

Унификация фрагментной разметки текста и оценивания моделей разметки: искусственный интеллект на пути к формализации гуманитарных знаний

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН • профессор ВМК МГУ,
руководитель лаборатории машинного обучения
и семантического анализа Института ИИ МГУ,
г.н.с. ФИЦ ИУ РАН, профессор МФТИ



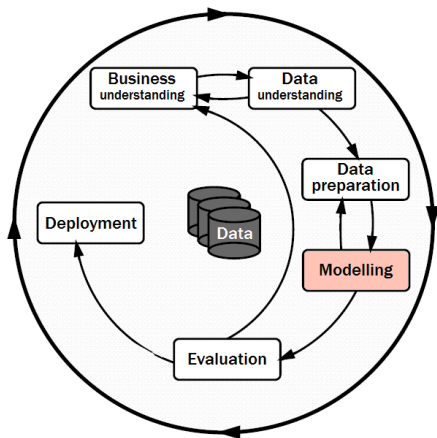
ВСЕМИРНЫЙ КОНГРЕСС
**ТЕОРИЯ СИСТЕМ,
АЛГЕБРАИЧЕСКАЯ БИОЛОГИЯ,
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ**
Математические основы и приложения

30 июня 2023

- 1 Задачи понимания естественного языка**
 - эволюция подходов в обработке текстов
 - большие предобученные языковые модели
 - чем GPT отличается от всего, что было раньше
- 2 Задачи разметки текстов**
 - задачи компьютерной лингвистики
 - задачи анализа новостных потоков
 - технологический конкурс ПРО//ЧТЕНИЕ
- 3 Унификация моделей разметки**
 - унификация структуры разметки
 - унификация моделей разметки
 - унификация оценивания качества разметки

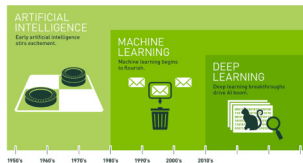
Эволюция подходов ИИ как автоматизации шагов CRISP-DM

CRISP-DM: Cross Industry Standard Process for Data Mining (1999)



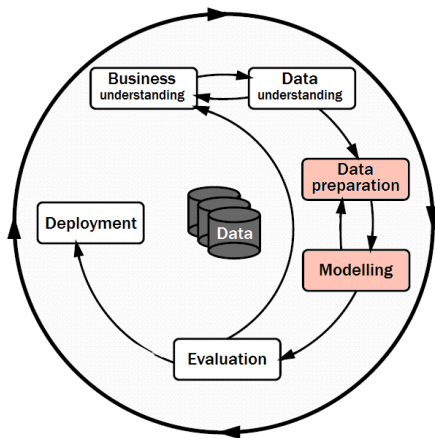
Эволюция ИИ:

- *Expert Systems*: жёсткие модели, основанные на правилах
- *Machine Learning*: параметрические модели, обучаемые по данным



Эволюция подходов ИИ как автоматизации шагов CRISP-DM

CRISP-DM: CRoss Industry Standard
Process for Data Mining (1999)

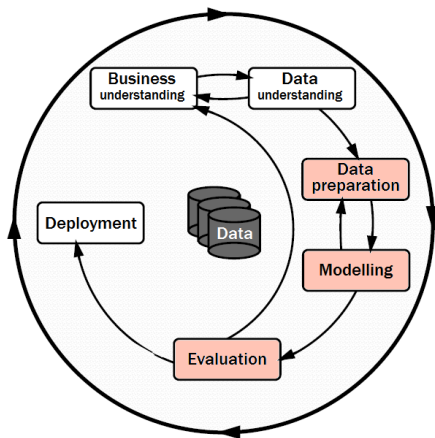


Эволюция ИИ:

- *Expert Systems*: жёсткие модели, основанные на правилах
- *Machine Learning*: параметрические модели, обучаемые по данным
- *Deep Learning*: модели с обучаемой векторизацией данных

Эволюция подходов ИИ как автоматизации шагов CRISP-DM

CRISP-DM: Cross Industry Standard Process for Data Mining (1999)

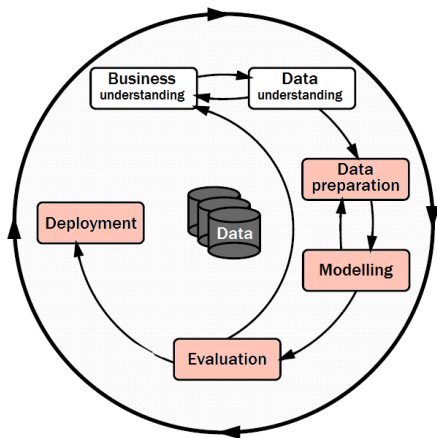


Эволюция ИИ:

- *Expert Systems*: жёсткие модели, основанные на правилах
- *Machine Learning*: параметрические модели, обучаемые по данным
- *Deep Learning*: модели с обучаемой векторизацией данных
- *AutoML*: автоматический выбор моделей и архитектур

Эволюция подходов ИИ как автоматизации шагов CRISP-DM

CRISP-DM: Cross Industry Standard Process for Data Mining (1999)



Эволюция ИИ:

- *Expert Systems:*
жёсткие модели,
основанные на правилах
- *Machine Learning:*
параметрические модели,
обучаемые по данным
- *Deep Learning:*
модели с обучаемой
векторизацией данных
- *AutoML:*
автоматический выбор
моделей и архитектур
- *Lifelong Learning:*
бесшовная интеграция
обучения и выбора
моделей в бизнес-процесс

Эволюция подходов машинного обучения в анализе текстов

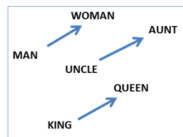
Декомпозиция задач по уровням пирамиды NLP

- 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]

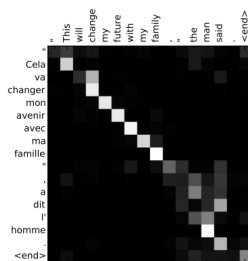
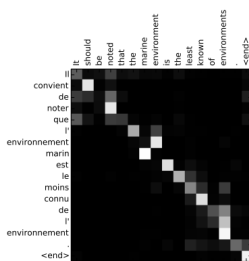
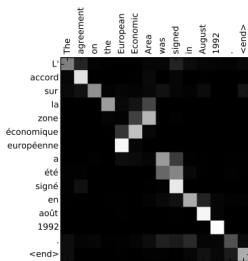


Нейросетевые модели локальных контекстов

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

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

Модели внимания для машинного перевода



Вход: $\{x_i\}$ — последовательность слов входного языка

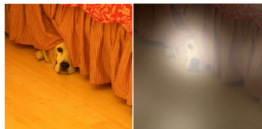
Выход: $\{y_t\}$ — последовательность слов выходного языка

Интерпретация модели внимания: матрица семантического сходства A_{ti} показывает, на какие слова x_i входного текста модель обращает внимание, генерируя слово перевода y_t

Модели внимания для аннотирования изображений



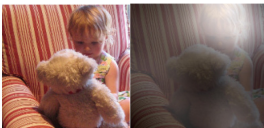
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



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



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



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



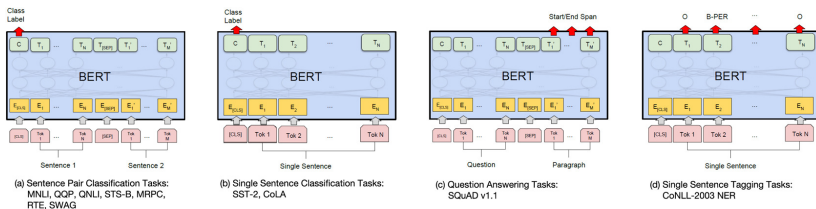
A giraffe standing in a forest with trees in the background.

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

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

Большие пред-обученные модели языка (трансформеры)

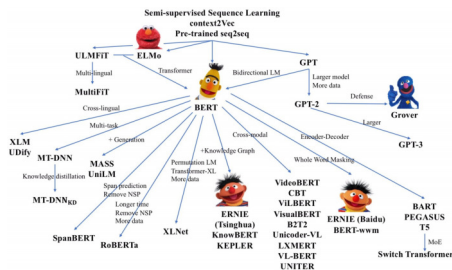
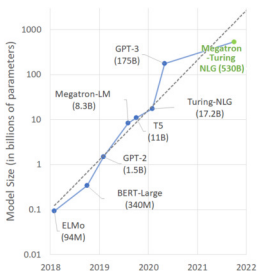
- обучены предсказывать слово по его контексту
- обучены по терабайтам текстов, «они видели в языке всё»
- способны выделять и классифицировать фрагменты текста, генерировать фейковые тексты, не отличимые от реальных
- *мультиязычны*: обучаются на десятках языков
- *мультизадачны*: для каждой новой задачи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



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

Большие пред-обученные модели языка

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

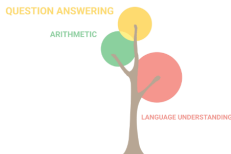
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research (27 March 2023)

Новые способности модели, не закладывавшиеся при обучении:

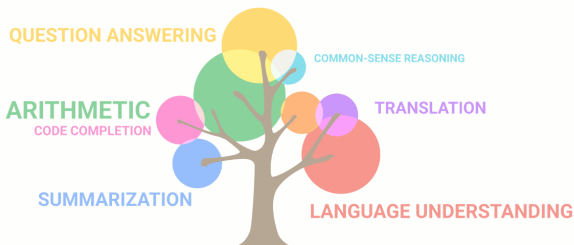
- объяснять свои ответы, перефразировать
- реферировать, генерировать планы, сценарии, шаблоны
- переводить на другие языки, строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать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

Появление у модели качественно новых способностей



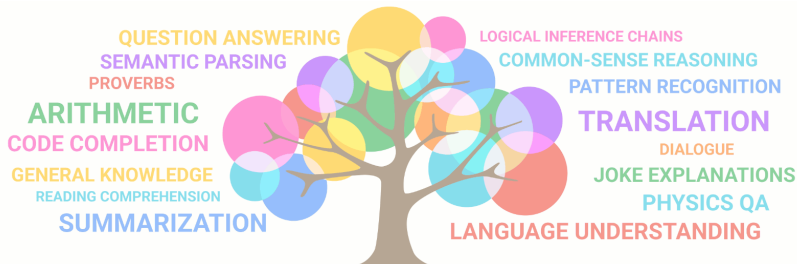
- GPT-2: 14/Feb/2019, контекст 768 слов (1,5 страницы)
- 1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb)
- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Появление у модели 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

Появление у модели качественно новых способностей

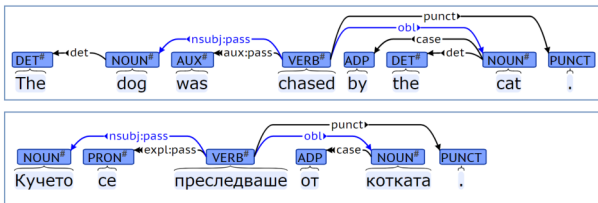


- GPT-4: 14/Mar/2023, контекст 24 000 слов (48 страниц)
- >1 трл. параметров, корпус >1Tb
- способность описывать и анализировать изображения,
- реагировать на подсказки вроде «Let's think step by step»,
- решать качественные физические задачи по картинке

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):

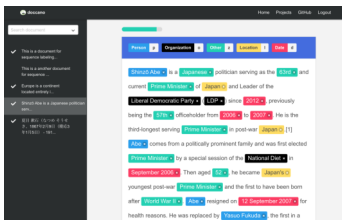
NOUN	noun	существительное	INTJ	interjection	междометие
PROPN	proper noun	имя собственное	ADP	adposition	предлог
ADJ	adjective	прилагательное	CONJ	conjunction	союз
VERB	verb	глагол	PART	particle	частица
ADV	adverb	наречие	PUNCT	punctuation	знак пунктуации
PRON	pronoun	местоимение	SYM	symbol	символ
NUM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

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

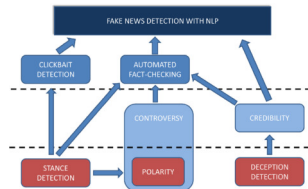


Stanford Named Entity Recognizer:

<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 Aquitania 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 comeant 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 proelis 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 конструкторов 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:

— 100М руб. русский язык

— 100М руб. 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)

Алгоритм должен выделять ошибки и давать их объяснения.



ФАКТИЧЕСКАЯ ОШИБКА
автор высказывания А.Франц

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

ЛОГИЧЕСКАЯ ОШИБКА
тезис не обоснован

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

Сравнение двух разметок (алгоритма и эксперта)

Алгоритмическая разметка

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

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

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

поступки
поступки
поступки
поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

Экспертная разметка 2

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

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

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

поступки
поступки
поступки
поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

поступки

- 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:

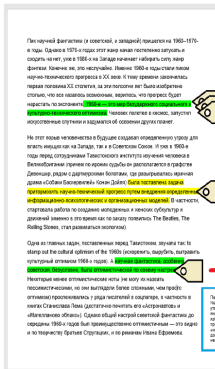
- 1 выделить фрагмент
- 2 классифицировать (тегировать) фрагмент по рубрикатору
- 3 связать несколько фрагментов
- 4 написать комментарий (затекст) к фрагменту или связи

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.

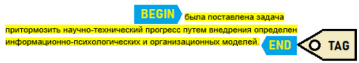


Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

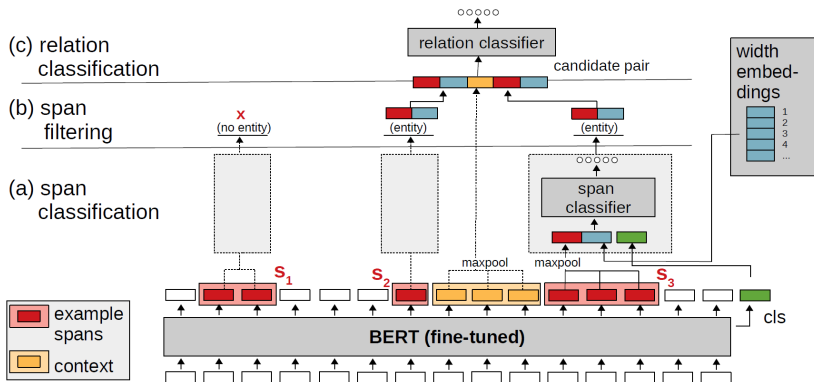
Теги (классы) выбираются из словаря тегов

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



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

Унификация нейросетевой архитектуры 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.

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 $\text{Con}_k(A, B)$
- Вводится их средневзвешенная согласованность $\text{Con}(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность $\text{Con}(A, E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность $\text{Con}(E_1, E_2)$ 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

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