Международная научная конференция
«Манипулятивные процессы в медиадискурсе: реальность, ментальные модели, язык»
10-11 апреля 2024 г., РГГУ

Нейросетевые модели разметки текста: от выявления манипулятивных воздействий к автоматизации контент-анализа

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. лабораторией *машинного обучения и семантического анализа*

Института искусственного интеллекта МГУ им. М.В. Ломоносова

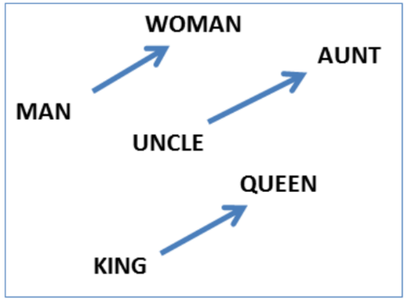
Эволюция подходов в обработке естественного языка

Как решали задачи анализа текстов 10 лет назад

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

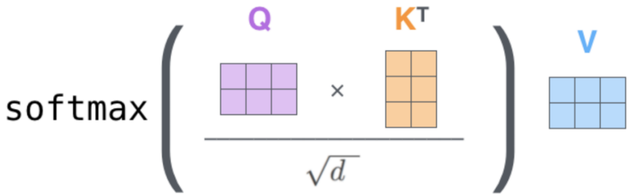
Модели векторизации слов (эмбединги слов)

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

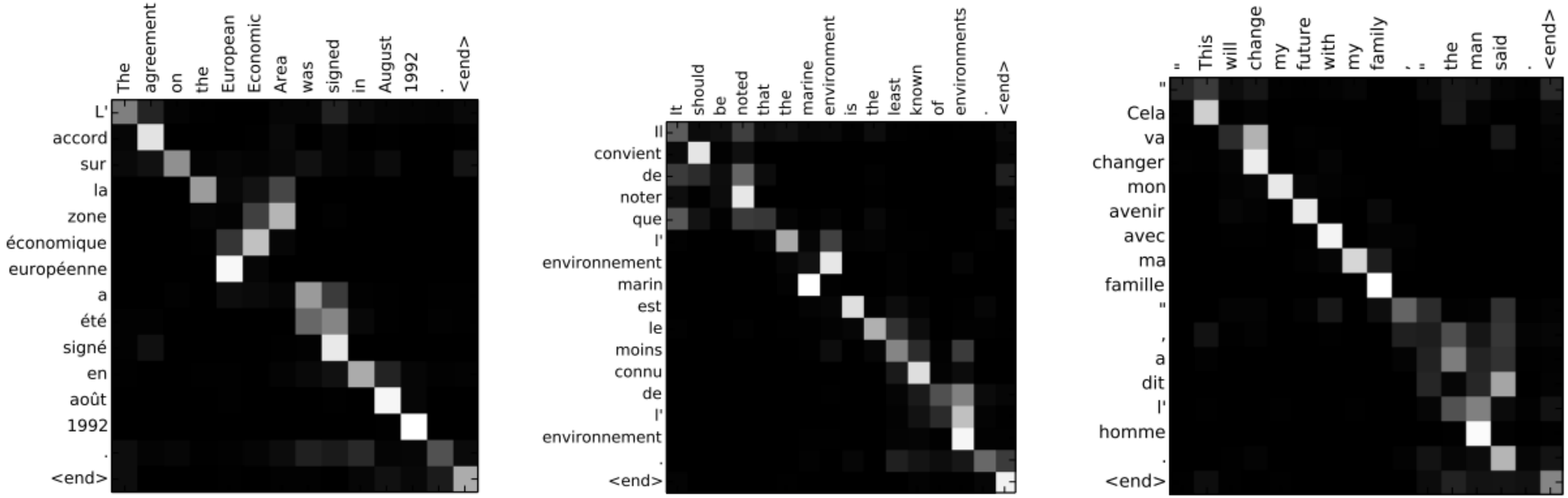


Нейросетевые модели контекстной векторизации

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



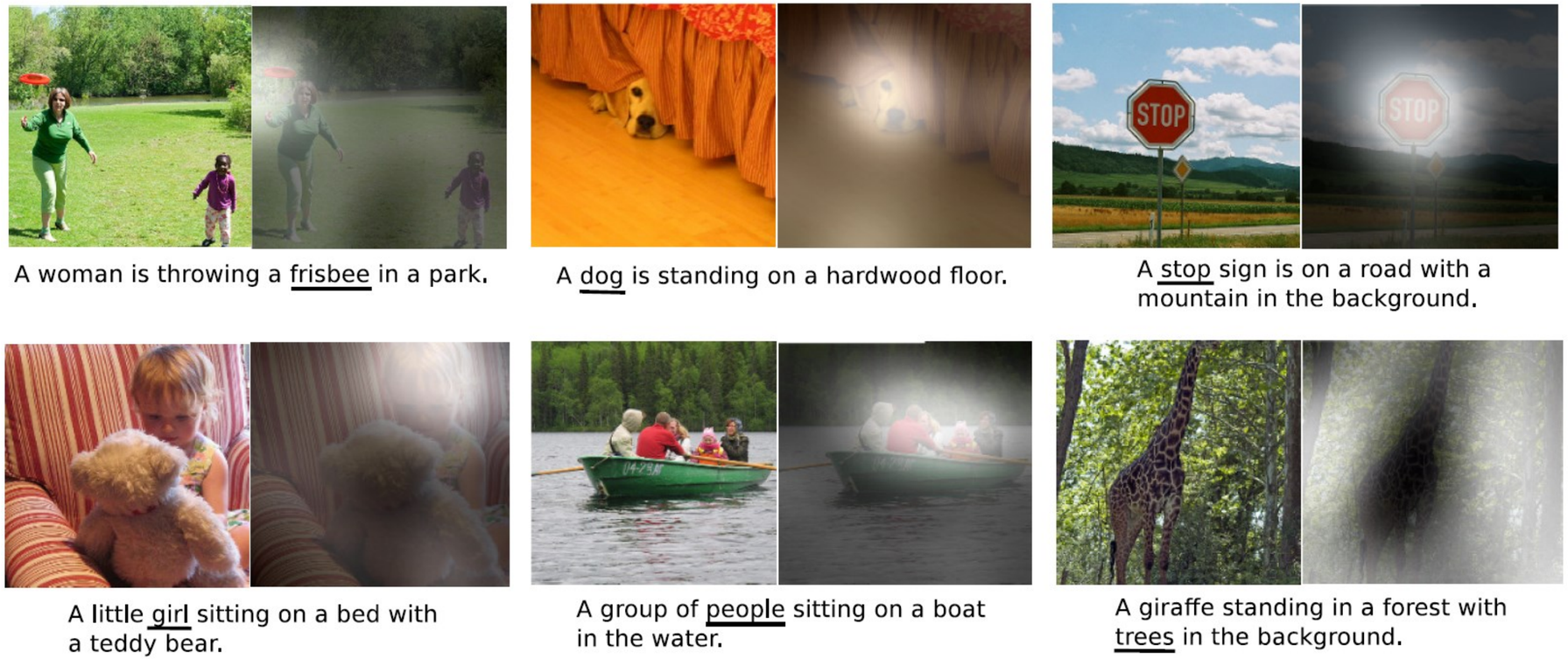
Модели внимания: машинный перевод



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

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015

Модели внимания: аннотирование изображений

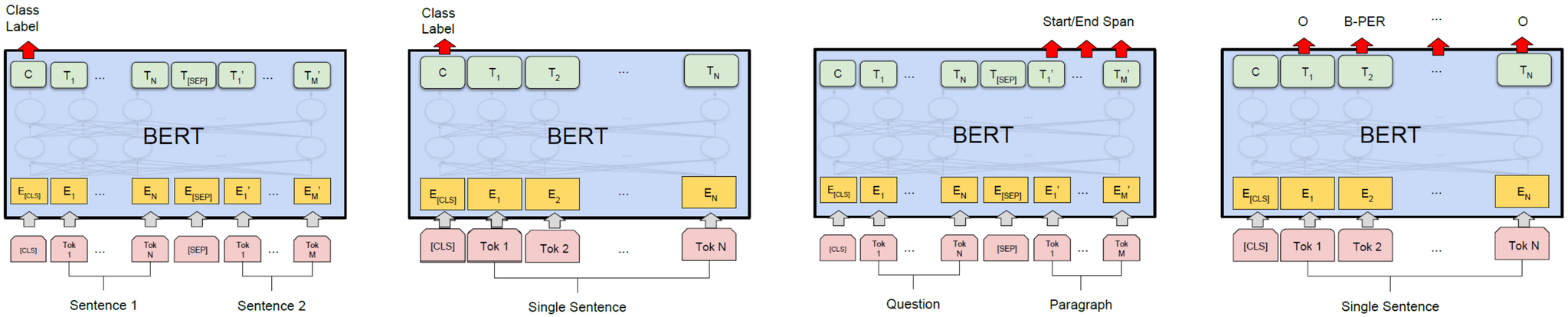


Интерпретация: на какие области модель обращает внимание, когда генерирует подчёркнутое слово в описании изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

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

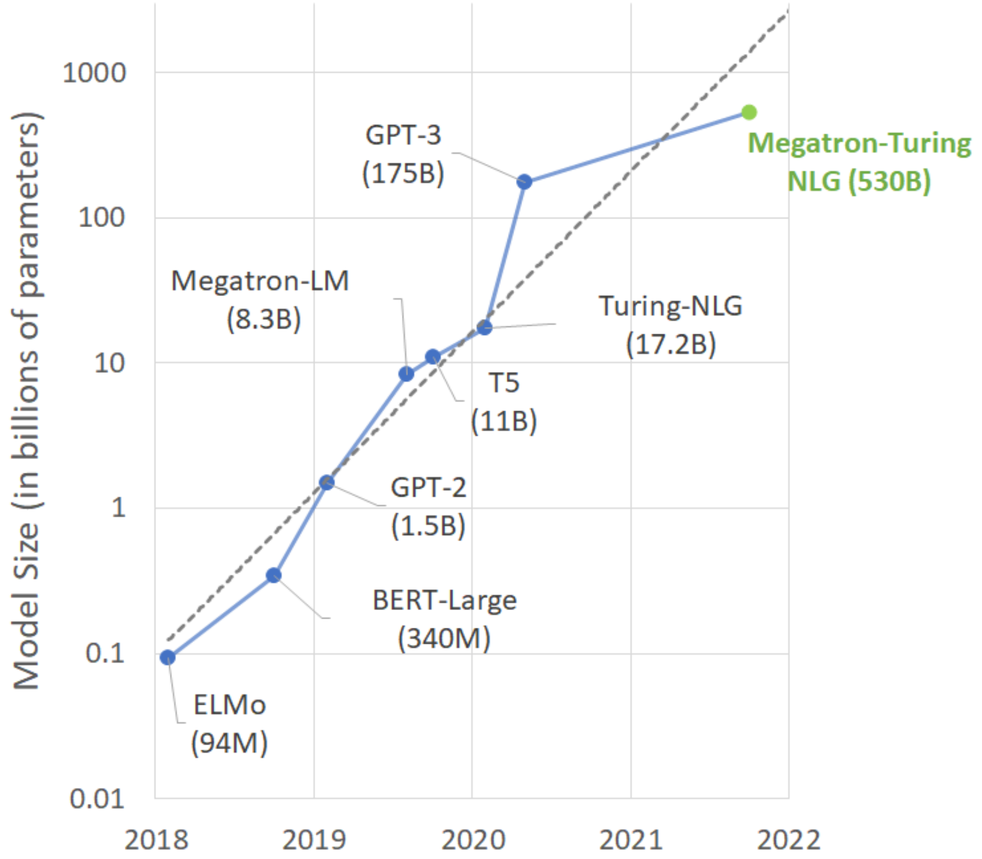
(b) Single Sentence Classification Tasks: SST-2, CoLA

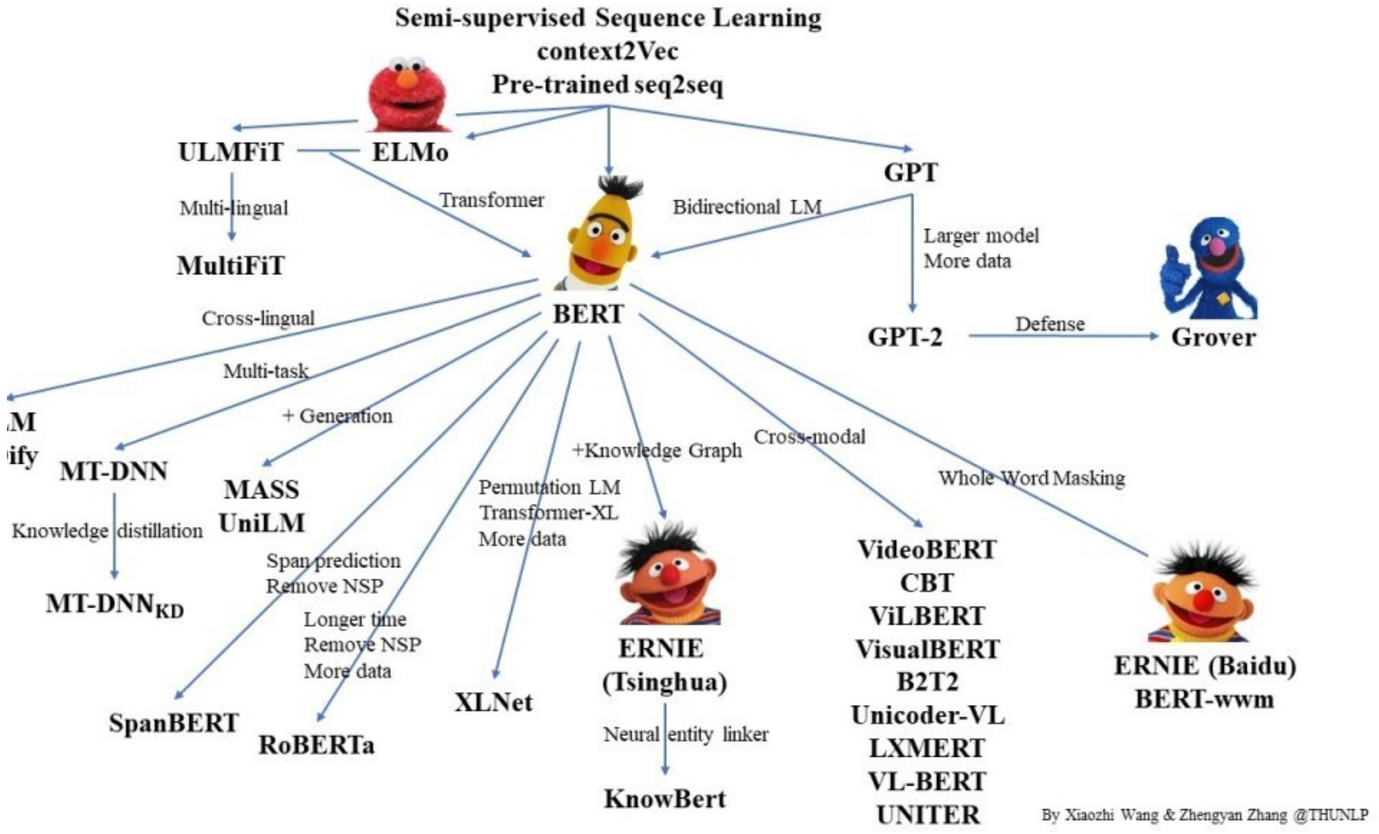
(c) Question Answering Tasks: SQuAD v1.1

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

Трансформеры: нейросетевые модели языка

Рост числа параметров нейросетевых трансформерных моделей языка





Проблески общего искусственного интеллекта

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

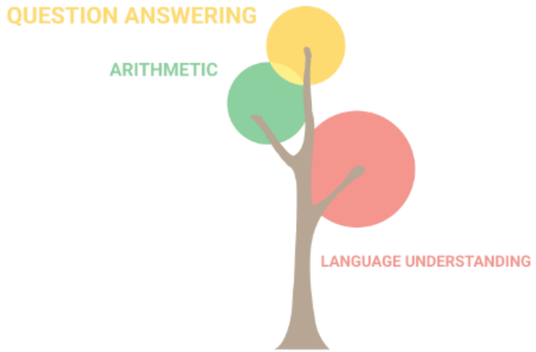
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Новые способности модели, не закладывавшиеся при обучении:

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

Эмерджентные (не ожидавшиеся) способности модели

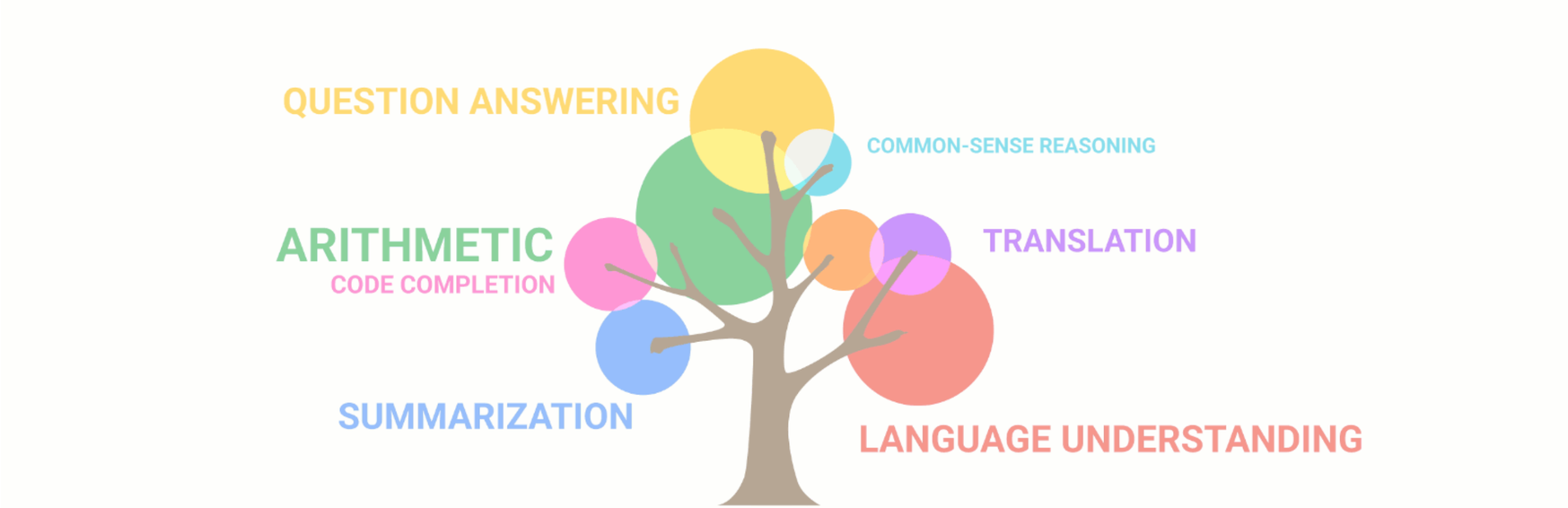


GPT-2: 14-Feb-2019

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

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

Эмерджентные (не ожидавшиеся) способности модели

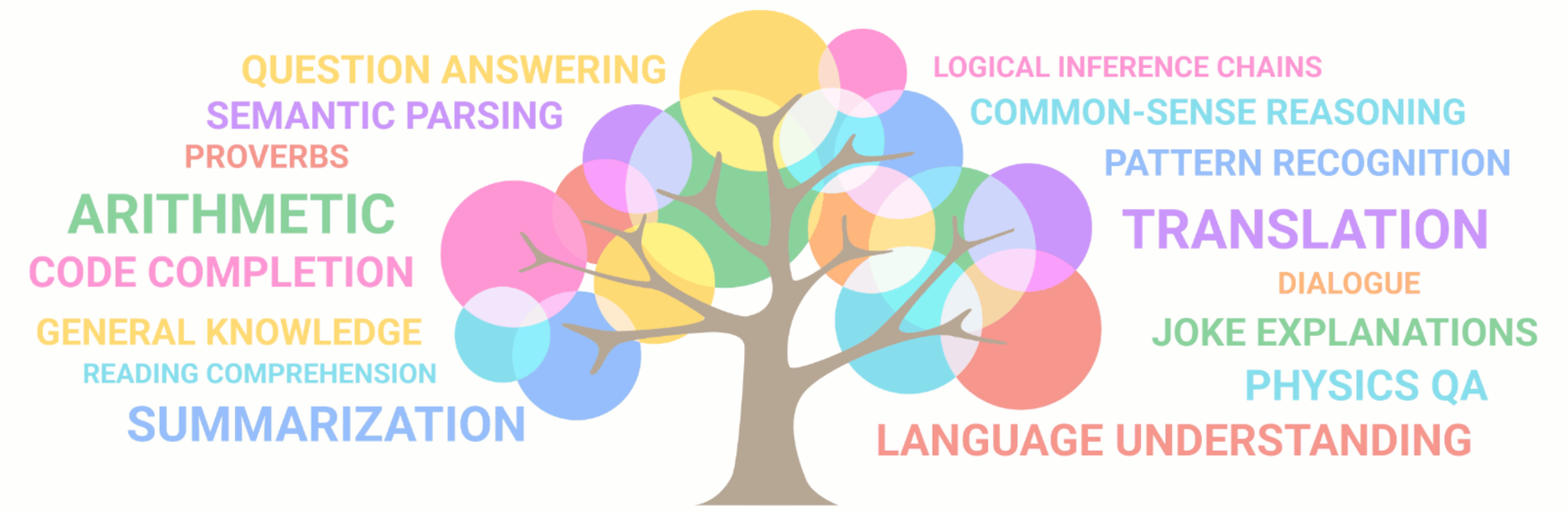


GPT-3: 11-Jun-2020

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

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

Эмерджентные (не ожидавшиеся) способности модели



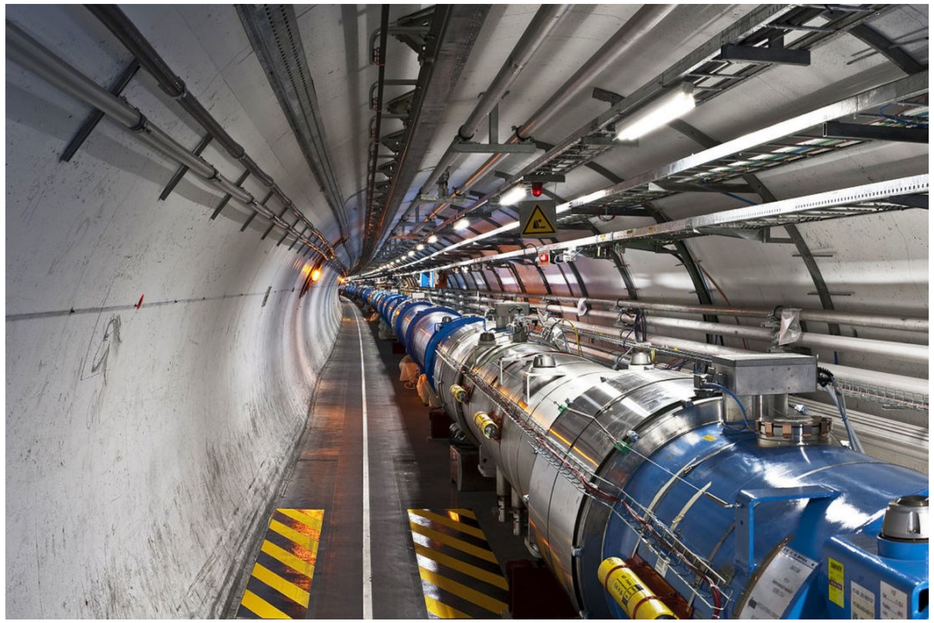
GPT-4: 14-Mar-2023

>1 трл. параметров, корпус >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

Концепция проекта «Новостной коллайдер»

Цель создания адронного коллайдера
— сталкивая потоки частиц,
узнать больше о строении материи



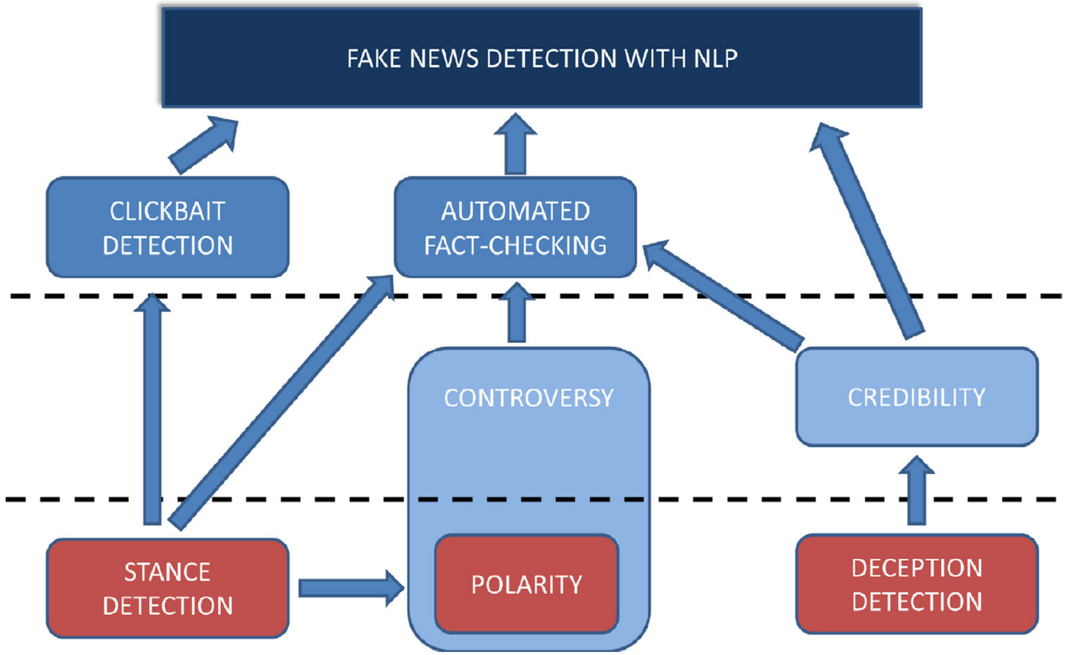
Цель создания новостного коллайдера
— сталкивая потоки новостей,
узнать больше о когнитивных войнах
и выработать защиты от манипуляций

Явления постправды

- Факты становятся менее значимы, чем эмоции и личные убеждения
- Явление «информационных пузырей»
- Явление фейков и «неопровержимой лжи»
- Явление замалчиваний: постправда маскируется под «другие грани истины»
- Постправда — инструментарий пропаганды и когнитивных войн

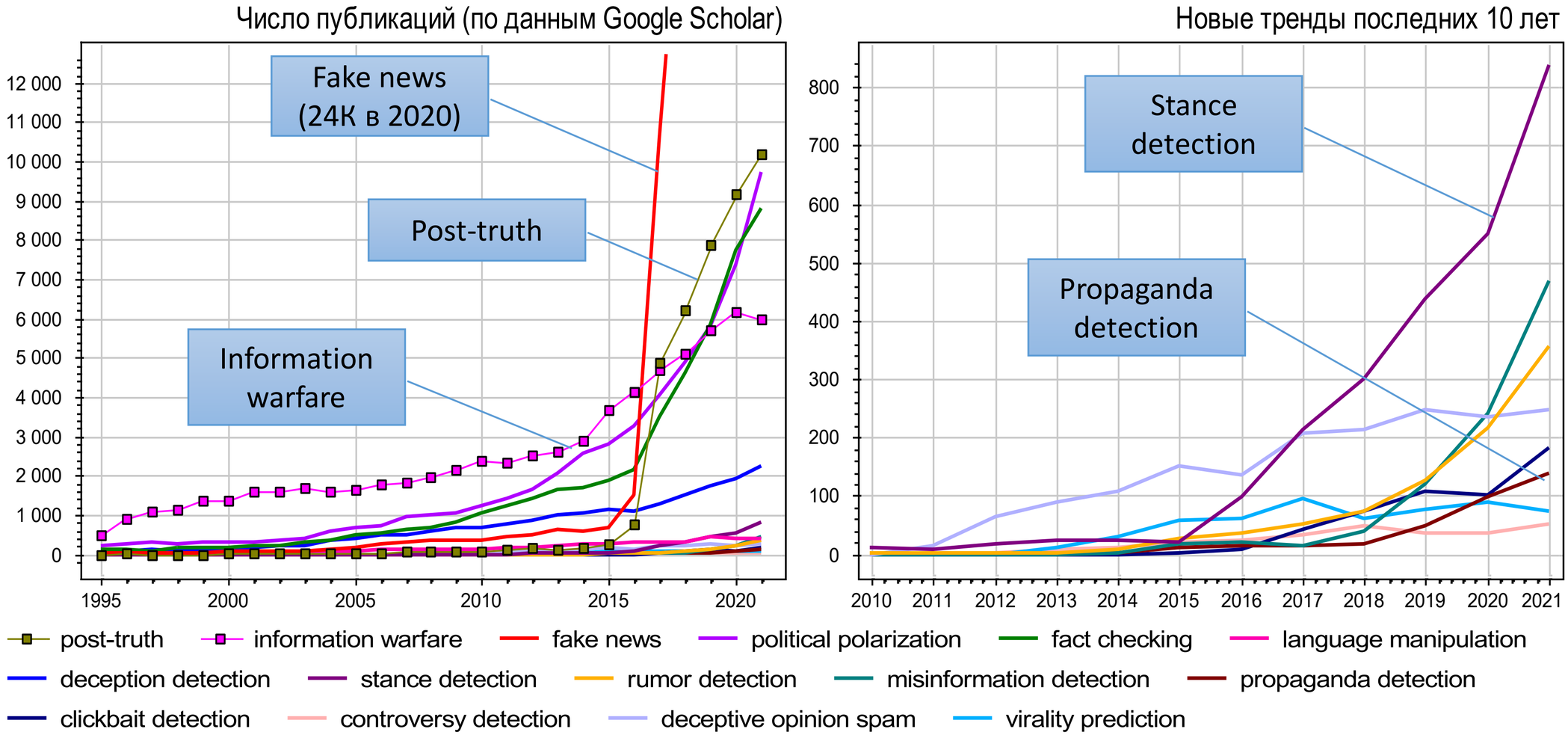
Область исследований «Fake News Detection»

1. Deception Detection
   выявление обмана в тексте новости
2. Automated Fact-Checking
   автоматическая проверка фактов
3. Stance Detection
   выявление позиции за/против запроса (claim)
4. Controversy Detection
   выявление и кластеризация разногласий
5. Polarization Detection
   классификация позиций по многим темам
6. Clickbait Detection
   выявление противоречий заголовка и текста
7. Credibility Scores
   оценка достоверности источника или новости



E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

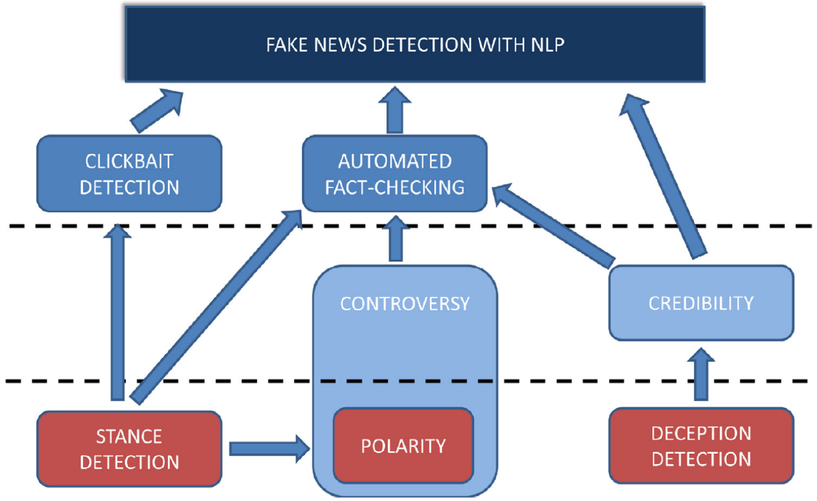
Fake News и смежные области исследований
(библиометрический анализ по данным Google Scholar)



Чего-то не хватает…

1. **Fake News** – не единственный, не главный и не самый сильный инструмент политики постправды
2. **Пропаганда** использует не только фейки, но и полуправду, замалчивание, манипулятивные воздействия и т.д.
3. **Когнитивные войны** нацелены на разрушение социокультурного кода и общественной идеологии

- Как распознавать манипулятивные воздействия и идеологические атаки?
- Как находить разногласия и замаливание?
- Насколько расширится типология задач?



E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

Типология деструктивного контента и система подзадач ML/NLP/NLU для его детекции

воздействия → фейки → пропаганда → когн.война

1.  детекция приёмов манипулирования
2.  детекция замалчивания
3.  **детекция обмана (deception detection), слухов (rumors det.), мистификаций (hoaxes det.)**
4.  **детекция кликбэйта (clickbait detection)**
5.  **автоматическая проверка фактов (auto fact-checking)**
6.  **детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)**
7.  выявление воздействий на социокультурный код и картину мира
8.  оценивание психо-эмоциональных воздействий и возможных реакций
9.  выявление целевых аудиторий воздействия
10.  **оценивание виральности (virality prediction)**
11.  **оценивание достоверности источников (credibility scores)**
12.  детекция побуждения к действиям (угрозы, призывы, провокации, вербовка, экстремизм)

Четыре основных типа подзадач ML/NLP/NLU

1. **Классификация текста (новости или предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Выделение и классификация (тегирование) фрагментов текста**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление идеологем, мифологем, ценностей социокультурного кода
   - выявление психо-эмоциональных реакций и целевых аудиторий
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление устойчивых сочетаний мнений (polarization detection)
   - выявление мнений как сочетаний слов, их семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний мнений, идеологем

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.

Задача выявления приёмов манипулирования

Структура манипуляции:

- фрагмент-мишень
- фрагмент-воздействие
- тип манипуляции

Пример из СМИ:
«**Зеленский** просто **играет роль президента, а не является президентом**[обесценивание], – считает экс-депутат Верховной рады Борислав Береза»

Типы манипуляций (всего 18 типов):

- негативизация (обесценивание, дисфемизмы, ярлыки, депрессивы и т.п.)
- позитивизация (героизация, эвфемизация, лозунги и т.п.)
- деавторизация (замалчивание источника, маскировка под ссылку и т.п.)
- паралогизация (алогизм, ложное следование, подмена тезиса и т.п.)

Классификация приёмов манипулирования

1. Негативизация

1.1 Навешивания ярлыков
1.2 Дисфемизмы
1.3 Аналогия с негативным объектом
1.4 Антифразис
1.5 Прием обесценивания
1.6 Негативирующая гиперболизация
1.7 Моделирование негативного сценария
1.8 Вкрапление депрессивов

2. Позитивизация

2.1 Эвфемизация
2.2 Лозунговые слова и словосочетания
2.3 Позитивирующая гиперболизация

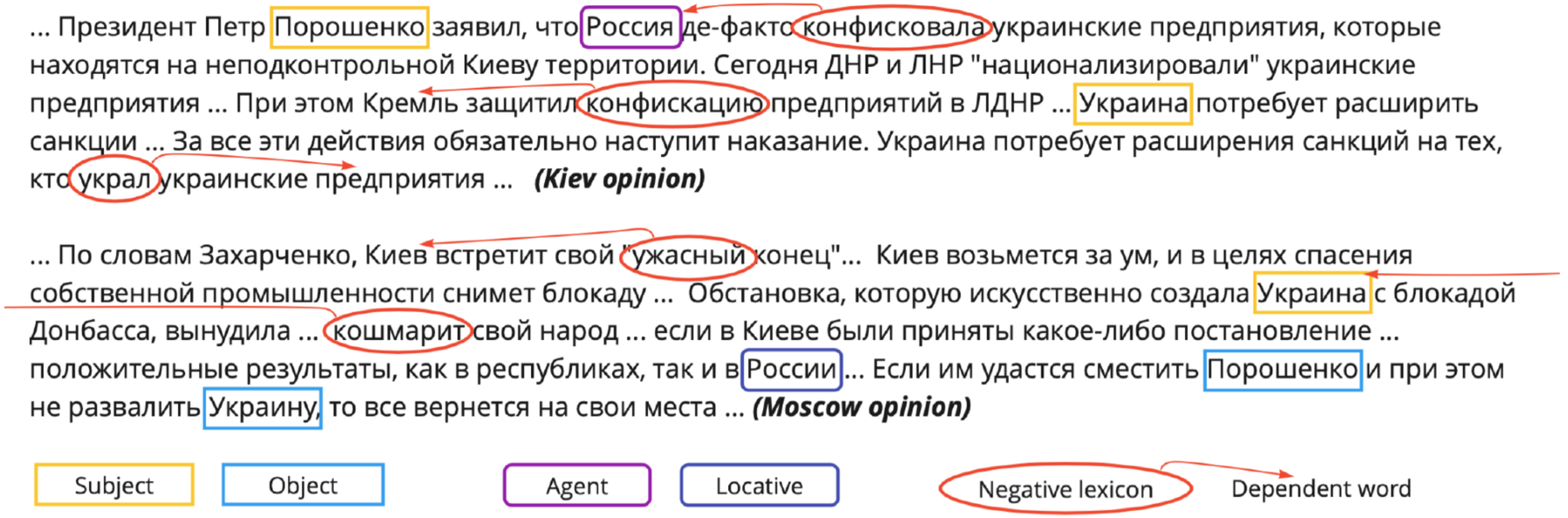
3. Деавторизация

3.1 Маскировка под ссылку на авторитет
3.2 Ссылки на неопределенный источник
3.3 Ссылки на неназванных свидетелей

4. Паралогизация

4.1 Ложная причинно-следственная связь
4.2 Прием «после этого не значит поэтому»
4.3 Подмена тезиса
4.4 Высказывание о состоянии другого

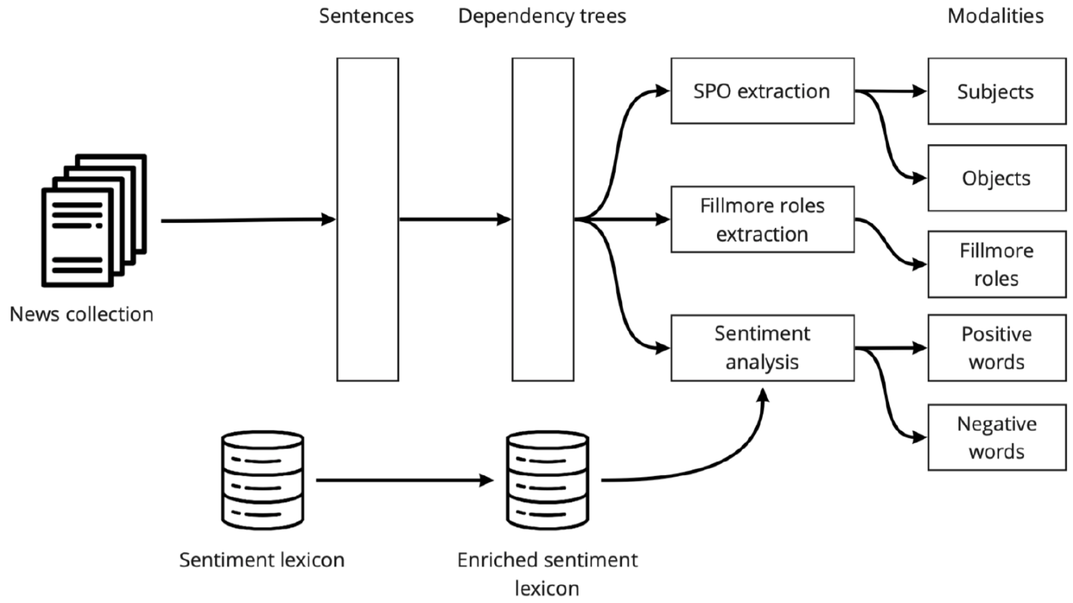
Задача выделения мнений в теме или событии



Слова «Порошенко», «Россия», «Украина» встречаются одинаково часто
«Порошенко» — субъект в первом тексте и объект во втором
«Россия» — агенс в первом тексте и локация во втором
Негативная тональность: «Россия», «Кремль» в 1-ом, «Киев», «Украина» во 2-ом

Feldman D. G., Sadekova T. R., Vorontsov K. V. Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining. Computational Linguistics and Intellectual Technologies. Dialogue 2020.

Задача выделения мнений в теме или событии



Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

LPR Business

Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	**0.77**	**0.94**	**0.85**

Paris Trump

Мнение формализуется как устойчивое сочетание слов, терминов, именованных сущностей, их семантических ролей по Филлмору и их тональных окрасок.
Все они используются в тематической модели как отдельные модальности.

Feldman D. G., Sadekova T. R., Vorontsov K. V. Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining. Computational Linguistics and Intellectual Technologies. Dialogue 2020.

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

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

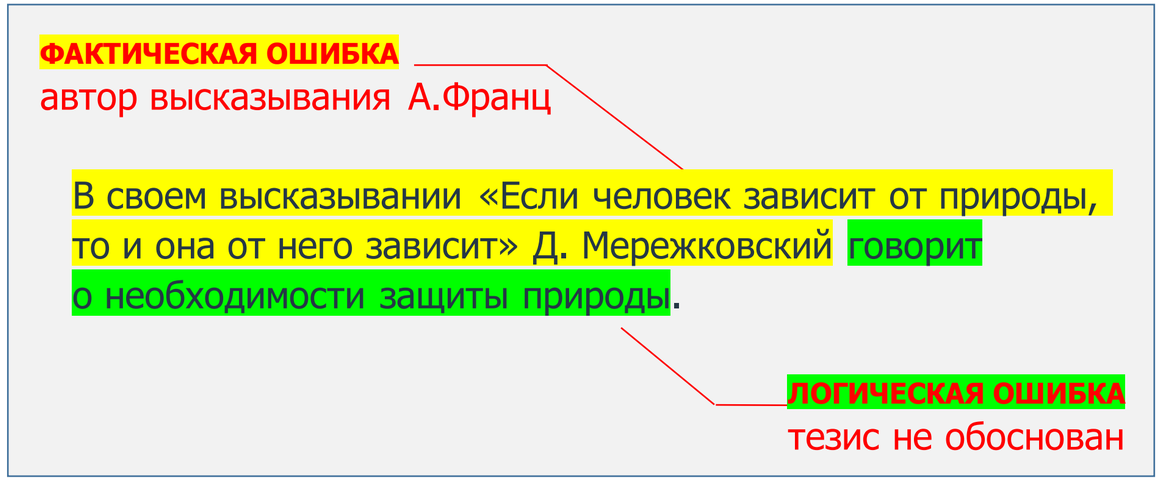
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

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

Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

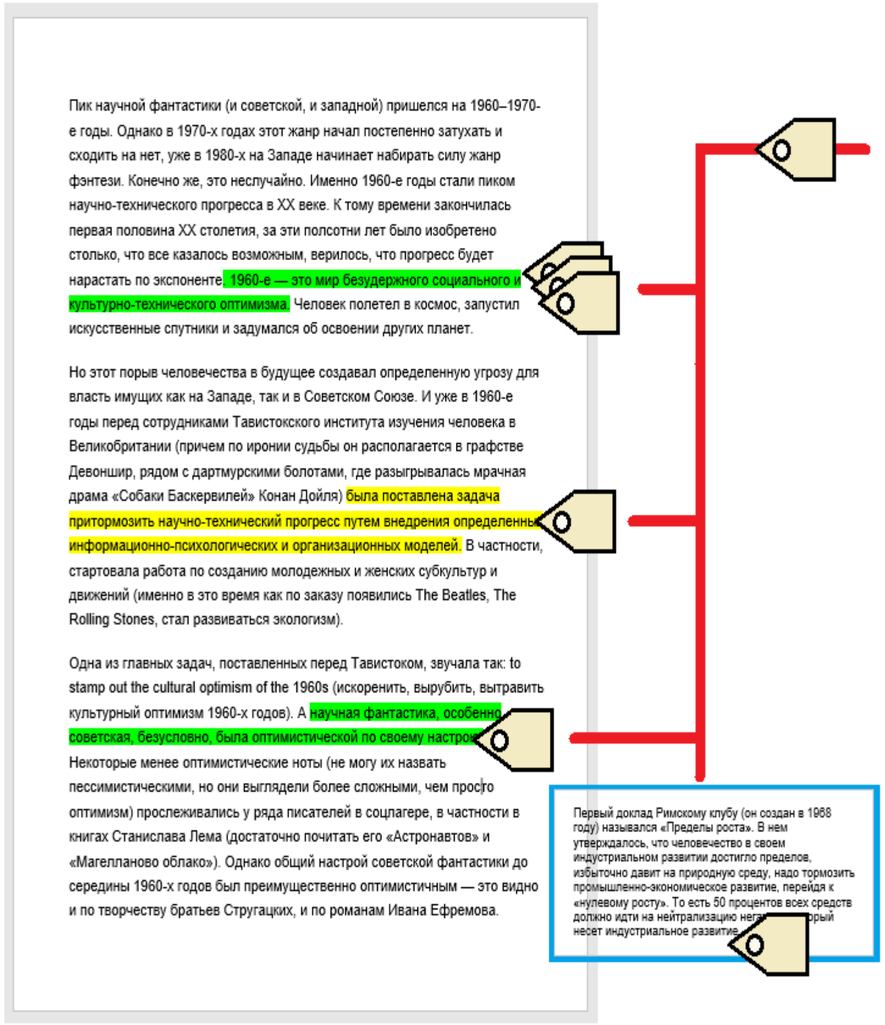
1) выделить фрагмент
2) классифицировать (тегировать) фрагмент по рубрикатору
3) связать несколько фрагментов
4) дать комментарий (затекст) к фрагменту или связи

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

Три задачи построения обучаемой модели разметки:

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

Разметка текста: обобщённый контент-анализ

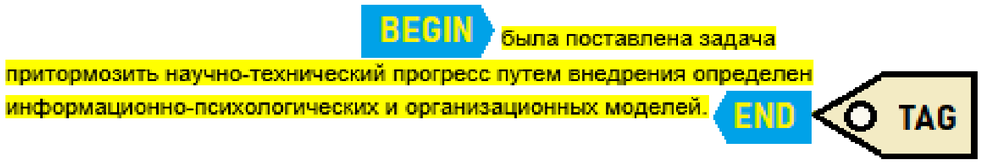


Разметка состоит из элементов

Элемент разметки — несколько взаимосвязанных фрагментов, затекстов и тегов

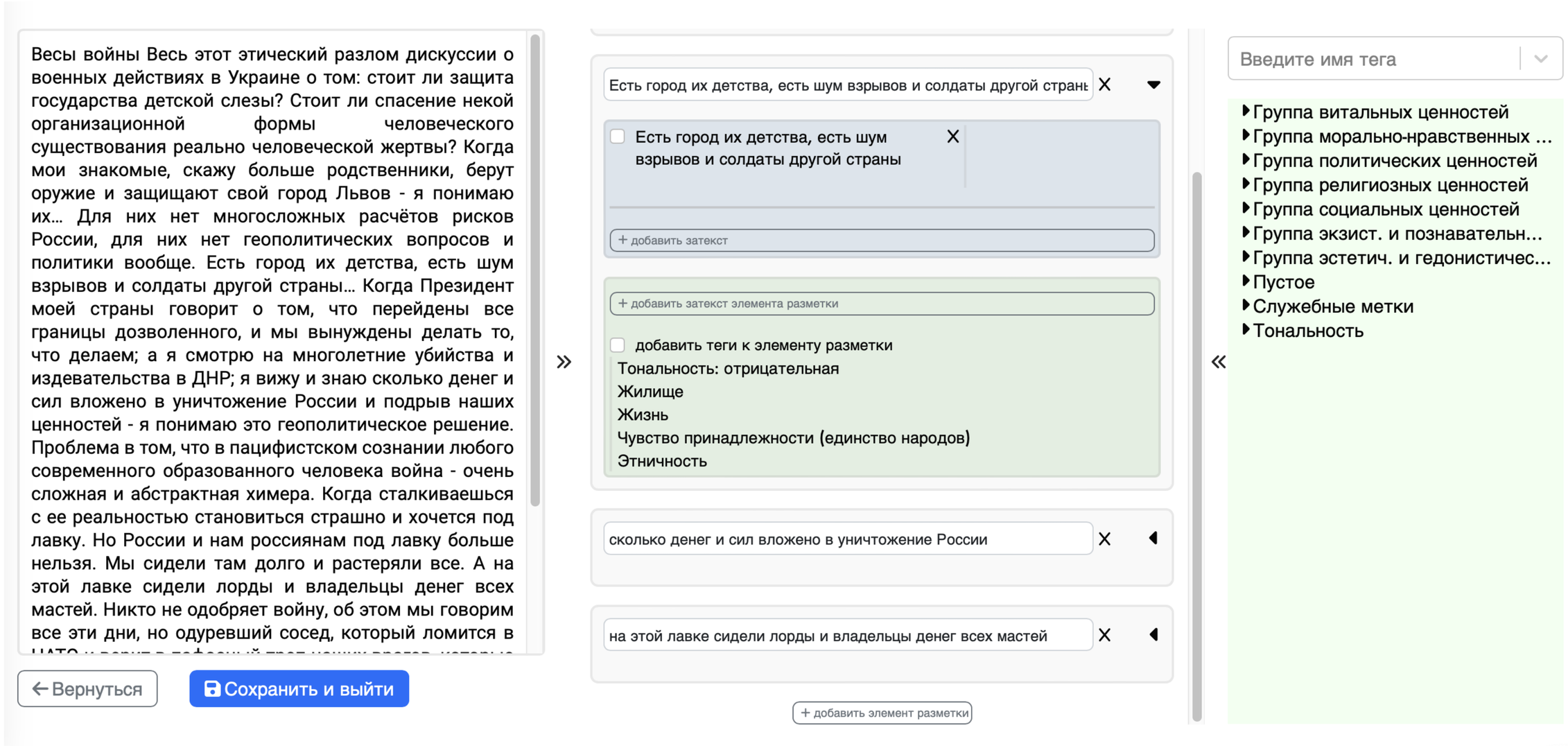
Теги (классы) выбираются из рубрикатора

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

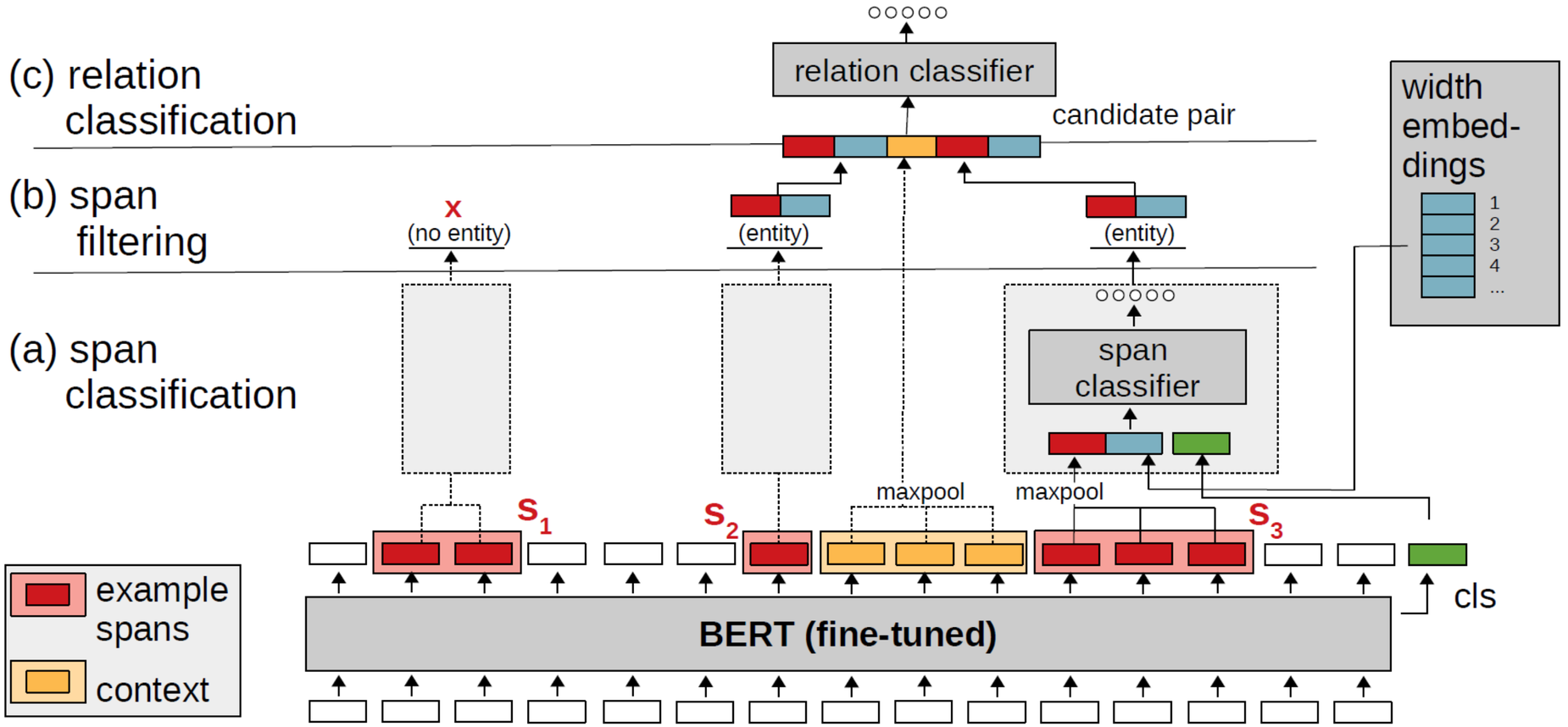


Затекст — комментарий, объяснение, дополнительная информация и т.п., может иметь один или несколько тегов

Инструмент разметки



Нейросетевые обучаемые модели разметки

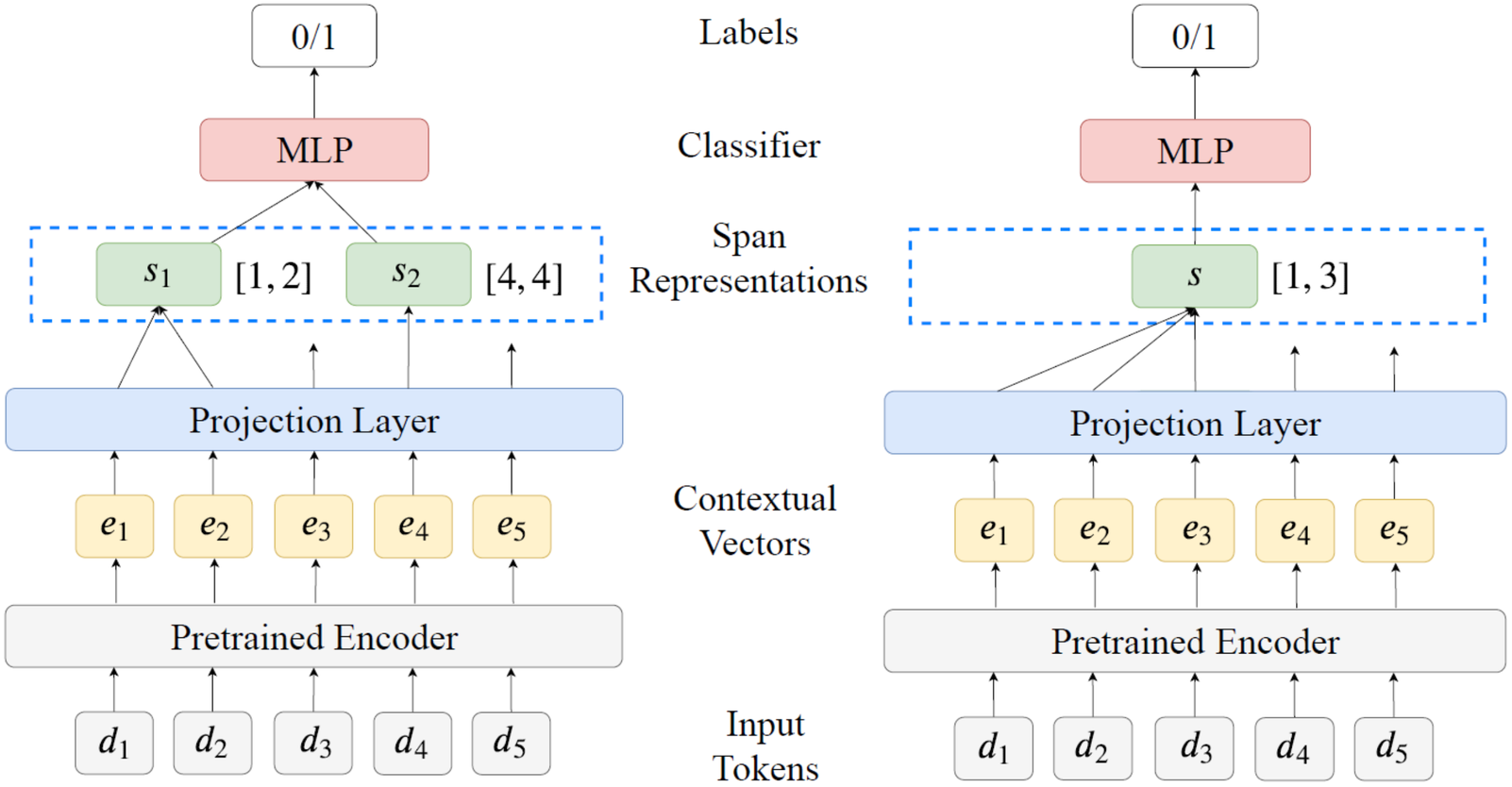


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

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.

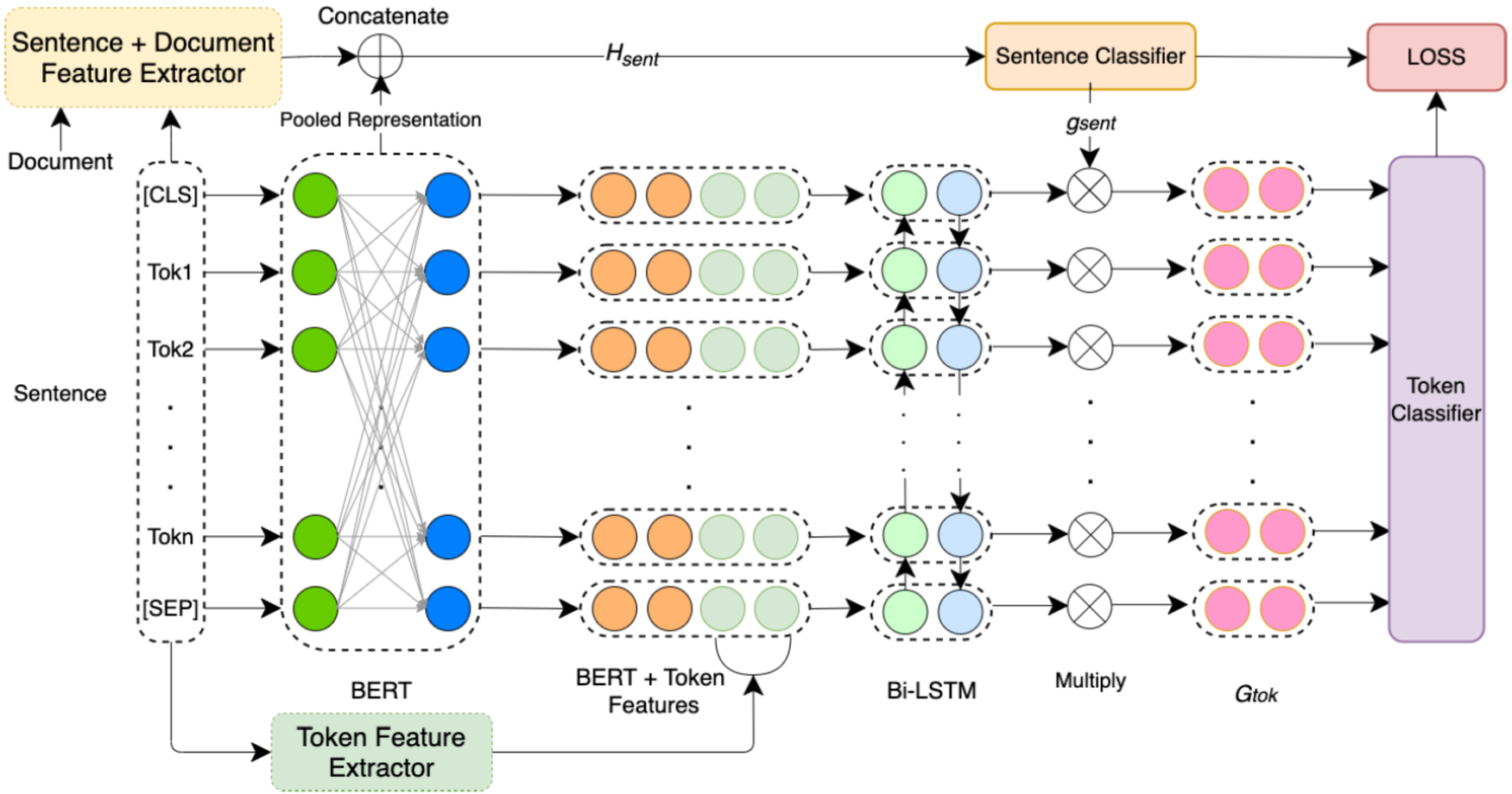
Нейросетевые обучаемые модели разметки



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



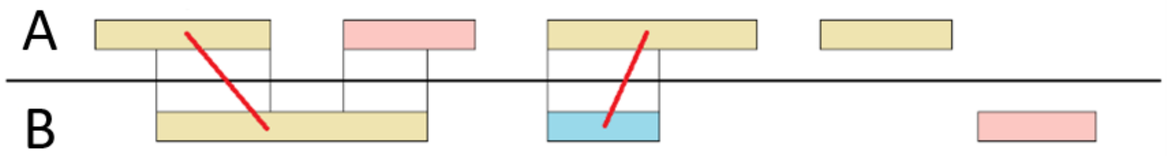
Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

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

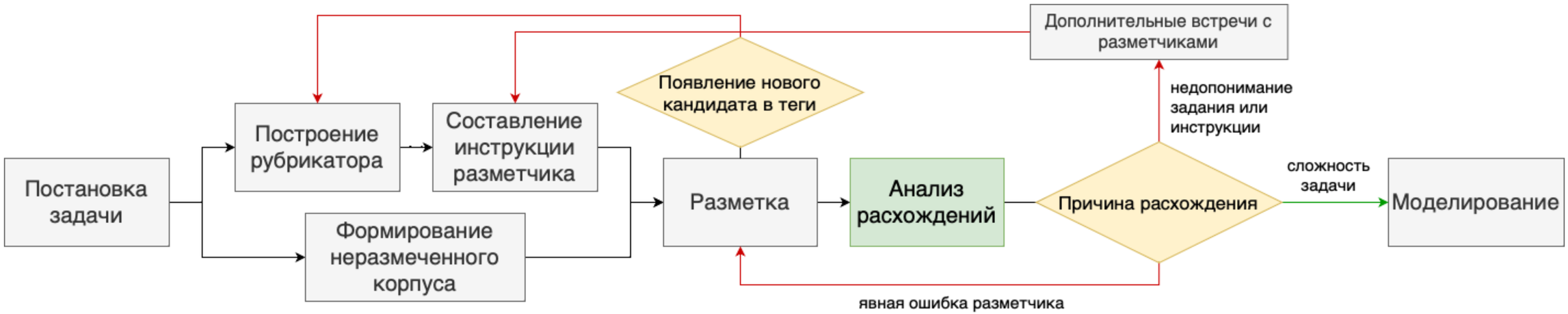
Con2 = точность наложения сопоставленных фрагментов

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 (2 или 3)
- документы ранжируются по согласованности экспертов Con(E,E’)
- наибольшие расхождения обсуждаются, вырабатывается консенсус
- происходит доработка инструкции и/или переразметка документов

Детекция ценностей социокультурного кода

Создание рубрикатора на основе кросс-дисциплинарного подхода

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

Указ президента Российской Федерации № 400 от 2-07-2021 «О стратегии нацбезопасности»
Указ президента Российской Федерации № 809 от 9-11-2022 «Об утверждении основ госполитики по сохранению и укреплению традиционных российских духовно-нравственных ценностей»

Какие ценности брать для рубрикатора: аксиоматический подход

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

Рубрикатор ценностей (1 из 2)

М. Рокич | Ш. Шварц | Г. Хофстеде

Ю.С. Степанов | Р. Инглхарт | Б.С. Ерасов

Группа социальных ценностей

- Социальные ценности
- Авторитет
- Альтруизм
- Благородство происхождения / аристократизм
- Важность общественного мнения
- Воспитание
- Гендерное разнообразие
- Дети
- Долгосрочная ориентация
- Дружба
- Избегание неопределённости
- Индивидуализм
- Интеллигентность
- Коллективизм
- Культура (нормы) поведения
- Личные границы
- Материальные ценности
- Патриархальность
- Патриотизм
- Пацифизм / мир во всём мире
- Полезность (созидательный труд)
- Профессиональный успех
- Репутация
- Семья
- Социальное признание
- Суеверия
- Трудолюбие / продуктивность
- Чувство принадлежности / единство народов
- Этничность
- Язык

Группа витальных ценностей

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

Группа политических ценностей

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

Рубрикатор ценностей (2 из 2)

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

Группа религиозных ценностей

- Религиозные ценности
- Благочестивость
- Бог
- Религиозность
- Эзотерика

Группа эстетических и гедонистических ценностей

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

Группа экзистенциальных и познавательных ценностей

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

Группа морально-нравственных ценностей

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

Ценностный ландшафт

Самые частотные теги (по количеству элементов)

40%

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

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

Выводы

- Большие языковые модели позволяют сегодня решать те задачи, которые ещё 5 лет назад считались непреодолимо трудными
- В том числе автоматизировать контент-анализ больших текстовых корпусов для масштабирования социогуманитарных исследований
- **Ближайшая задача:** составление мультизадачного бенчмарка для тестирования универсальности обучаемых моделей разметки

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН,
зав. лабораторией МОСА Института ИИ МГУ,
зав. кафедрой ММП ВМК МГУ,
voron@mlsa-iai.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov