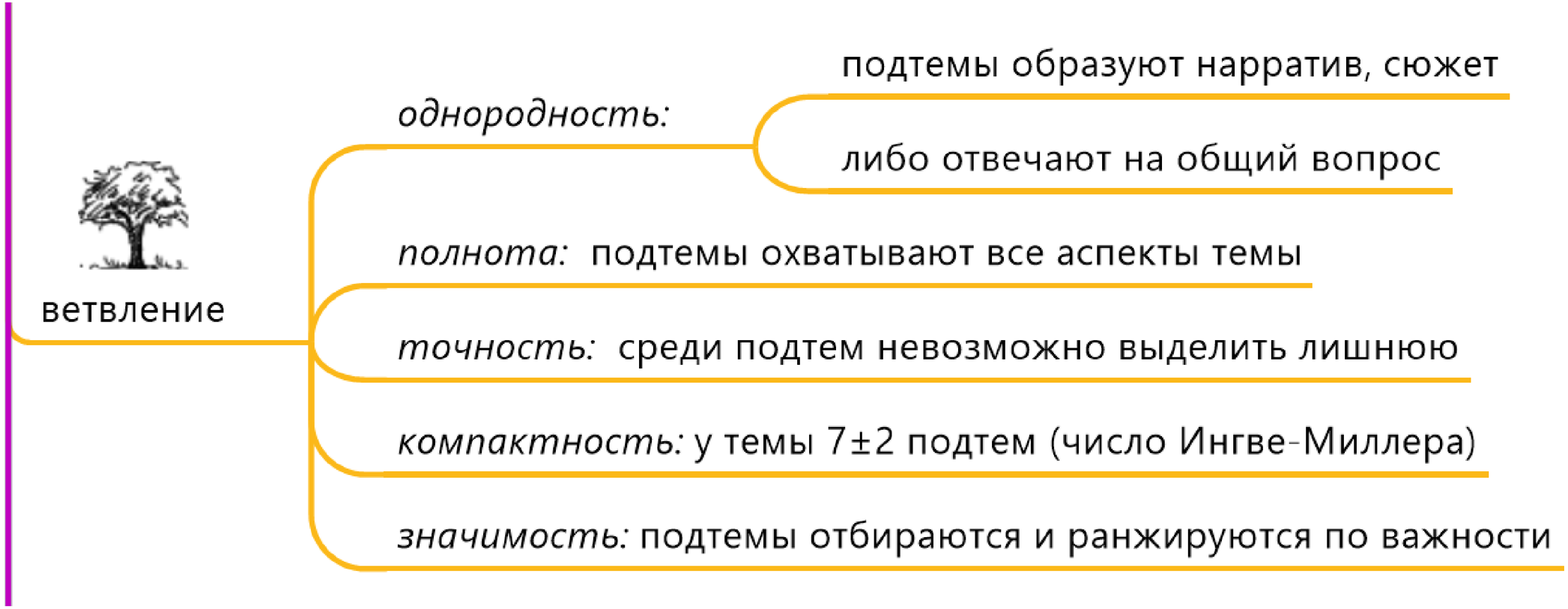


От интеллект-карт (mind-maps) к картам знаний



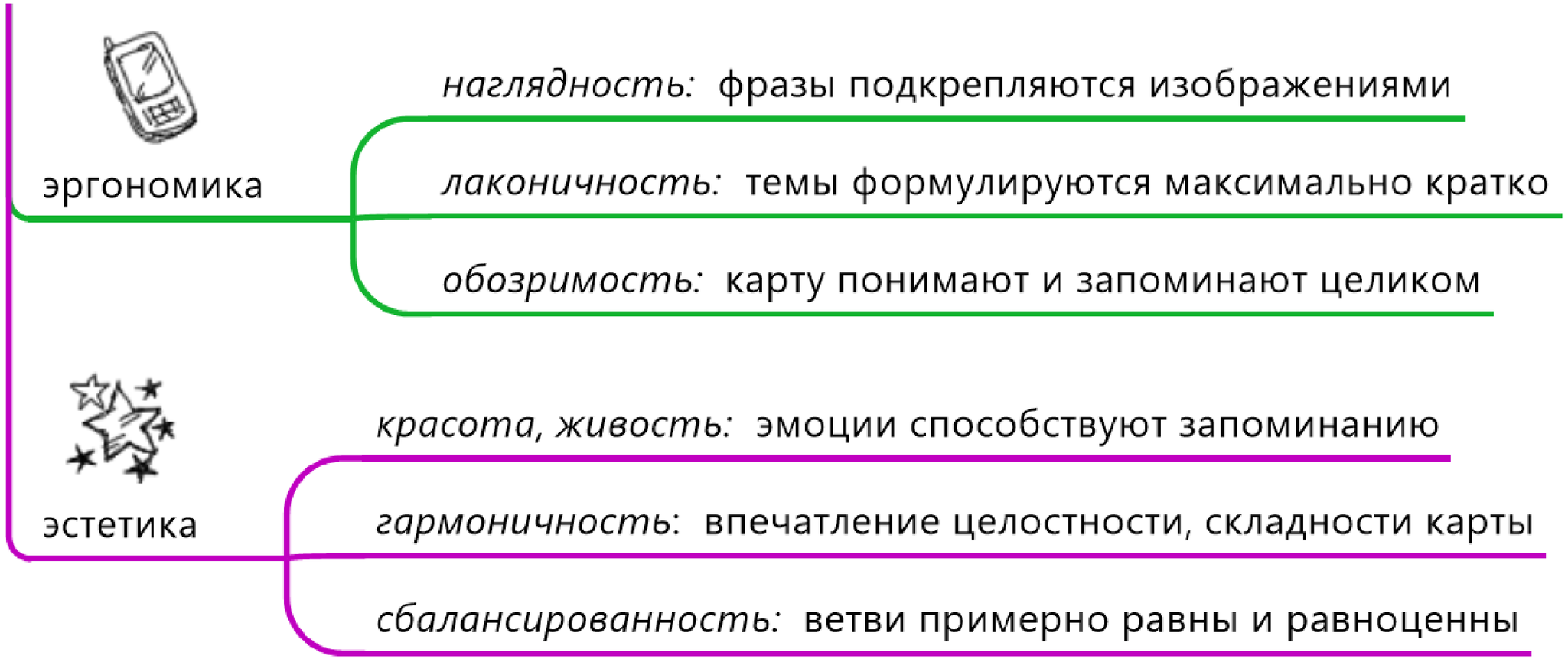
От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



От интеллект-карт (mind-maps) к картам знаний

6 принципов, усиливающих интеллект-карты до **карт знаний**

(1)
читабельность

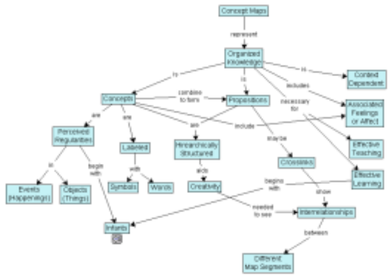
компромисс с
лаконичностью
и обозримостью

любой фрагмент карты
читается как нарратив

легко и однозначно

даже автоматически

в отличие
от других
способов
представления
знаний

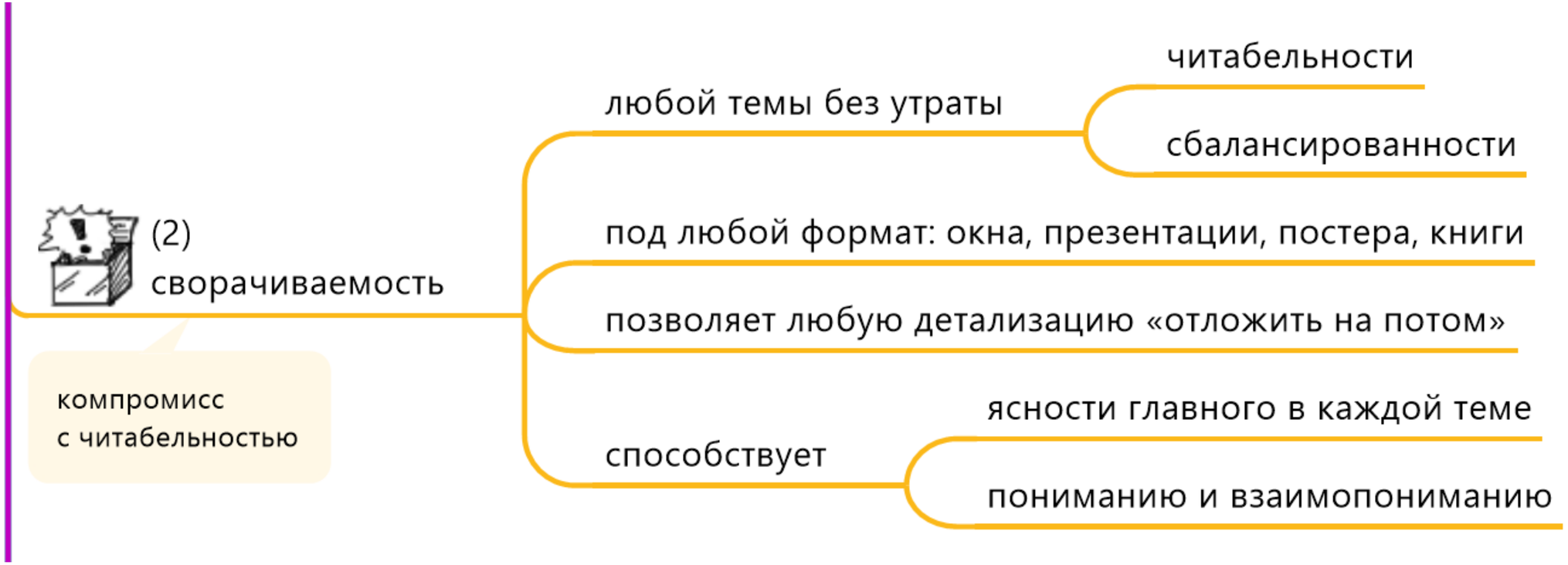


онтологий

фреймов и др.

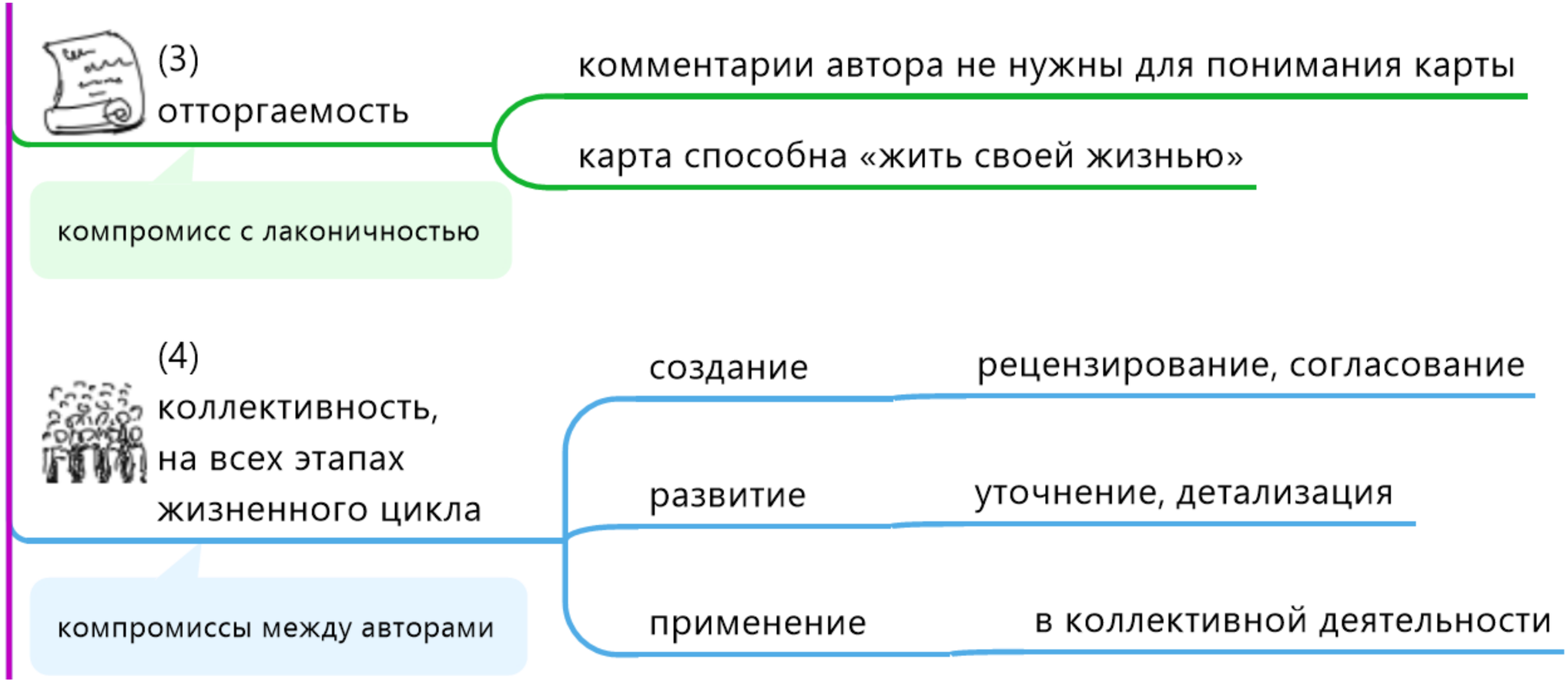
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



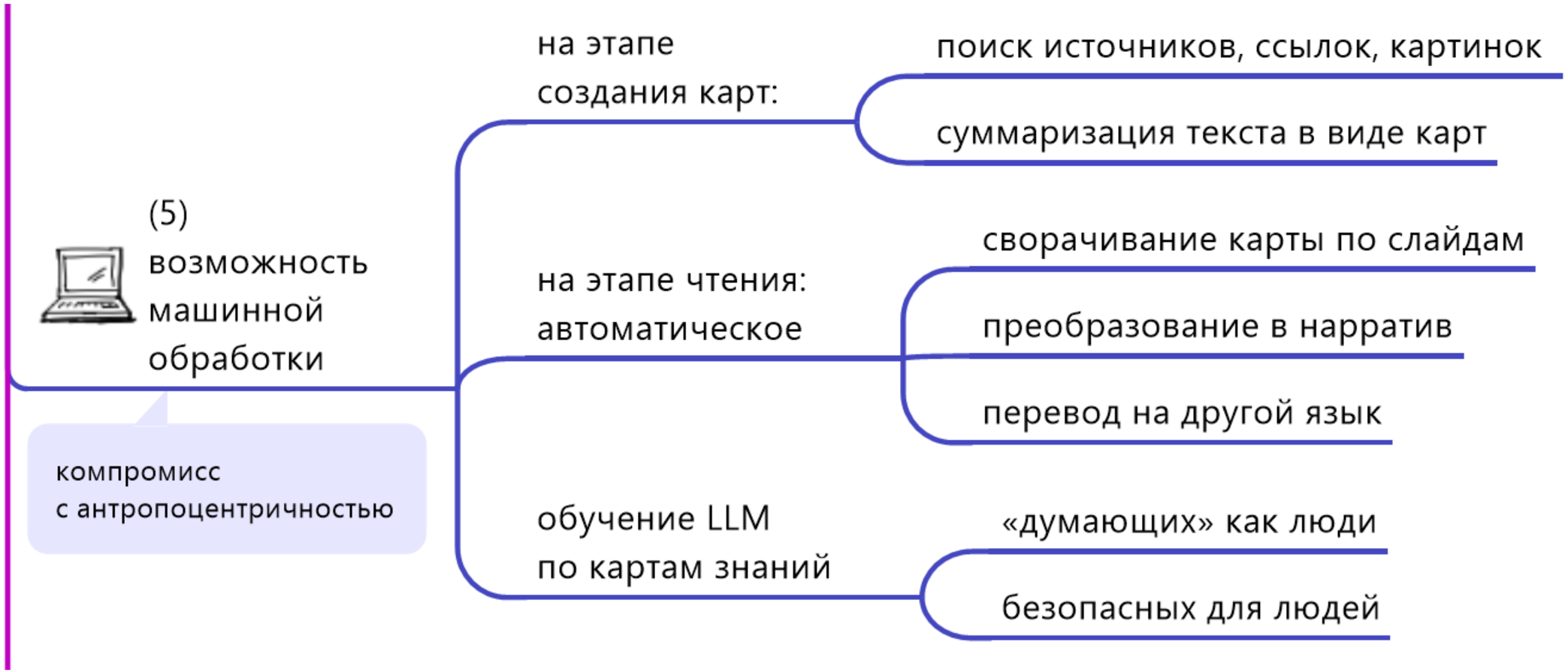
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



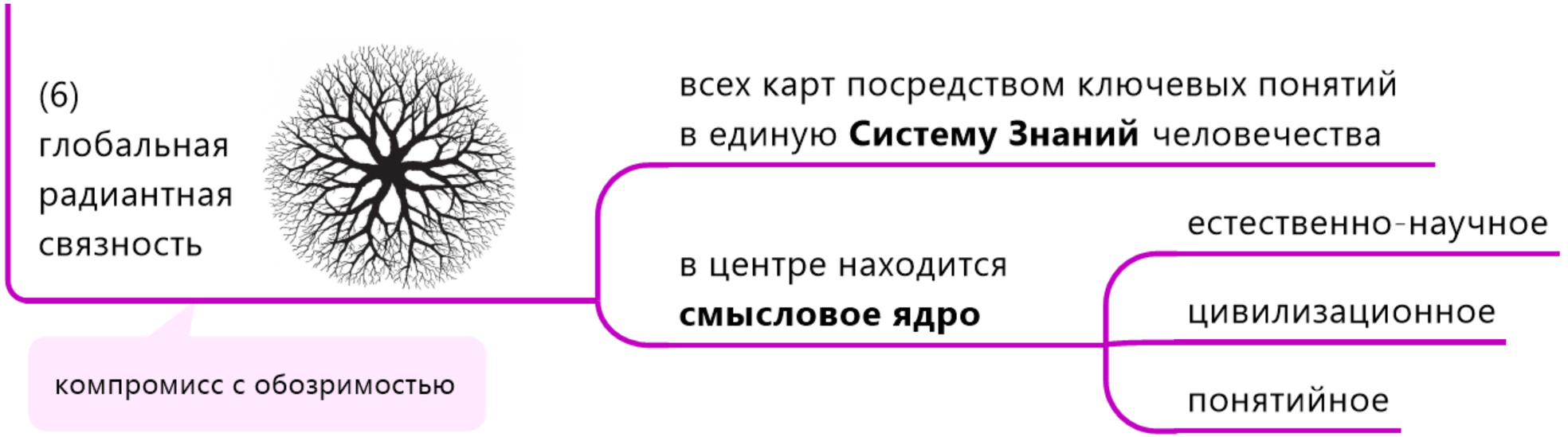
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



От интеллект-карт (mind-maps) к картам знаний

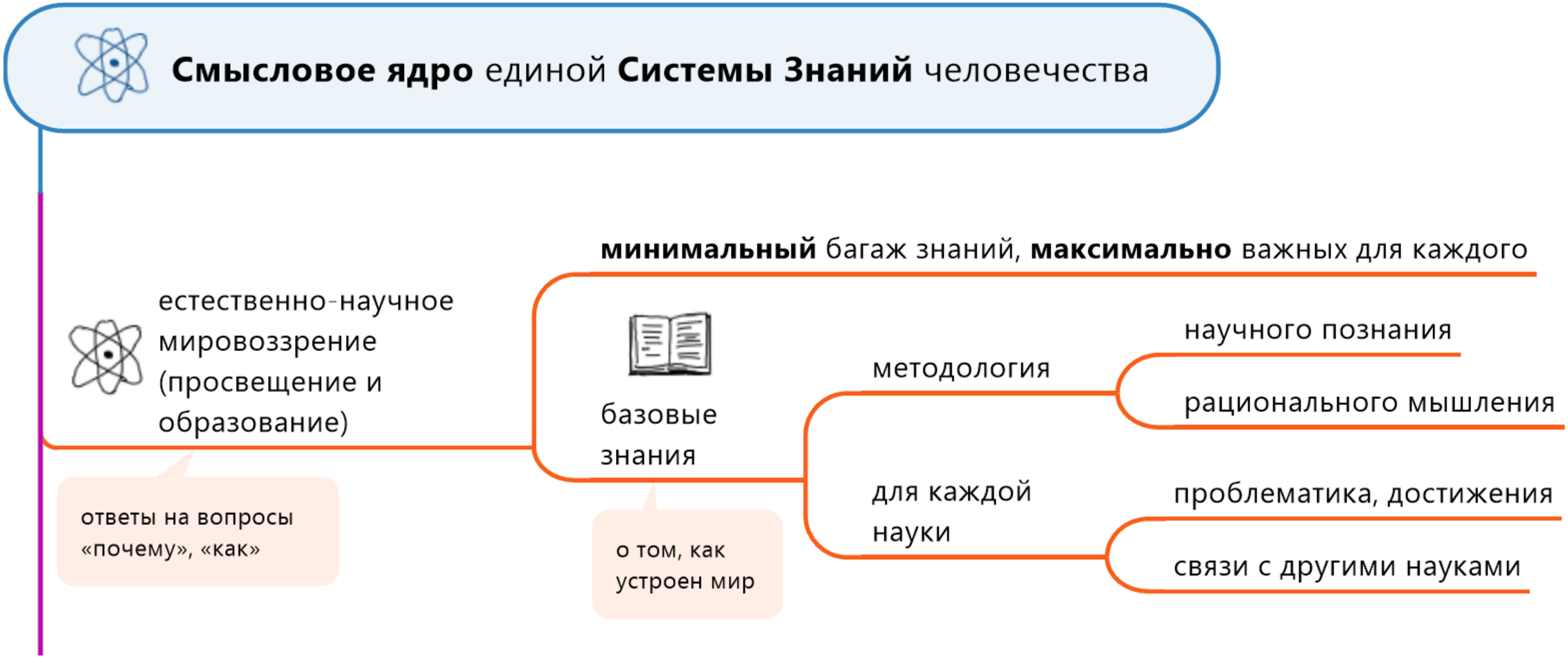
(6 принципов, усиливающих интеллект-карты до **карт знаний**)



Воронцов К.В., Курилов В.А. Карты знаний: на пути к доверенным языковым моделям и системам представления знаний. BIS Journal, №3 (54), 2024.

Курилов В. А. Теория счастья. Азбука жизни. Красноармейск, 2006. 208с. ISBN 5-91270-001-1

От интеллект-карт (mind-maps) к картам знаний



От интеллект-карт (mind-maps) к картам знаний

ответы на вопросы «ради чего», «зачем»

цивилизационная идеология (воспитание)

система ценностей человеческой цивилизации

как прямое следствие и продолжение естественно-научной картины мира

знание и понимание

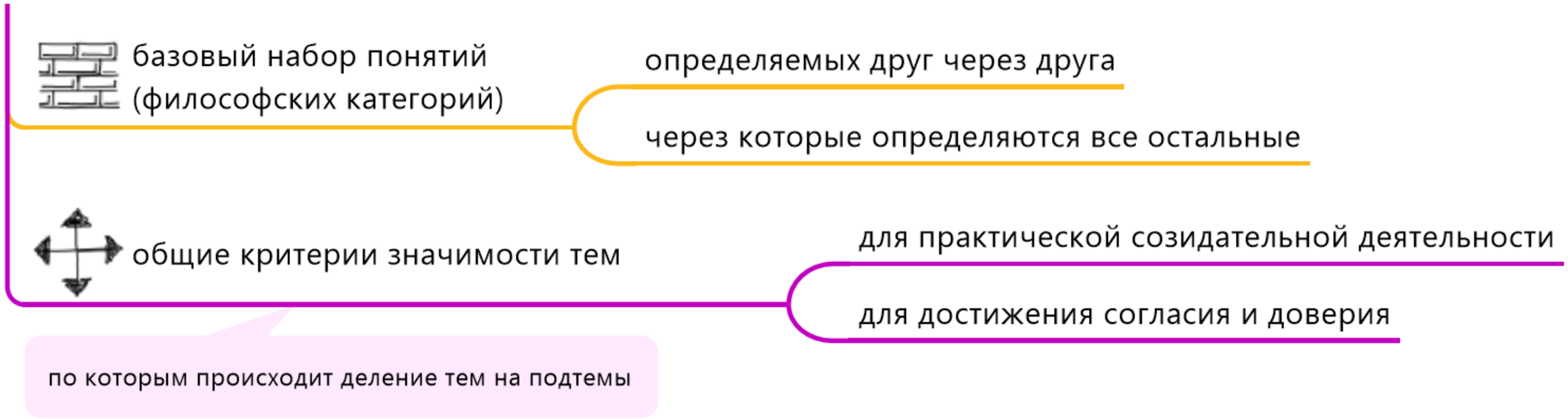
- целей и задач цивилизации
- законов сохранения цивилизации
- позитивных образов будущего

ответы на важнейшие личностные вопросы

- в чём смысл жизни
- как быть счастливым
- зачем нужны знания
- зачем я работаю

необходимые для мотивации деятельности

От интеллект-карт (mind-maps) к картам знаний



Как активировать визуальное аналитическое мышление (эволюционно обусловленное, намного более мощное)

1. порядка сотни карт: просмотреть, обсудить, поспорить, принять
2. десятки карт: построить самому, следуя 16+6 принципам
3. испытать «моменты ясности», инсайты, когда карта
   - «красиво сложилась»
   - привела к согласию
   - легко и ярко запомнилась,
   - легла в основу деятельности

индивидуальная практика и опыт

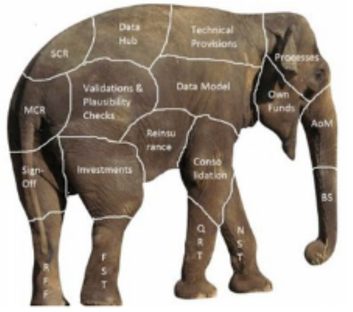
4. сделать построение карт регулярной профессиональной практикой
   - индивидуальной
   - коллективной

Задачи: ближайшие и перспективные

методологические

- отработать методологию использования карт знаний
  - при анализе учебных и научных текстов
    - отдельных статей
    - подборок статей
  - при ведении проектной деятельности
  - при решении значимых проблем
    - научно-технических
    - экономических
    - общественно-политических
- разработать структуру и прототипировать наполнение смыслового ядра единой Системы Знаний
- отработать на прототипе принципы построения карт знаний и их встраивания в единую Систему Знаний

Задачи: ближайшие и перспективные



организационные

- организовать сбор данных по индивидуальному и коллективному построению карт по учебным и научным текстам
- организовать сообщество «Картографии Знаний»
- создать фабрику мысли (think tank), вооружённую методологией визуального аналитического мышления (не только карт знаний)
- внедрять Систему Знаний для коллективного решения значимых научно-технических и экономических проблем

Задачи: ближайшие и перспективные

технологические

- разработать инструментарий для коллективного редактирования карт знаний и контроля версий
- разработать инструментарий для визуализации и навигации по единой Системе Знаний
- разработать прототип единой Системы Знаний
- разработать модель иерархической суммаризации для (полу)автоматического наполнения карт знаний
- реструктурировать Википедию, преобразовав её в карту знаний (полу)автоматически
- построить LLM (большую языковую модель) по текстографическому корпусу Системы Знаний
- вводить LLM в роли интеллектуального помощника при построении карт знаний в проектной деятельности

ВЫВОДЫ: **КАРТЫ ЗНАНИЙ**

основаны на интеллект-картах (mind-map) Тони Бьюзена

отличаются более строгими принципами построения (16+6)

активируют **визуальное аналитическое мышление**

способствуют взаимопониманию в коллективной интеллектуальной деятельности

при накоплении образуют размеченную выборку
для обучения LLM навыкам выделять главное, общаться с людьми
на языке **радиантно структурированного текста**

могут

- строиться эффективнее при автоматизации средствами ИИ / LLM
- стать методологической основой единой Системы Знаний
- обеспечить доверенность следующего поколения LLM
- стать базовым инструментом **коллективного разума**
- внедрить в ИИ человеческую цивилизационную систему ценностей
- обеспечить доверенность в **человеко-машинной цивилизации**

в порядке усиления гипотез

Проблемы, решаемые картами Знаний

Проблема неэффективности текста

(текст из драйвера развития цивилизации становится его тормозом)

как средства передачи знаний от головы к голове,
в условиях экспоненциального роста объёма накопленных знаний

из-за затрат на кодирование знаний в текст и декодирование их из текста обратно

при том, что в головах знания структурированы

- иерархически, по важности, образно
- нечётко, неточно, у всех по-разному

избыточность текста может скрывать

- нелогичность, неточность, неполноту
- заблуждения, обман, флуд, демагогию, графоманию

решаемая путём

- явного выделения главного, всегда и во всём
- иерархической структуризации текста в виде карт
- инструментализации нелинейного чтения (non-linear reading)

Проблемы, решаемые картами Знаний

Проблема
недостаточного доверия людей к LLM (и друг к другу)

почему люди доверяют друг другу

оценивание компетентности, деловых качеств

накопление деловой репутации, соц. капитала

структурированность, системность мышления, умение выделять главное *(elevator pitch)*

почему люди станут доверять моделям LLM

оценивание: измеримая обратная связь от пользователей

накопление статистики по многокритериальным бенчмаркам

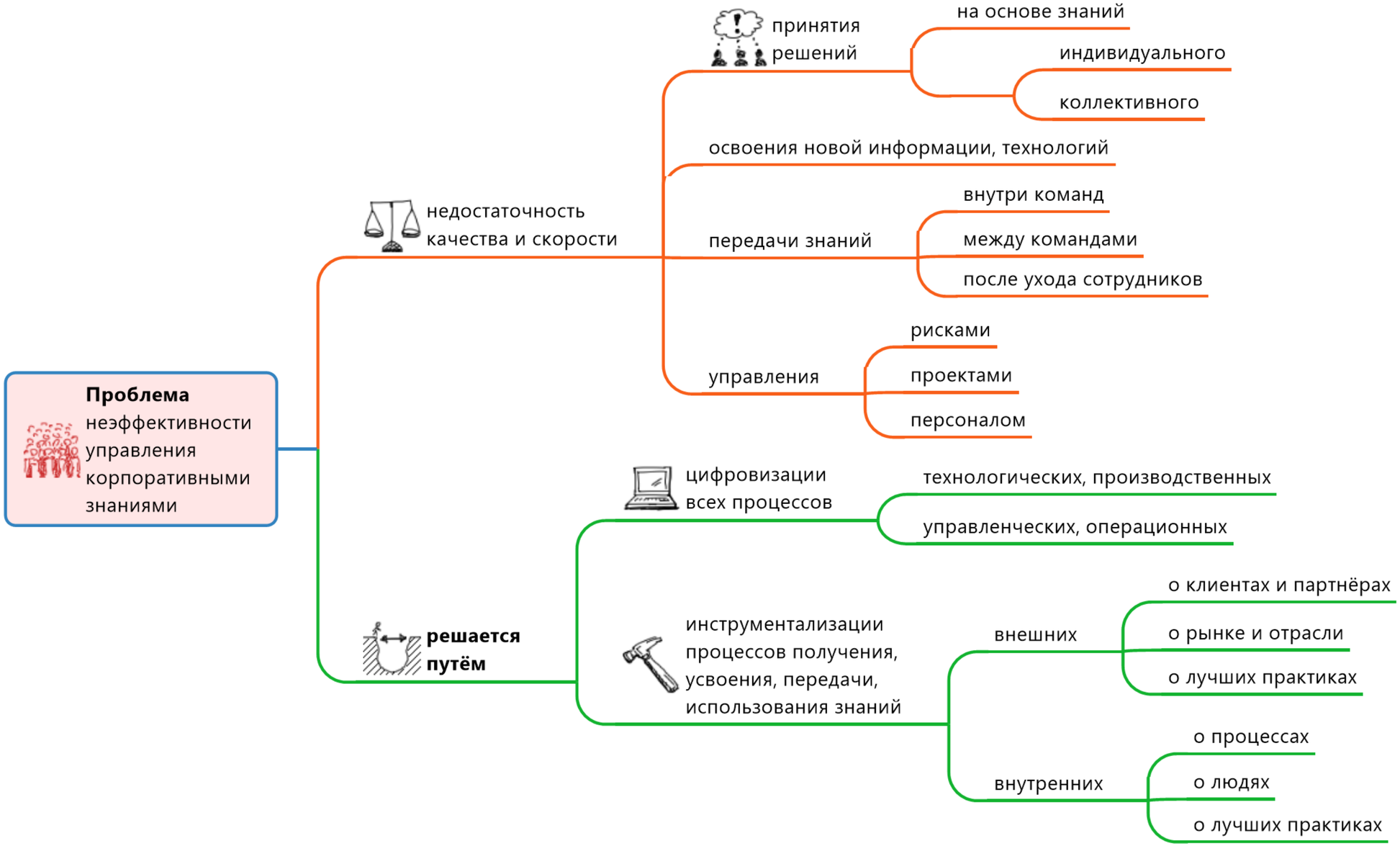
структурированность, выделение главного
- в обучающих данных для LLM
- при коммуникации с людьми

решается путём

текстографического структурирования профессиональной коммуникации

массового перехода к **нелинейному чтению** (non-linear reading) по принципу «сначала главное, детализация по требованию»

Проблемы, решаемые картами Знаний



Проблемы, решаемые картами Знаний

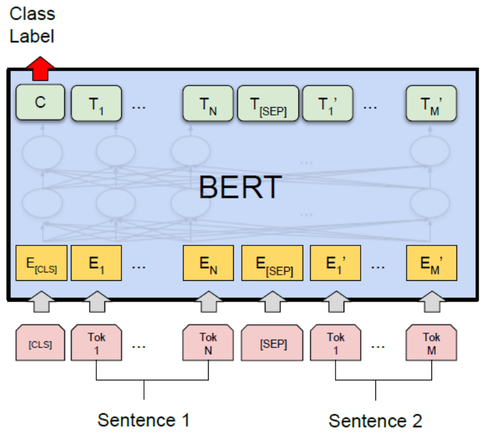
Выводы про карты знаний

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

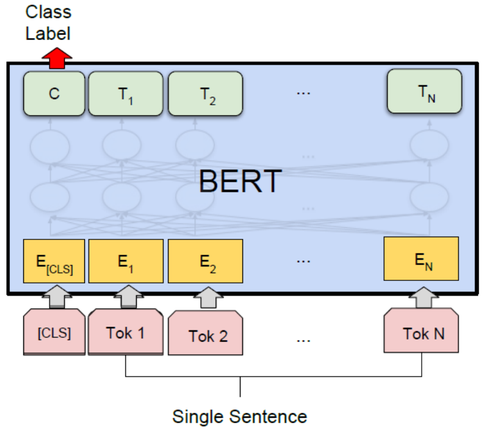
- **Важные навыки** для работы с научной информацией
  — во всём выделять главное (7±2),
  — делать это быстро, формулировать лаконично
- **Прежде чем обучать ИИ** по тексто-графическим представлениям,
  — необходимо освоить их самим,
  — в своей практической деятельности,
  — как индивидуальной, так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

Трансформеры: нейросетевые модели языка

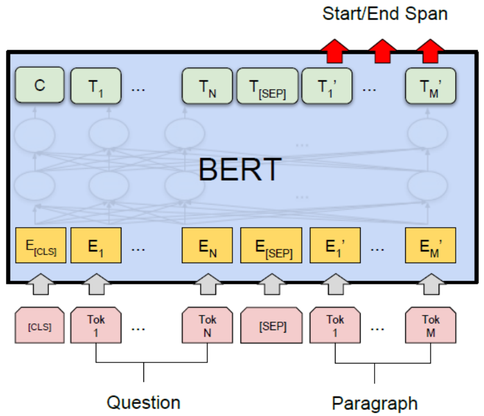
- обучаются векторизовать и предсказывать слова по контексту
- обучаются по терабайтам текстов, «они видели в языке всё»
- мультиязычны: обучаются на десятках языков
- мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



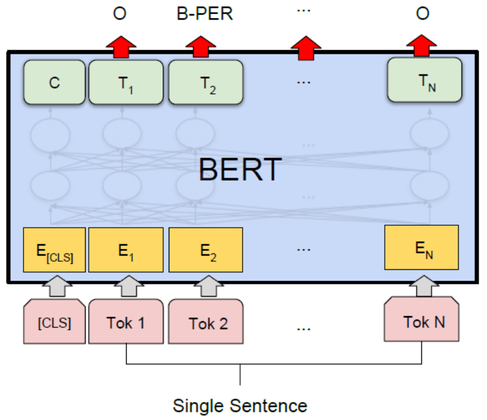
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



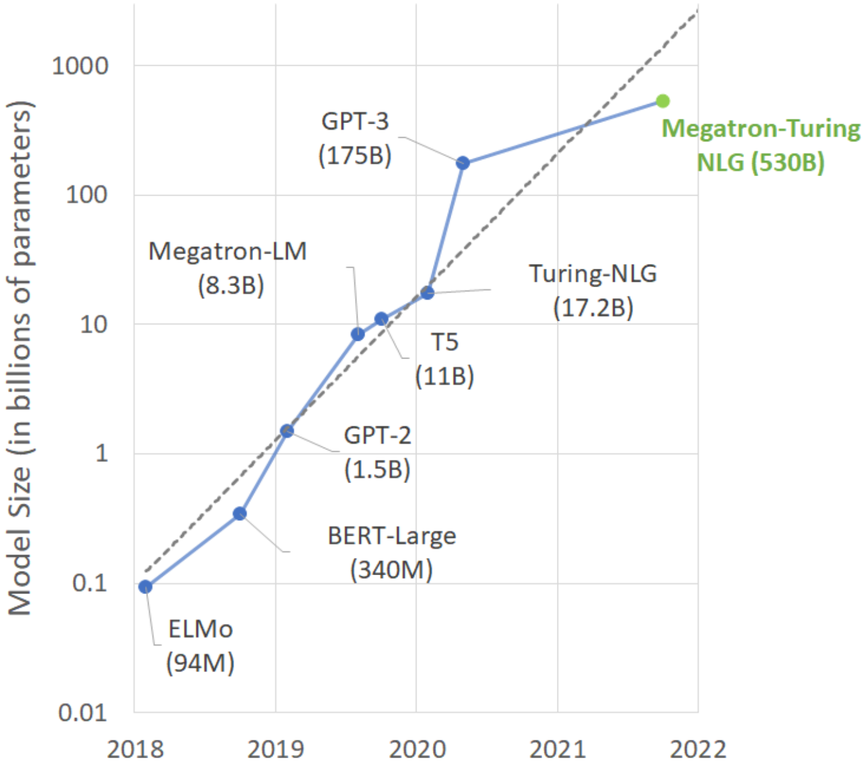
(c) Question Answering Tasks: SQuAD v1.1

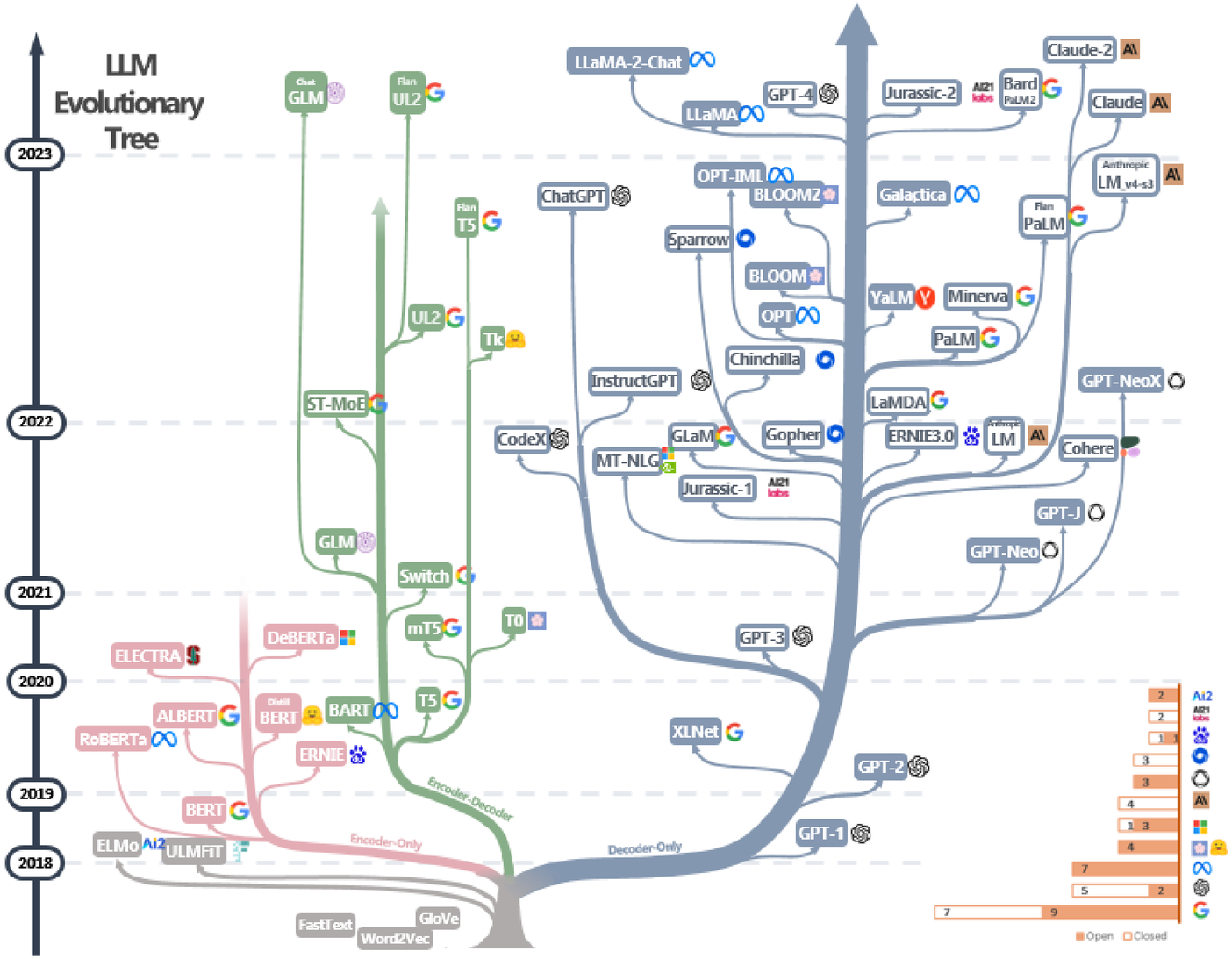


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Трансформеры: размер имеет значение

Рост числа параметров больших языковых моделей





Проблески общего искусственного интеллекта

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

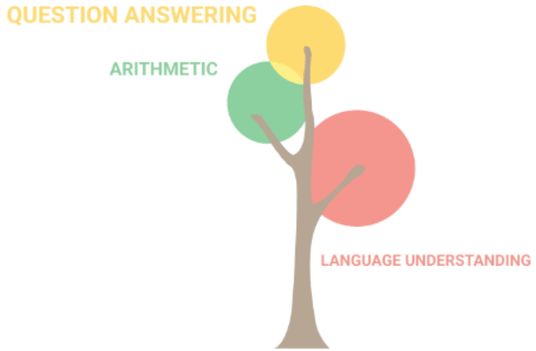
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Новые способности модели GPT, не закладывавшиеся при её 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

Новые (эмерджентные) способности модели

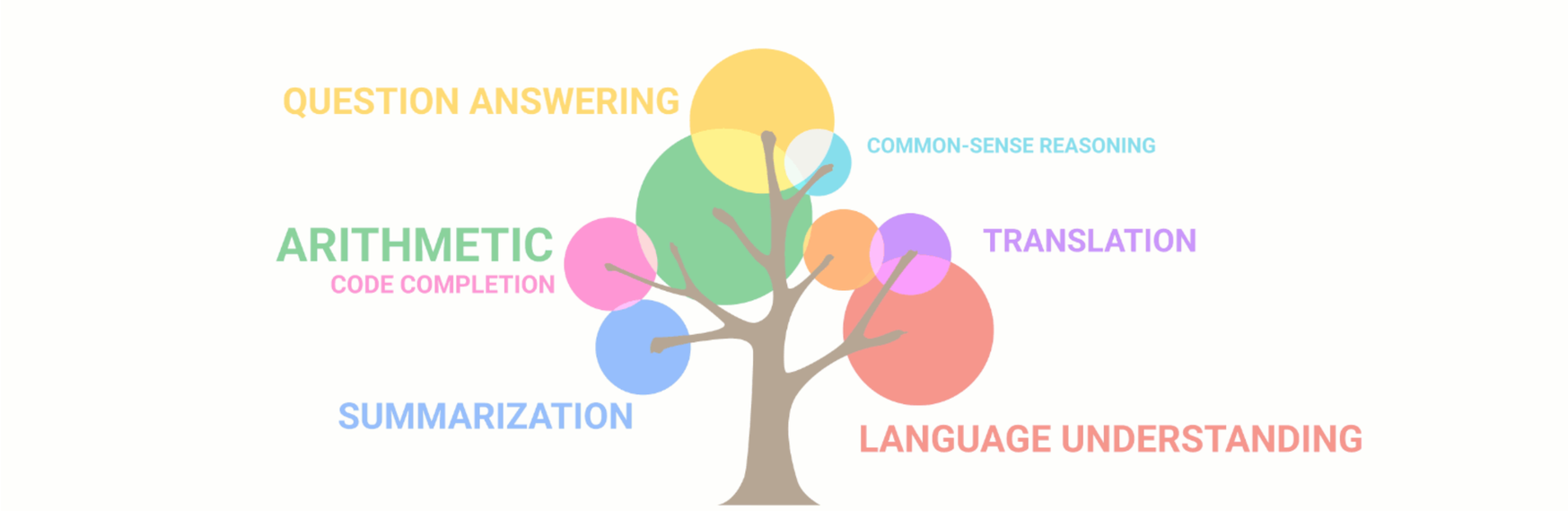


GPT-2: 14-Feb-2019

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

Новые (эмерджентные) способности модели

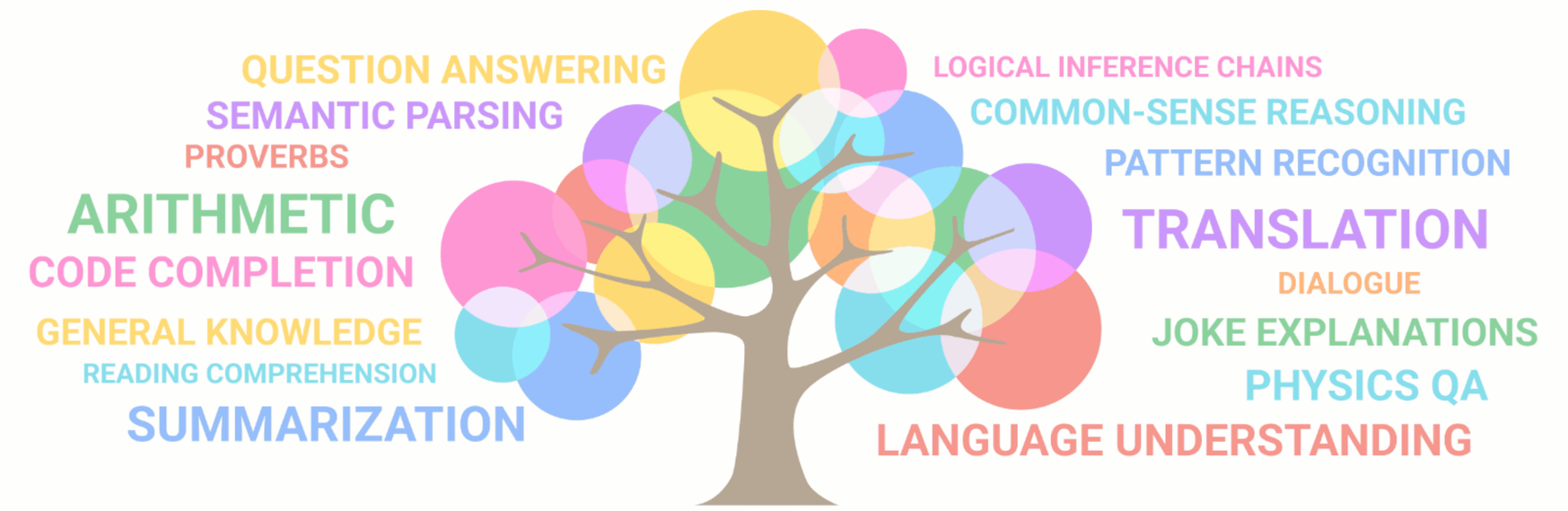


GPT-3: 11-Jun-2020

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

Новые (эмерджентные) способности модели



GPT-4: 14-Mar-2023

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

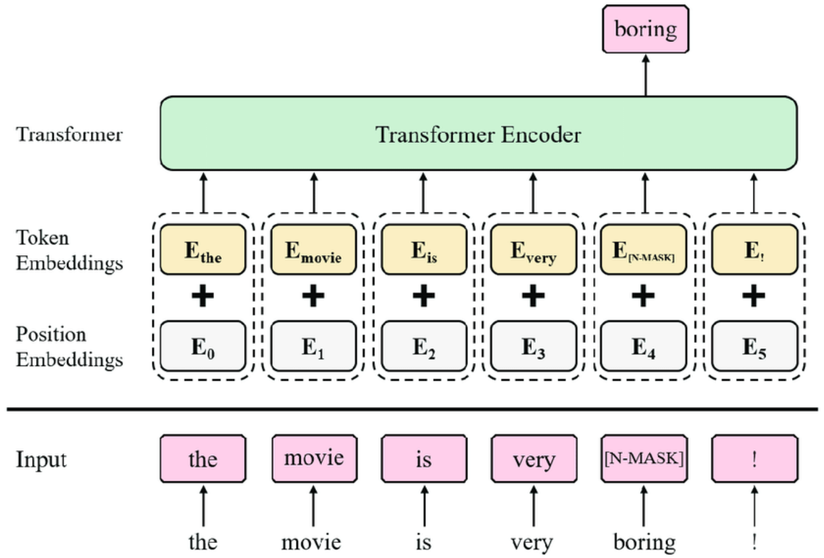
- способность описывать и анализировать изображения
- способность реагировать на подсказки вроде «Let's think step by step»
- способность решать качественные физические задачи по картинке

Возможности и угрозы: чаты GPT способны

- помогать с рутинной интеллектуальной работой
- искать и структурировать профессиональную информацию
- делать обзоры, рефераты, сводки на разных языках
- генерировать тексты и сайты по тех. заданию, в том числе технические, медицинские, юридические
- генерировать программный код по описанию
- обсуждать новости, поддерживать разговор по теме
- разговаривать с детьми с учётом возрастных особенностей
- выполнять функции воспитателя, учителя, наставника
- оказывать психологическую помощь

- «галлюцинировать», давать неверные сведения, касающиеся здоровья человека, законов, событий, технологий, других людей
- вызывать необоснованное доверие и манипулировать человеком
- переубеждать, побуждать человека к действиям, не выгодным ему
- поддерживать предрассудки и лженаучные представления
- поддерживать пропагандистские медиа-кампании
- неконтролируемо влиять на формирование мировоззрения у детей, подростков
- оказывать депрессивное воздействие на психику

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

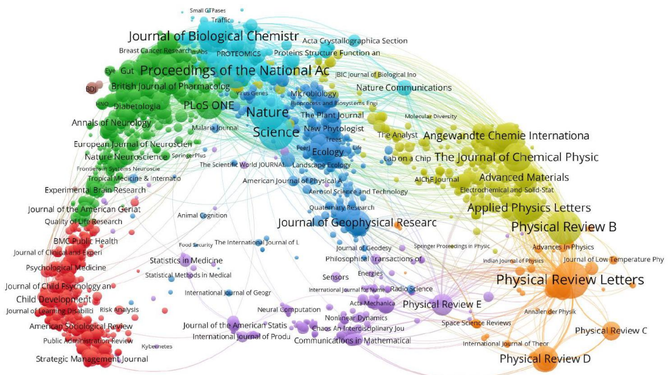
SEMANTIC SCHOLAR

- **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)

SEMANTIC SCHOLAR

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: 12 задач

- 6 задач классификации OECD/ГРНТИ по аннотации ru / en / ru+en
- 2 задачи кросс-язычного поиска ru→en / en→ru
- 2 задачи предсказания цитирования / социтирования
- 2 задачи регрессии: предсказание года и цитируемости публикации

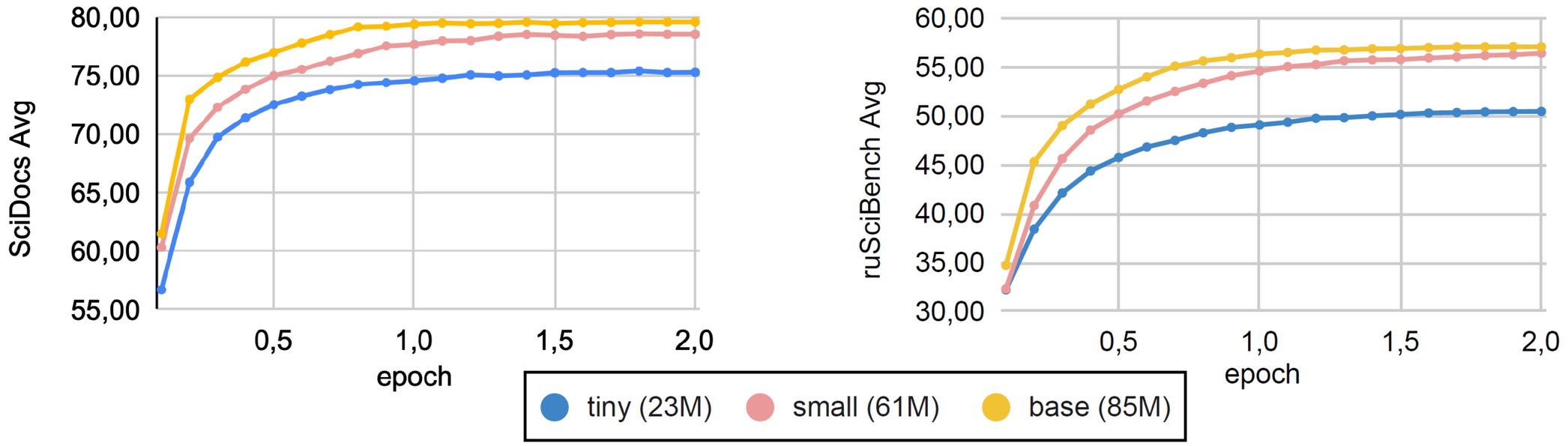
МГУ

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)

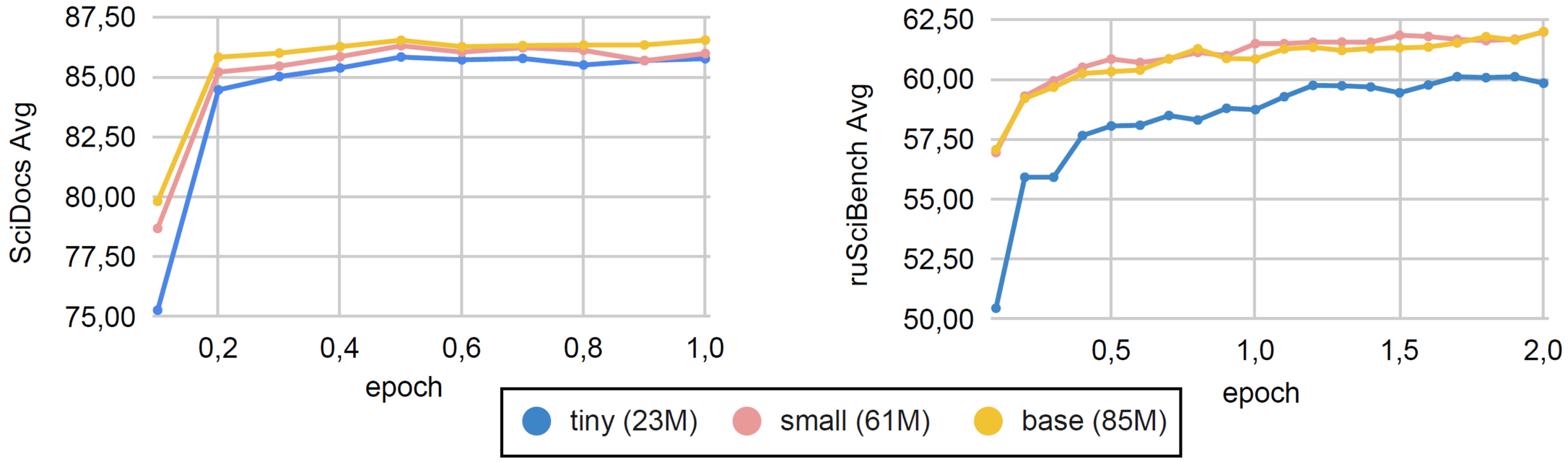
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



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



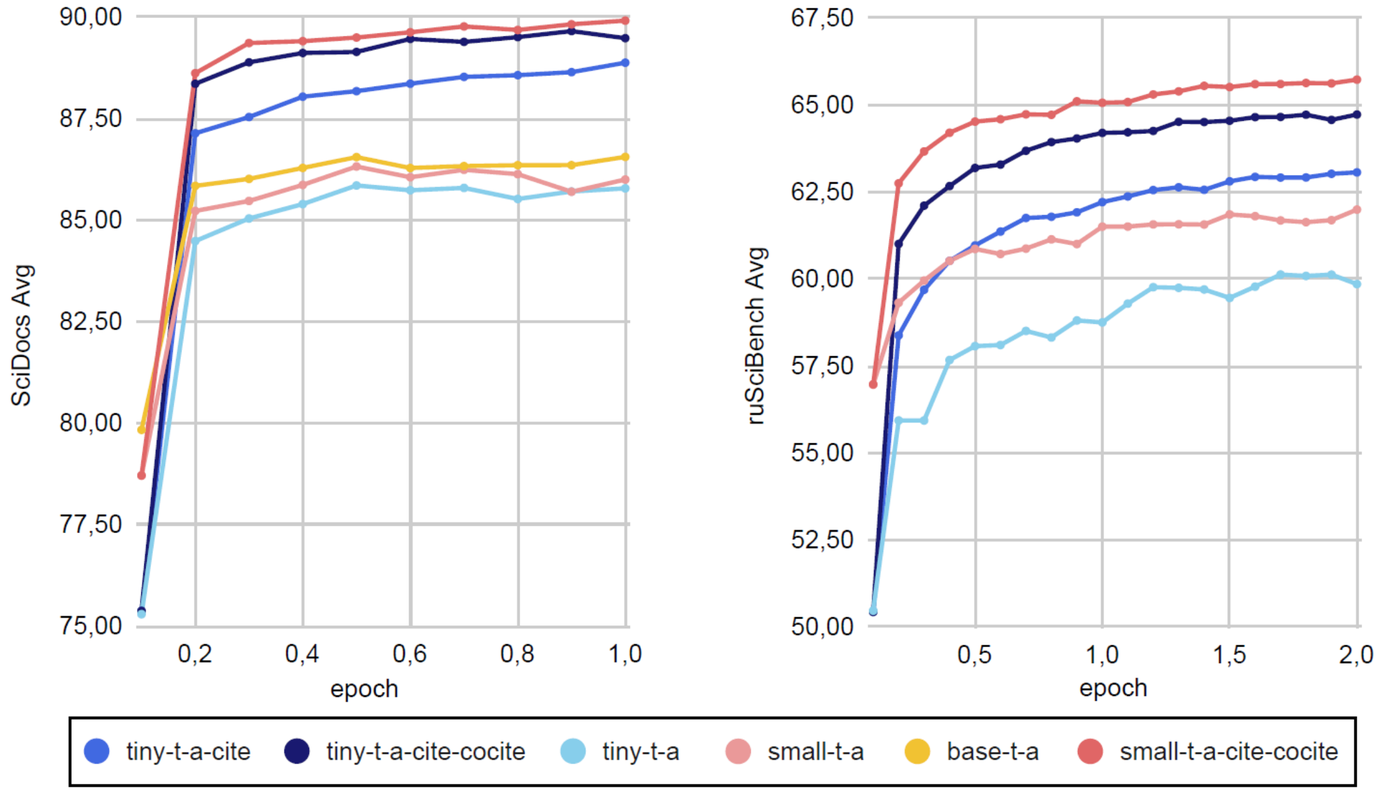
L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

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 ruSciBench

	model_name	Model size	elibrary_oecd_full	translation_search	
			macro_f1	ru_en recall@1	en_ru recall@1
SOTA	e5-mistral-7b-instruct	7.11B	**67,28**	3,65	18,11
	scirus-tiny3.1	23M	65,40	97,40	98,80
	multilingual-e5-large	560M	63,70	**99,19**	**99,37**
	scirus-tiny2	23M	62,02	96,70	95,11
	multilingual-e5-base	278M	62,00	97,00	98,00
	LaBSE	471M	60,21	98,31	97,20
	LaBSE-en-ru	128M	60,05	98,26	96,93
	paraphrase-multilingual-mpnet-base-v2	118M	60,03	66,33	78,18
	FRED-T5-large	360M	59,80	22,25	0,79
	distiluse-base-multilingual-cased-v1	135M	58,69	92,04	90,83
	paraphrase-multilingual-MiniLM-L12-v2	118M	56,48	72,87	77,49
	mfaq	280M	54,84	86,75	90,11
	scirus-tiny	23M	54,83	88,00	88,00

- Сильнее модели, которая в ~20 раз больше
- Приблизились вплотную к SOTA, которую держит модель в ~300 раз больше

Сравнение моделей по метрикам SciRepEval

Model name	Model size	SciDocs	Out-of-Train	In-Train
all-mpnet-base-v2	110M	**91,03**	50,2	53,12
scincl	110M	90,84	**51,8**	55,6
scirus-tiny3.1	23M	90,1	50,08	**57,2**
SPECTER	110M	89,10	50,6	54,7
e5-large-v2	335M	88,70		
e5-base	109M	88,58		
e5-base-v2	109M	88,43		
multilingual-e5-large	560M	87,53	49,32	55,65
e5-small-v2	33.4M	86,99		
multilingual-e5-base	278M	86,91		
e5-mistral-7b-instruct 4byte	7.11B	86,03		
scirus-tiny2	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	23M	67,92		
nreimers/MiniLM-L6-H384-uncased (e5-small-v2 pretrain)	33.4M	65,68		

🏆 **SOTA (In-Train)**

- Топ-3 в **SciDocs** и **Out-of-Train** (конкуренты в ~5 раз больше), SOTA в **In-Train**

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 91.03
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)

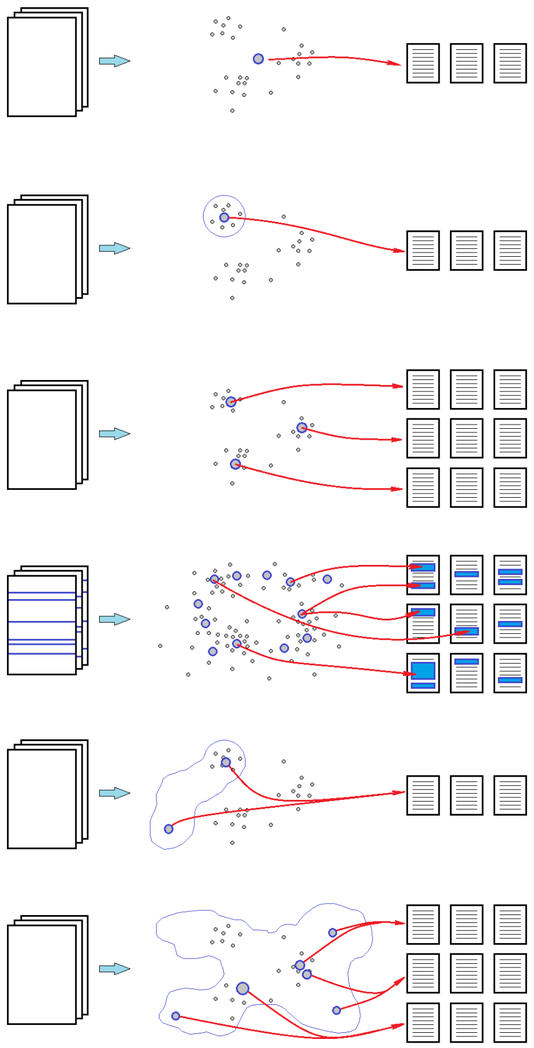
Первое внедрение

«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

Стратегии векторного поиска документов



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

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

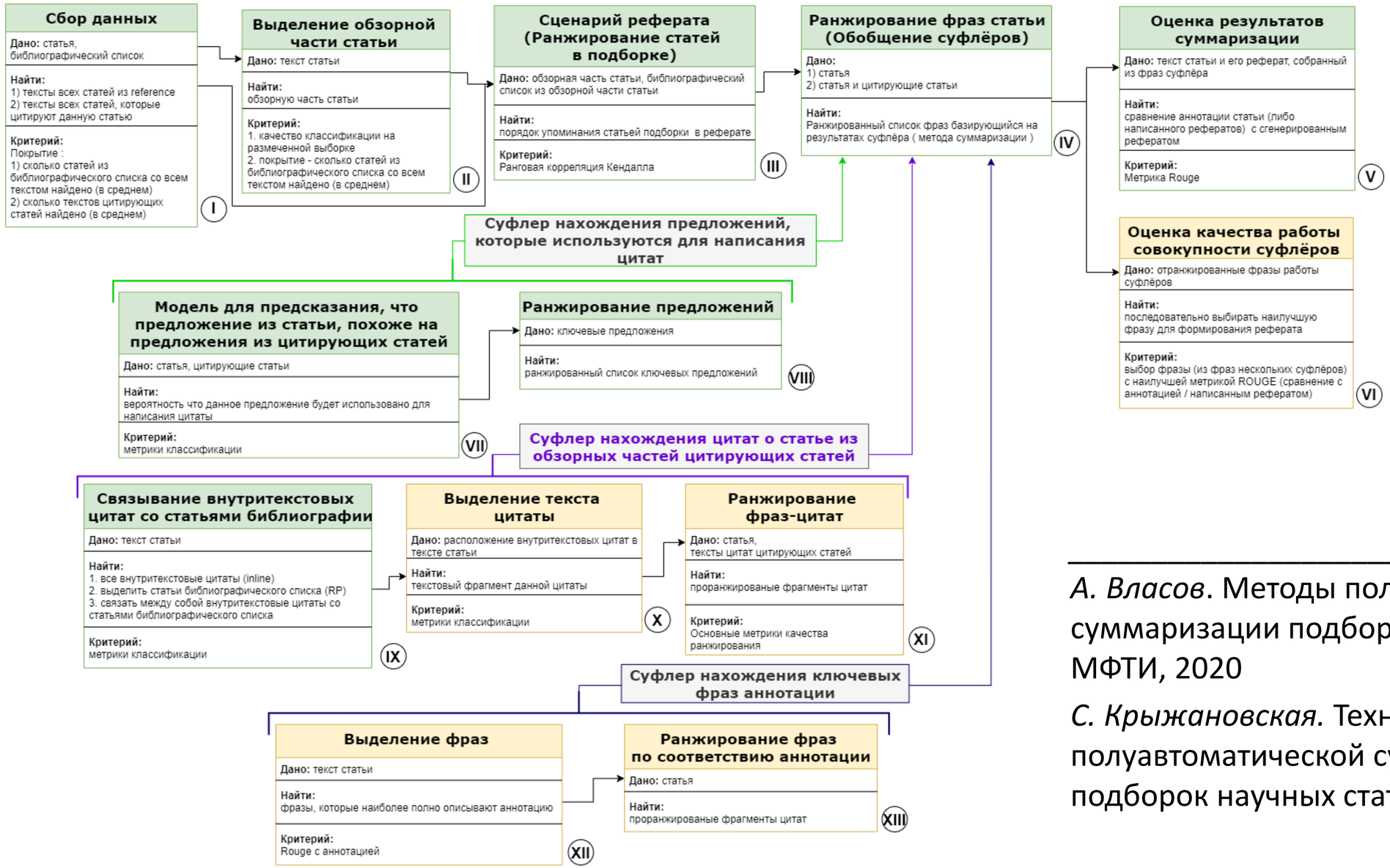
Основные задачи машинного обучения:

- 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

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:

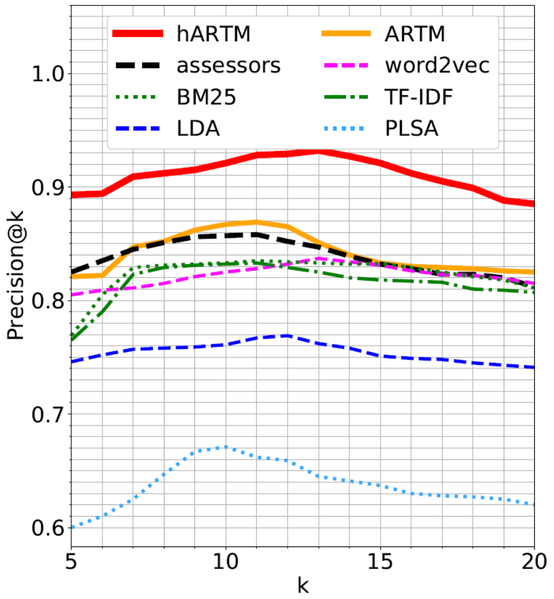
- длинные запросы (1 стр. A4)
- 100 запросов на коллекцию
- 3 асессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:

(170K Russian docs)

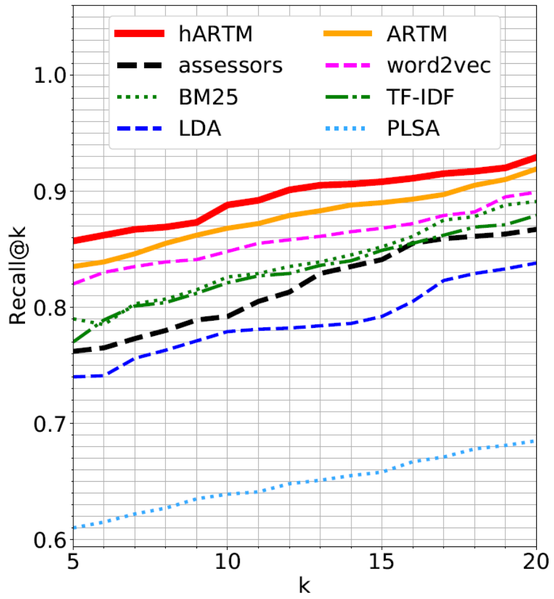
(750K English docs)

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.

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

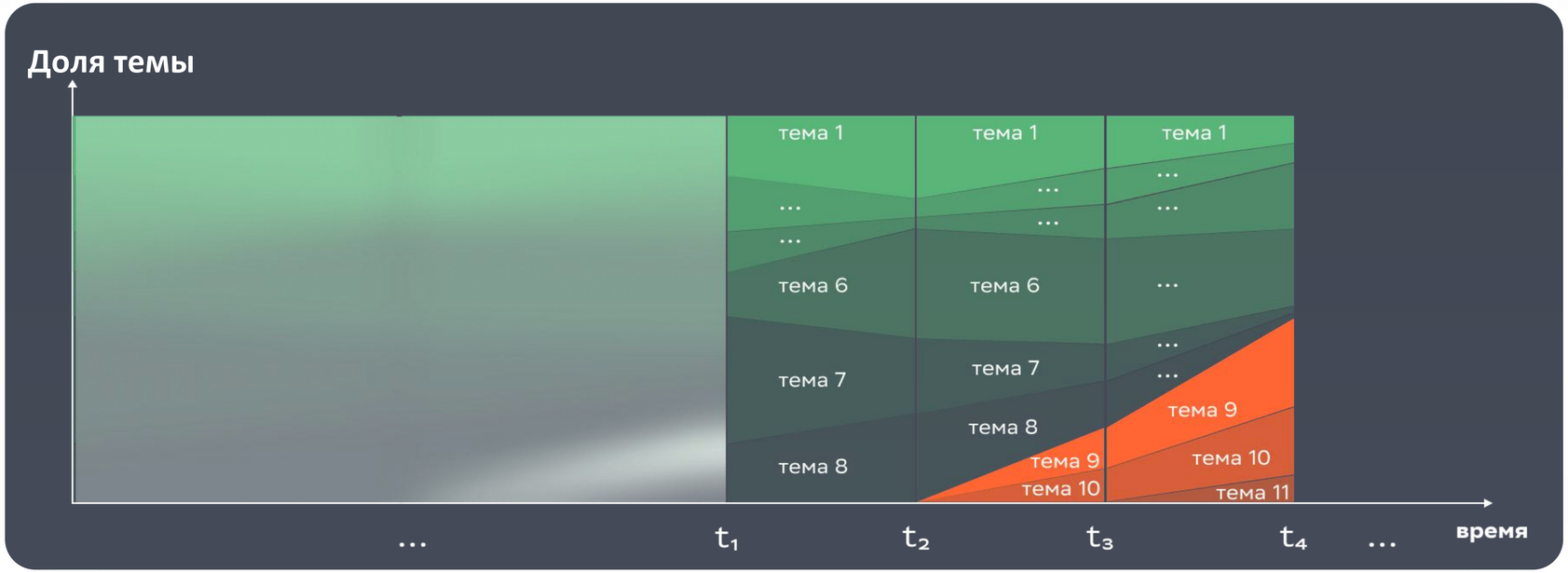
Особенности соревнования:

- 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: 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

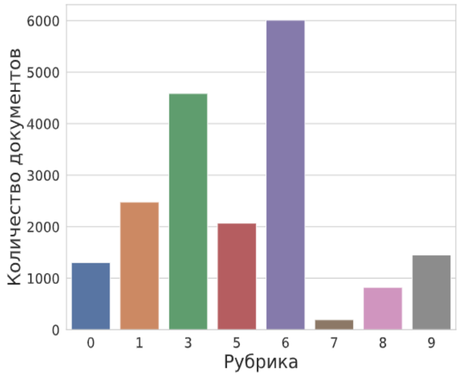
Поиск научных трендов: примеры тем

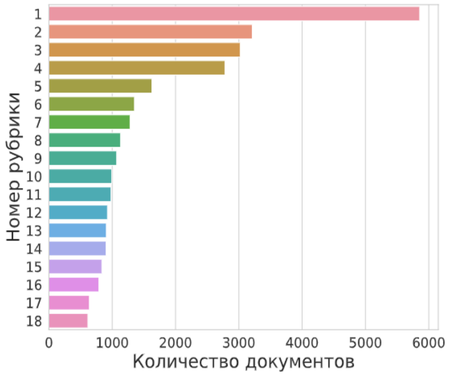
StyleGAN	Meta Learning	NERF
stylegan	meta model	neural radiance field
latent code	meta train	accurate depth estimation
mapping network	meta optimization	additional qualitative result
ablation study	meta update	novel loss function
text generation	meta testing	optical flow prediction
generation quality	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

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%

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

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

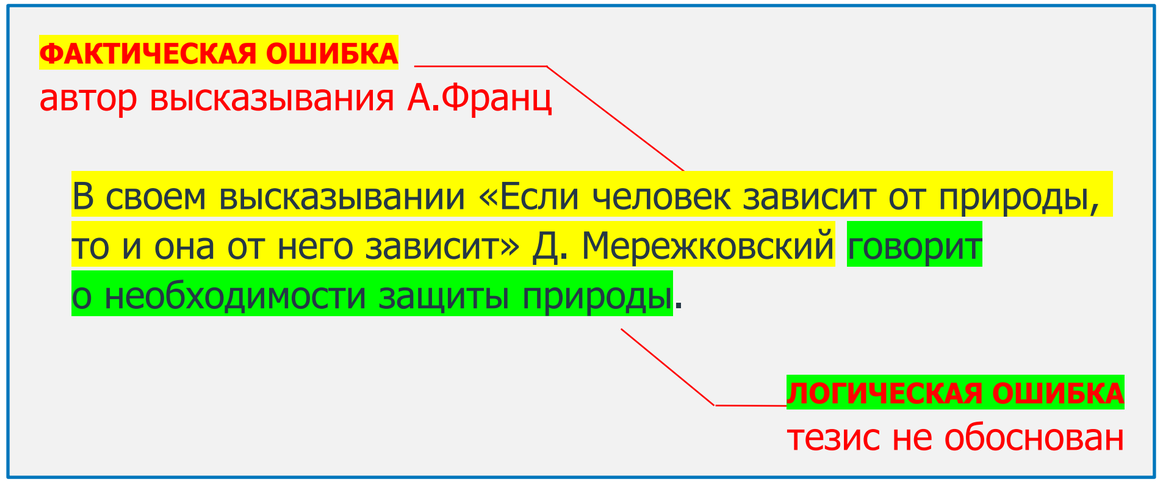
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

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

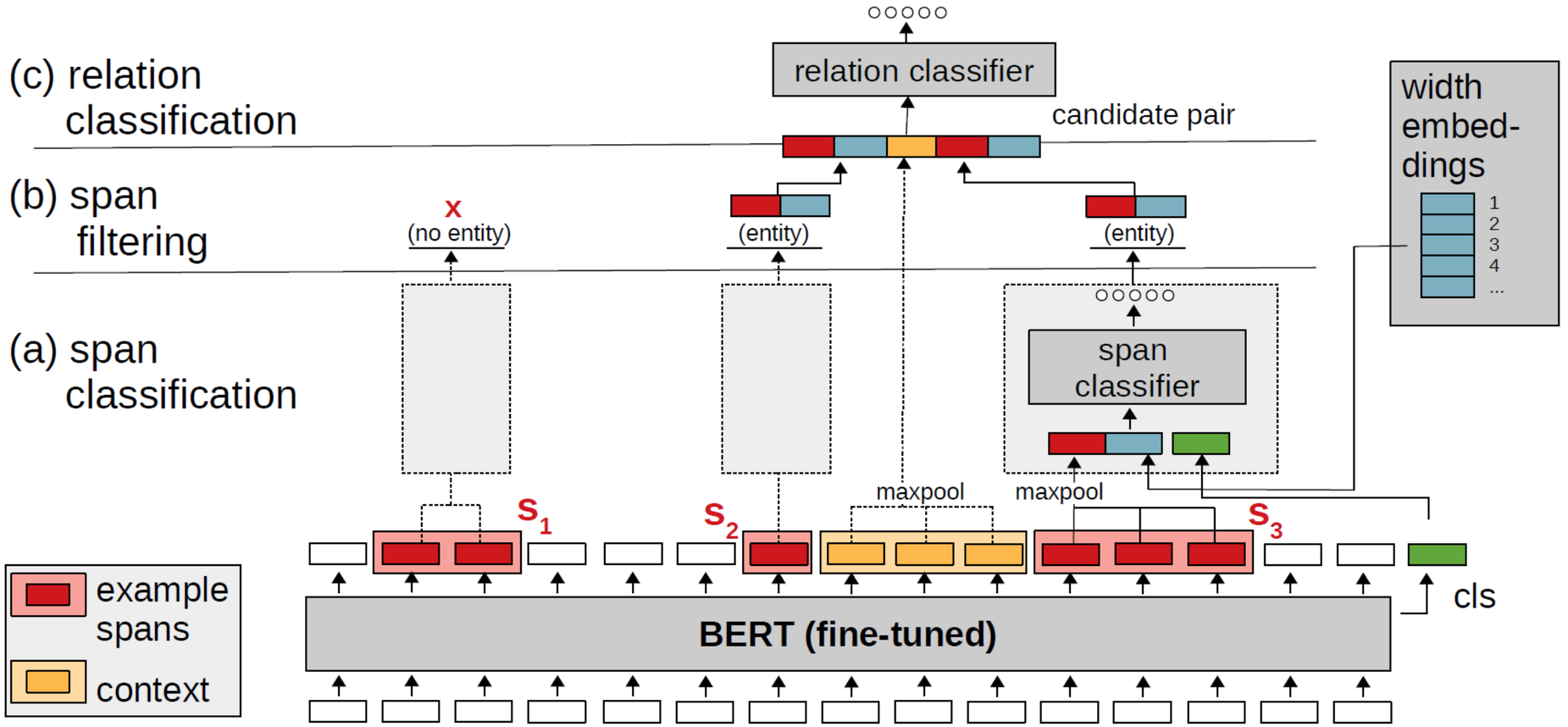
Commissio
PopulusQue
Europaea

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.

Нейросетевые обучаемые модели разметки

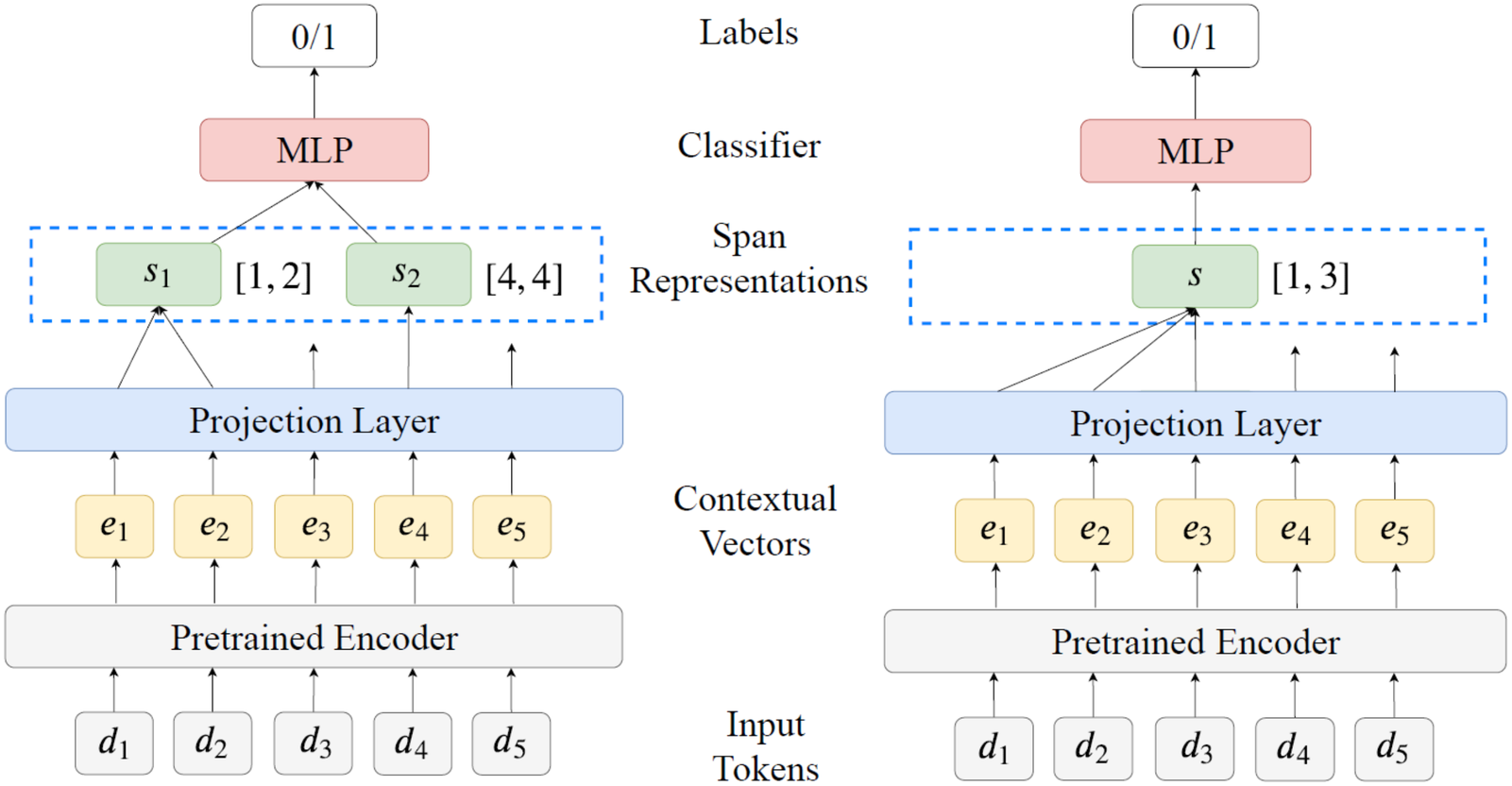


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

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.

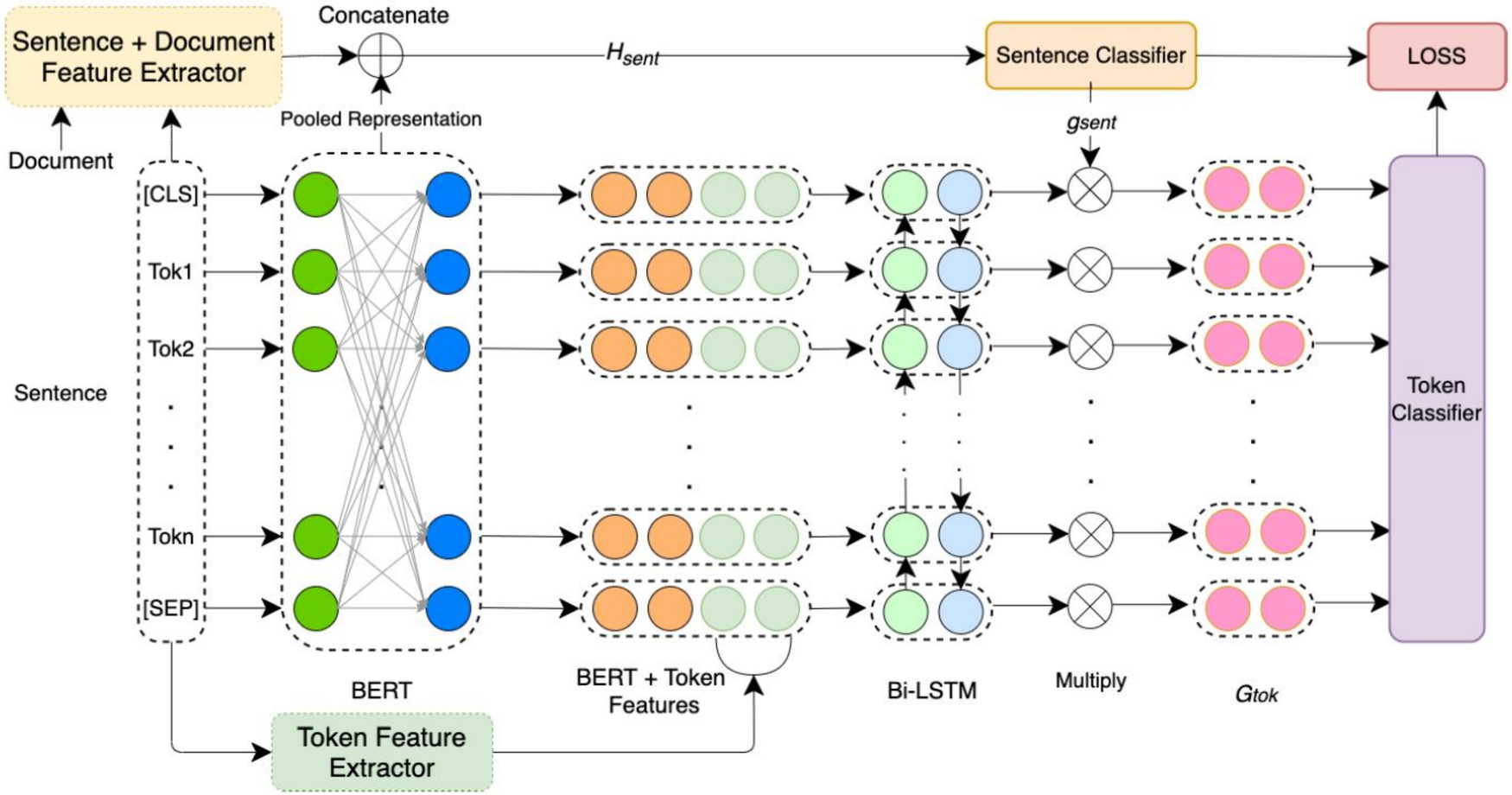
Нейросетевые обучаемые модели разметки



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



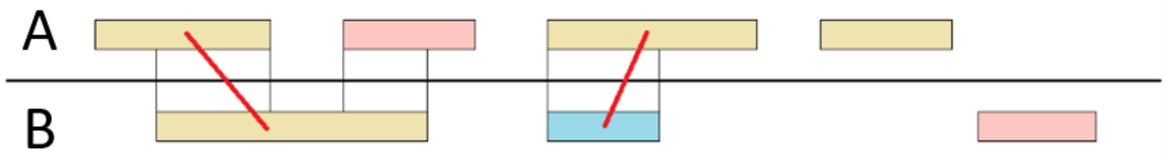
Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

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

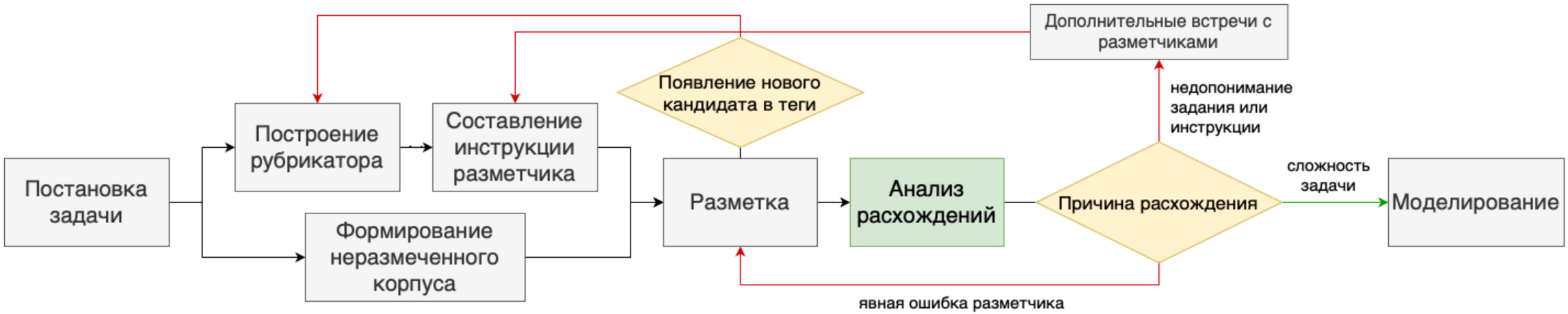
Con2 = точность наложения сопоставленных фрагментов

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

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:** 1:поиск, 2:мониторинг, 3:реферирование, 4:тематизация, 5:картирование, 6:онтологизация, 7:хронологизация, 8:персонализация, 9:анализ трендов, 10:контент-анализ
- **источники:** научные статьи, патенты, документация, новости,…
- **мультиязычность:** русский—английский—китайский—…