Большие языковые модели для автоматизации разметки текстов

Воронцов Константин Вячеславович
voron@mlsa-iai.ru
зав. лаб. Машинного интеллекта и семантического анализа Института ИИ МГУ; проф., зав. каф. ММП ВМК МГУ; проф., зав. каф. МОЦГ МФТИ; г.н.с. ФИЦ ИУ РАН

Форум «Открытые данные» • Панельная сессия
«Информационное воздействие. ИИ для анализа культурных кодов»
Томск • 10–11 ноября 2023

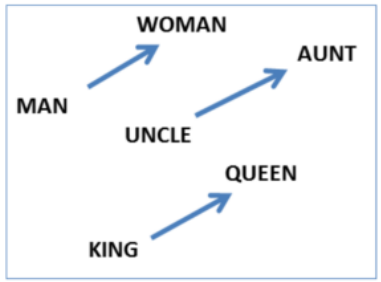
Содержание

Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

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

Модели векторных представлений (эмбедингов) слов на основе матричных разложений



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

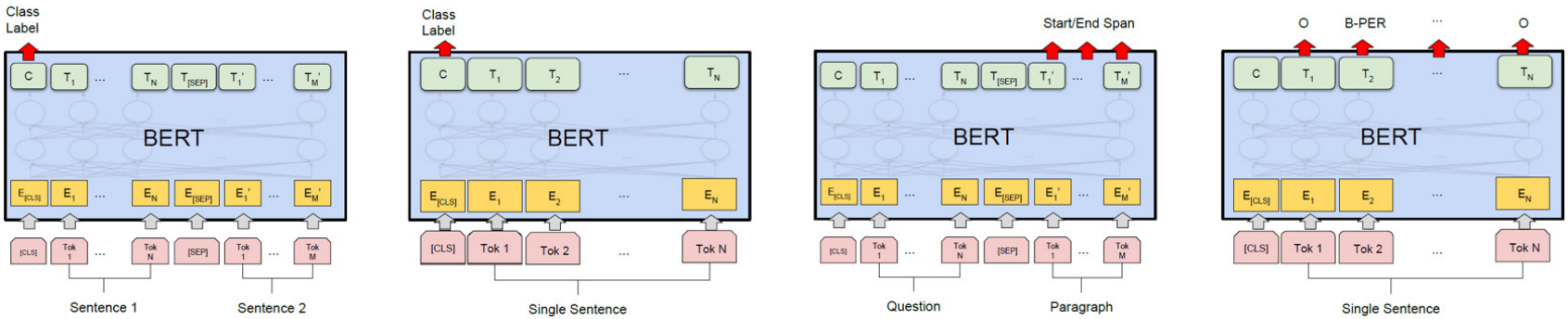
Нейросетевые модели локальных контекстов

- рекуррентные нейронные сети
- модели внимания и трансформеры:
  BERT [2018], GPT-3 [2020], GPT-4 [2023]

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

Большие пред-обученные модели языка (трансформеры)

- обучены предсказывать слово по его контексту
- обучены по терабайтам текстов, «они видели в языке всё»
- способны выделять и классифицировать фрагменты текста, генерировать фейковые тексты, не отличимые от реальных
- *мультиязычны*: обучаются на десятках языков
- *мультизадачны*: для каждой новой задачи NLP/NLU достаточно дообучения на малой размеченной выборке



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

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

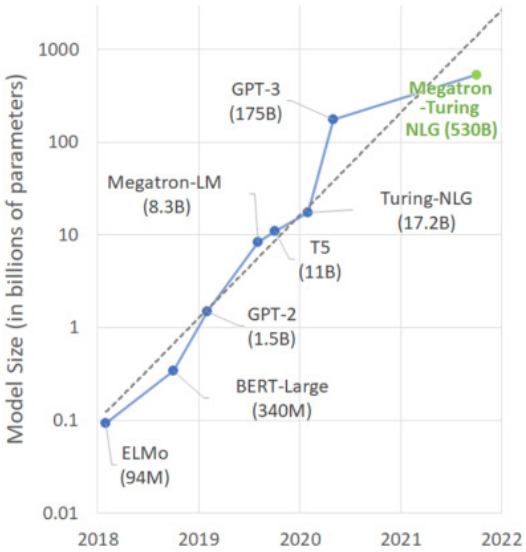
(c) Question Answering Tasks: SQuAD v1.1

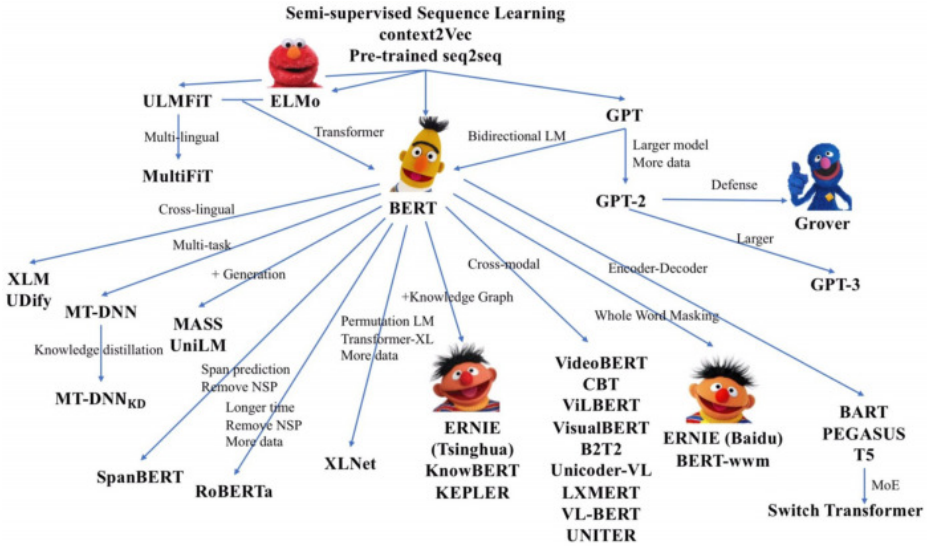
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

J.Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019.

Рост больших языковых моделей — быстрее закона Мура

Рост числа параметров в трансформерных моделях языка





- *трансформер-кодировщик* преобразует последовательность слов в числовые векторы, зависящие от контекста
- *трансформер-декодировщик* преобразует векторную последовательность в последовательность слов

Число параметров сети сопоставимо с объёмом исходных данных

ChatGPT и GPT-4: проблески общего искусственного интеллекта

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

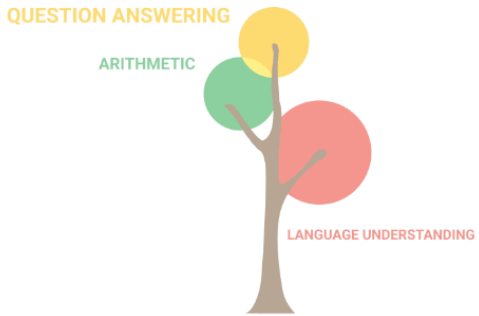
Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Microsoft Research *(27 March 2023)*

Новые способности модели, не закладывавшиеся при обучении:

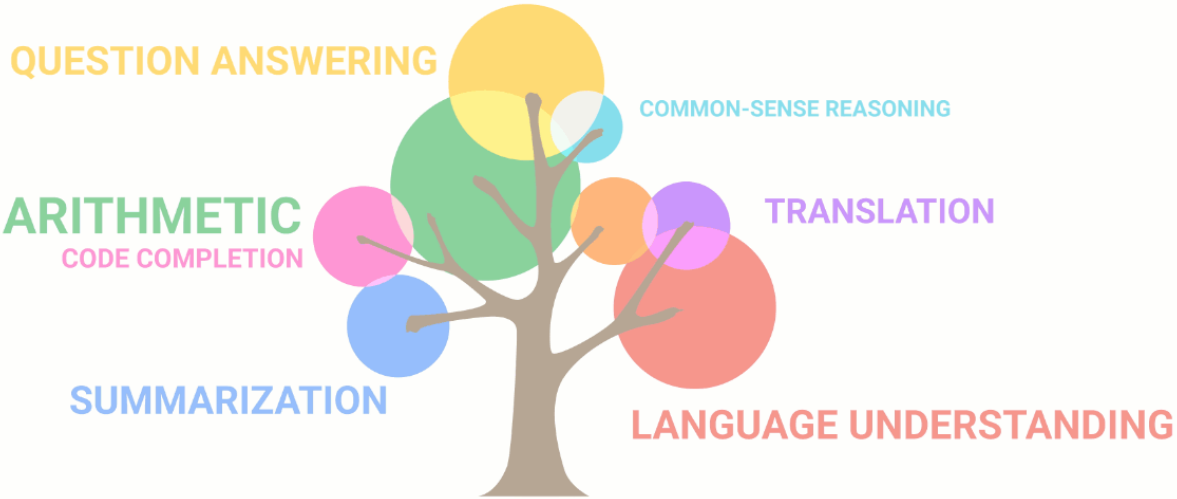
- объяснять свои ответы, перефразировать
- реферировать, генерировать планы, сценарии, шаблоны
- переводить на другие языки, строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать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

Эмерджентность — появление качественно новых способностей



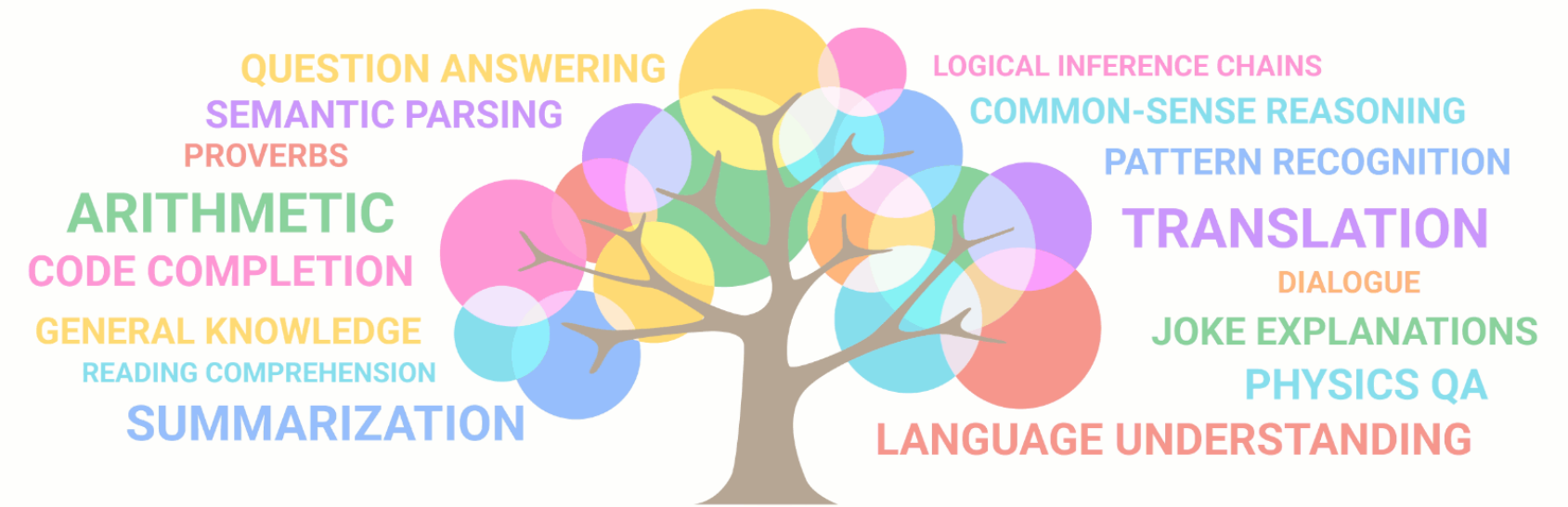
- GPT-2: 14/Feb/2019, контекст 768 слов (1,5 страницы)
- 1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb)
- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Эмерджентность — появление качественно новых способностей



- GPT-3: 11/Jun/2020, контекст 1536 слов (3 страницы)
- 175 млрд. параметров, корпус 500 млрд. токенов
- способность делать перевод на другие языки,
- решать логические и математические задачи,
- генерировать программный код по описанию

Эмерджентность — появление качественно новых способностей

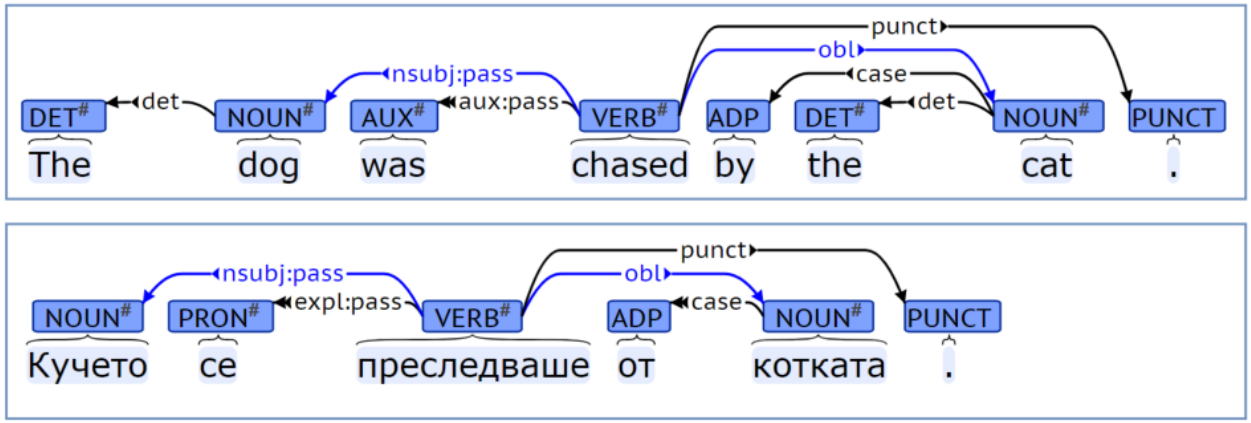


- GPT-4: 14/Mar/2023, контекст 24 000 слов (48 страниц)
- >1 трл. параметров, корпус >1Tb
- способность описывать и анализировать изображения,
- реагировать на подсказки вроде «Let's think step by step»,
- решать качественные физические задачи по картинке

Примеры задач разметки и сегментации

- распознавание частей речи (part of speech tagging, POS)
- неглубокий синтаксический разбор (shallow syntax parsing)
- распознавание именованных сущностей (named entity, NER)
- выделение семантических ролей (semantic role labeling)
- анализ тональности заданной сущности (sentiment analysis)
- выделение текстовых полей данных (slot filling)
- выделение полей в библиографических записях
- сегментация научных или юридических текстов
- поиск кореференций и разрешение анафор
- поиск и разрешение эллипсиса (гэппинга)
- перевод речевого сигнала в текст
- перевод музыкального сигнала в нотную запись
- выделение генов в нуклеотидных последовательностях

Пример частеречной и синтаксической разметки



Теги частей речи (не все, и могут зависеть от языка):

NOUN	noun	существительное	INTJ	interjection	междометие
PROPN	proper noun	имя собственное	ADP	adposition	предлог
ADJ	adjective	прилагательное	CONJ	conjunction	союз
VERB	verb	глагол	PART	particle	частица
ADV	adverb	наречие	PUNCT	punctuation	знак пунктуации
PRON	pronoun	местоимение	SYM	symbol	символ
NUM	numeral	числительное	X	other	иное

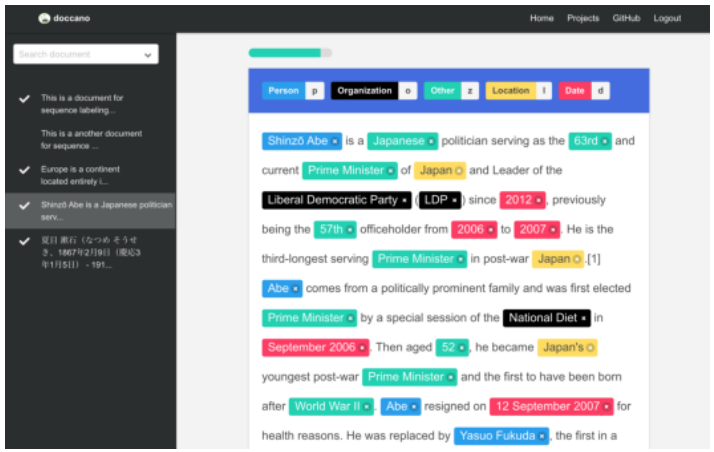
http://universaldependencies.org/

Разметка именованных сущностей (Named Entity Recognition)

Named entity — объект (сущность) реального мира, имеющий *наименование* и относящийся к определённой *категории*.

Примеры категорий:

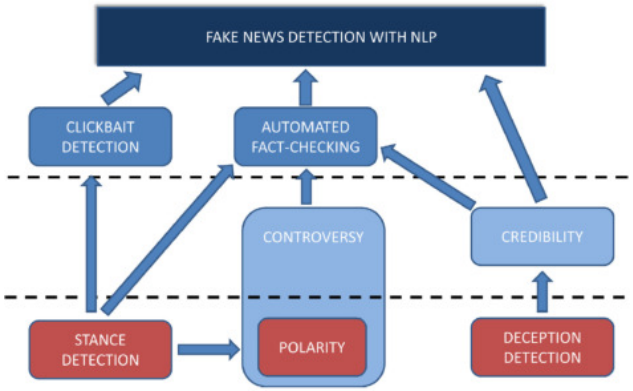
- персона, организация, локация
- профессия, должность, звание
- болезнь, симптом, препарат
- химическое вещество
- биологический вид
- астрономический объект
- артикул, изделие, стандарт
- ссылка на нормативно-правовой акт



Stanford Named Entity Recognizer:
`http://www-nlp.stanford.edu/software/CRF-NER.shtml`

Область исследований «Fake News Detection»

1. **Deception Detection**
   выявление обмана в тексте
2. **Automated Fact-Checking**
   автоматическая проверка фактов
3. **Stance Detection**
   выявление позиции за или против
4. **Controversy Detection**
   выявление и кластеризация разногласий
5. **Polarization Detection**
   выявление полярных позиций
6. **Clickbait Detection**
   противоречия заголовка и текста
7. **Credibility Scores**
   оценка достоверности источников



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

Задачи Propaganda/Manipulation/Persuasion Detection

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.

Типология угроз и задачи их автоматической детекции

воздействия → фейки → пропаганда → инф.война

1. детекция приёмов манипулирования
2. детекция замалчивания
3. детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
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 конструктов картины мира: идеологем, мифологем
8. оценивание возможных психо-эмоциональных реакций
9. выявление целевых аудиторий воздействия
10. оценивание и предсказание скорости распространения (virality prediction)
11. оценивание достоверности источников (credibility scores)
12. детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

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

Типы задач ML/NLU для мониторинга медиа-пространства

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

ПРО//ЧТЕНИЕ — технологический конкурс Up Great

Задача: поиск смысловых ошибок в сочинениях ЕГЭ по русскому, литературе, истории, обществознанию, английскому

Период: декабрь 2019 — декабрь 2022

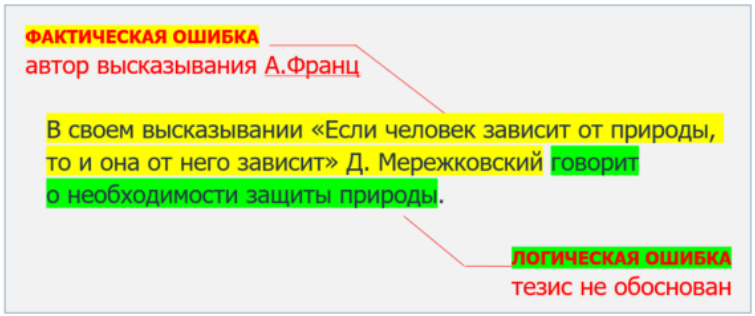
Призовой фонд:
— 100M руб. русский язык
— 100M руб. английский язык

Up Great technology contest

READ // ABLE
comprehension technology

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)

Алгоритм должен выделять ошибки и давать их объяснения.



Официальный сайт конкурса: http://ai.upgreat.one

Сравнение двух разметок (алгоритма и эксперта)

1. насколько точно предсказана оценка за сочинение
2. насколько точно предсказаны фрагменты ошибок и блоков
3. насколько точно совпадают границы фрагментов
4. совпадают ли типы и подтипы ошибок
5. насколько содержательны сгенерированные пояснения

Разметка как способ формализации гуманитарных знаний

Цель — автоматизация обработки текстовых источников (контент-анализа и др.) в социогуманитарных исследованиях.

Гипотеза: достаточно четырёх базовых операций разметки:

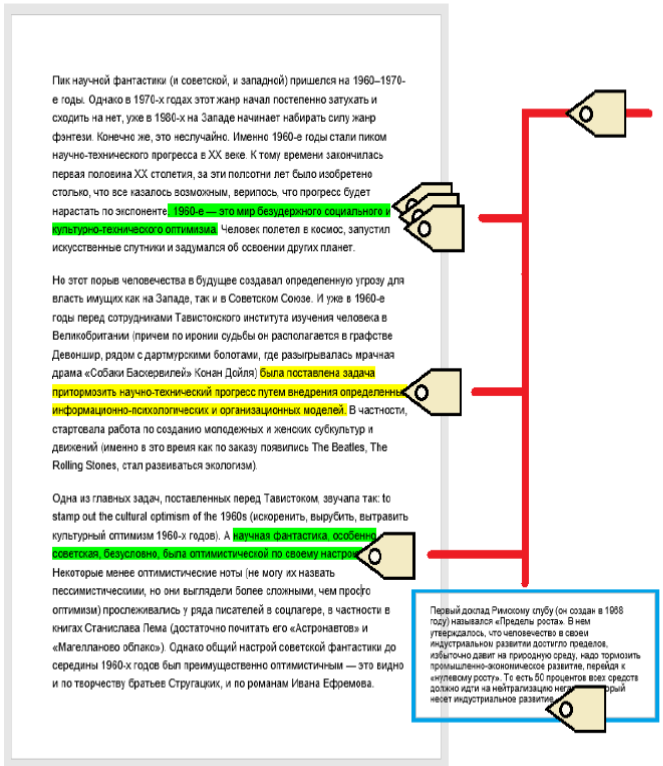
1. выделить фрагмент
2. классифицировать (тегировать) фрагмент по рубрикатору
3. связать несколько фрагментов
4. дать комментарий (затекст) к фрагменту или связи

Задачи универсализации обучаемой модели разметки:

1. унификация правил разметки и инструментария разметки
2. унификация нейросетевой архитектуры модели разметки
3. унификация методики оценивания моделей разметки

Унификация правил разметки и инструментария разметки

Обобщение классических задач компьютерной лингвистики (NER, SentAn, SemRL, SyntPars), задач выявления манипуляций, поляризации, смысловых ошибок в академических эссе и др.



Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Теги (классы) выбираются из словаря тегов

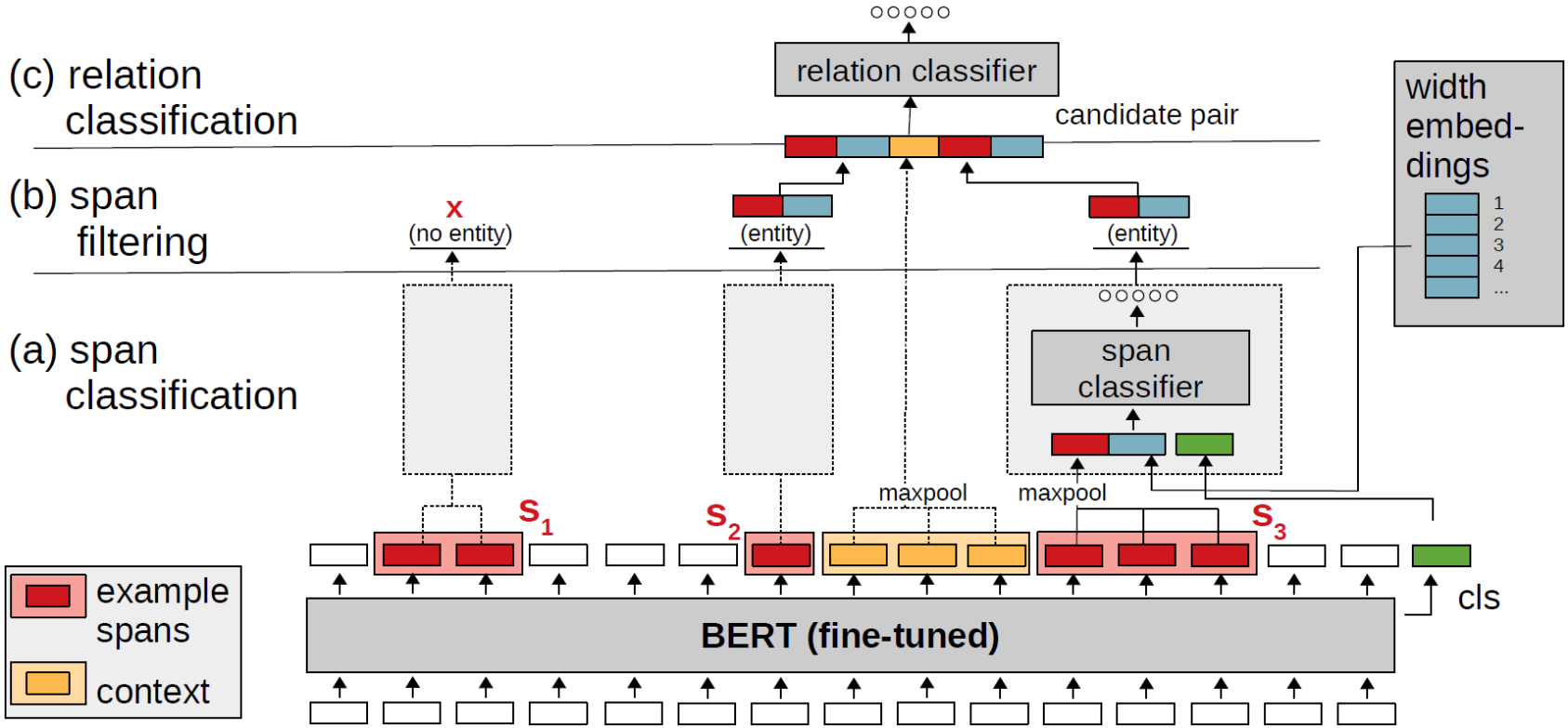
Фрагмент задаётся началом и концом, может иметь один или несколько тегов:

BEGIN была поставлена задача притормозить научно-технический прогресс путем внедрения определен информационно-психологических и организационных моделей. END TAG

Затекст может выбираться из словаря фраз или свободно генерироваться по контексту, может иметь один или несколько тегов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Унификация нейросетевых архитектур моделей разметки

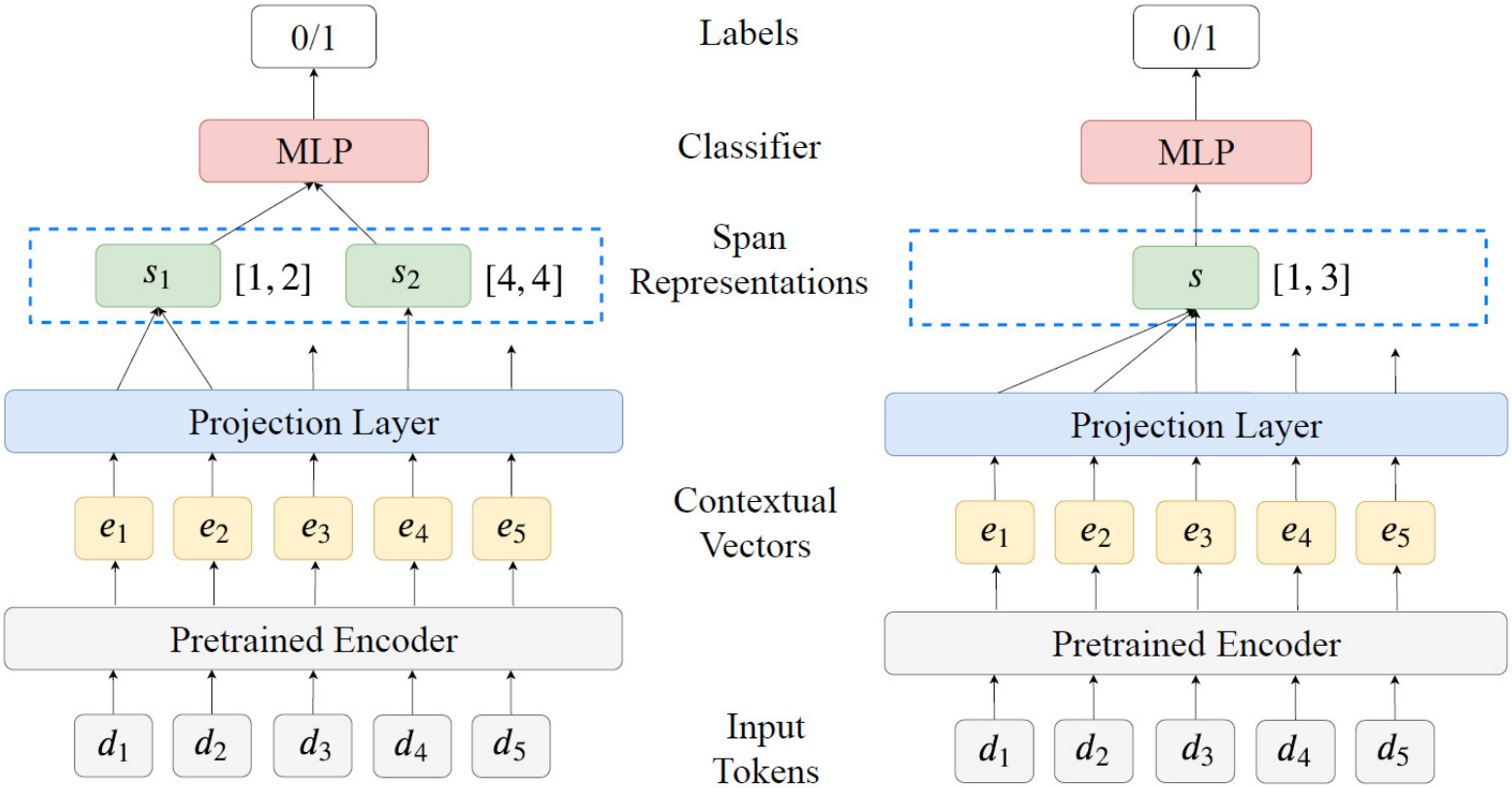


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

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.

Сравнение методов формирования эмбедингов фрагментов



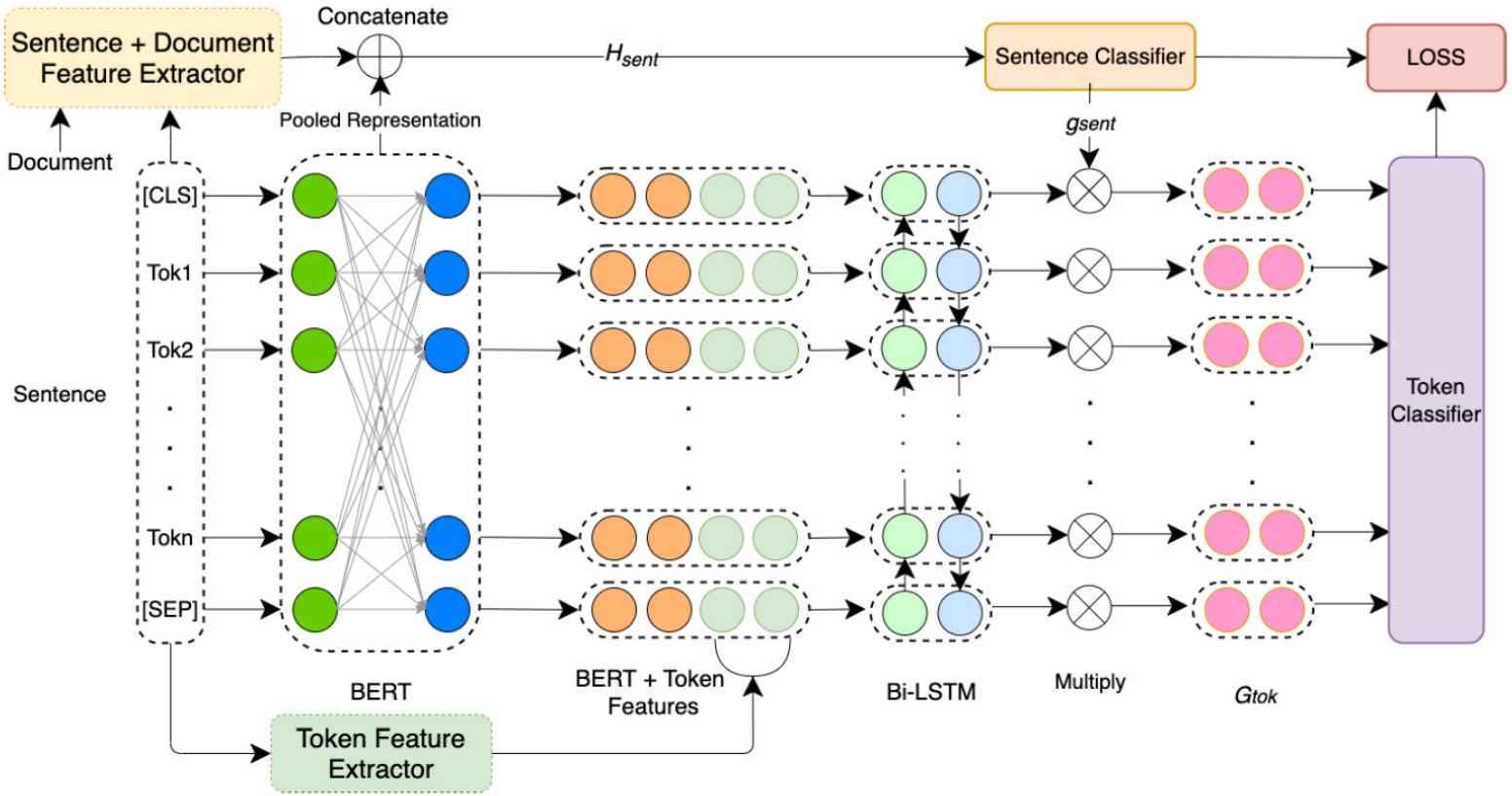
Полнота (recall) до 90% на задачах NER, SRL, Mention Detection

---

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.

Извлечение признаков предложений и документов

Задача детекции фрагментов с приёмами пропаганды



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

Унификация методики оценивания моделей разметки

- В основе методики — сравнение пар разметок текста: «алгоритм – эксперт», «эксперт-1 – эксперт-2», путём оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $\mathrm{Con}_k(A, B)$
- Вводится их средневзвешенная согласованность $\mathrm{Con}(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(A, E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность $\mathrm{Con}(E_1, E_2)$ разметок двух экспертов, E_1 и E_2
- ОТАР = СТАР / СТЭР (Относительная Точность Алгоритмической Разметки) — если выше 100%, то это означает, что алгоритм работает не хуже экспертов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Выводы

- Нынешний бум искусственного интеллекта обязан развитию методов обучаемой (по большим данным) векторизации сложно структурированных объектов.
- В анализе текстов это большие языковые модели, размер которых сопоставим с размером обучающих данных.
- Эти модели позволяют сегодня решать те задачи, которые ещё 5 лет назад считались непреодолимо трудными.
- В том числе задачи понимания текста для автоматизации и масштабирования социогуманитарных исследований.
- Причём «гибридный интеллект» не заменяет специалиста, а уменьшает объём рутинной работы и ускоряет её.

Воронцов Константин Вячеславович • voron@mlsa-iai.ru