

The International Congress Humanities vs Sciences &
the Knowledge Accelerating in Modern World: Parallels and Interaction

**Стандартизация разметки текста и
оценивания предсказательных моделей
в задачах понимания естественного языка**

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. кафедрой математических методов прогнозирования МГУ им. М.В. Ломоносова,

зав. лабораторией машинного обучения и семантического анализа ИИИ МГУ им. М.В. Ломоносова,

зав. кафедрой машинного обучения и цифровой гуманитаристики МФТИ, Москва, Россия

voron@mlsa-iai.ru

Эволюция подходов в обработке текстов

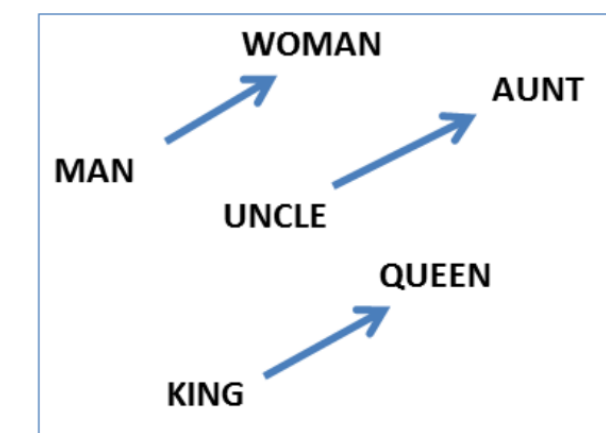
Декомпозиция задач по уровням «пирамиды NLP»

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



Модели векторизации слов (эмбедингов)

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



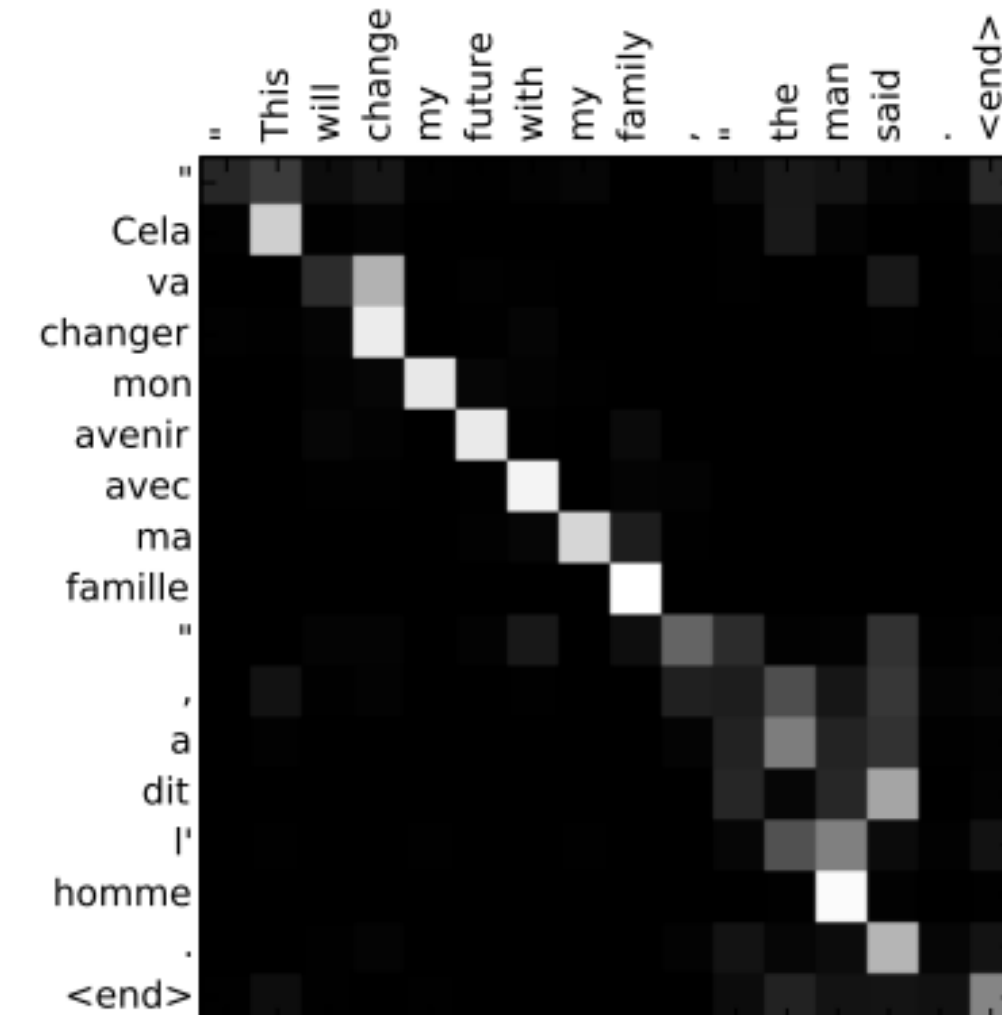
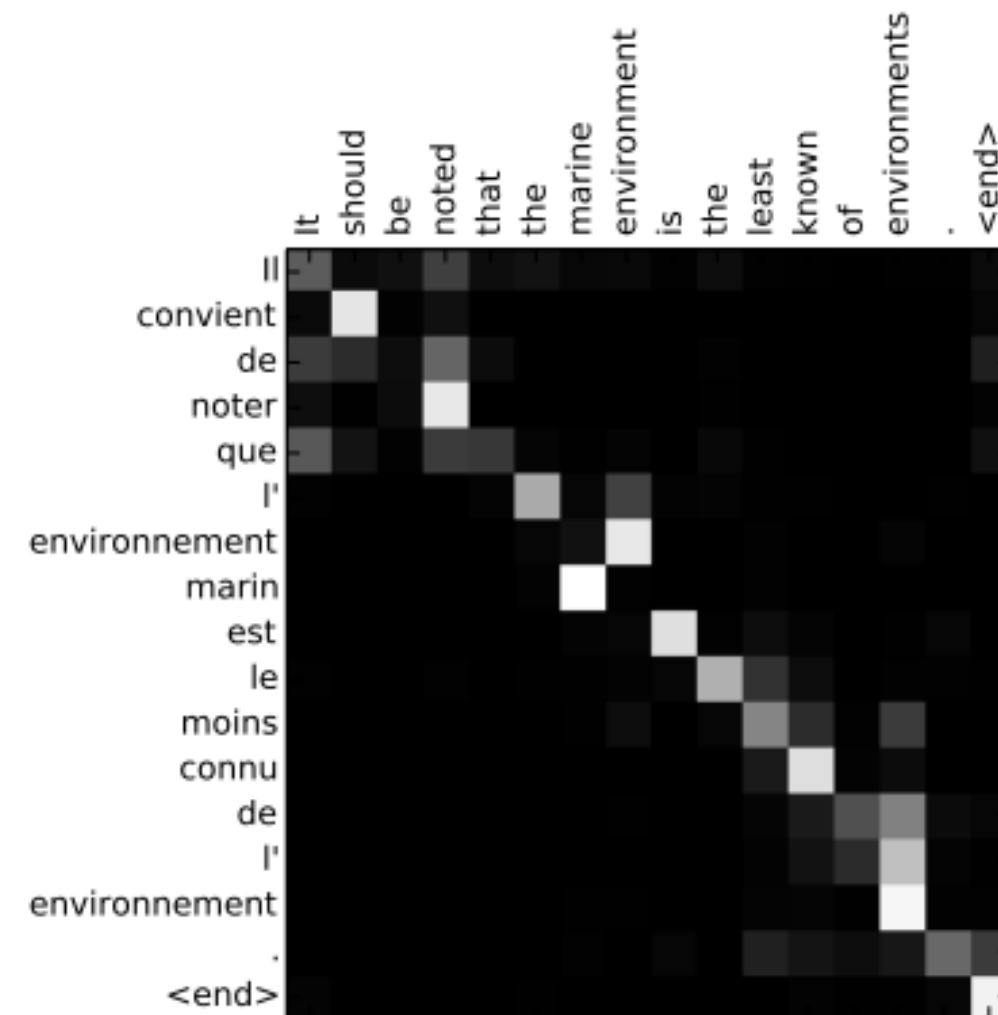
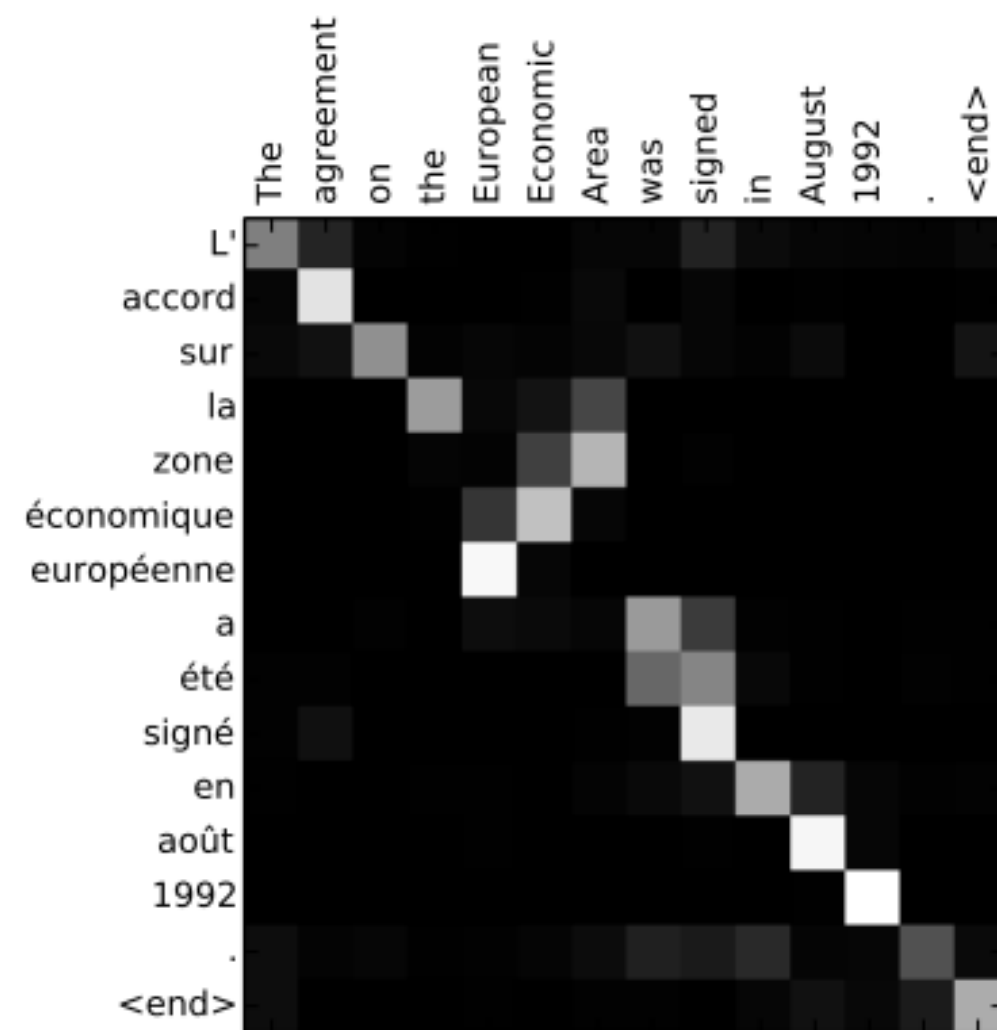
Нейросетевые модели контекстной векторизации

- рекуррентные нейронные сети: LSTM, GRU, ...
- «end-to-end» модели внимания и трансформеры: машинный перевод [2017], BERT [2018], GPT-3 [2020], ...

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

The diagram shows a matrix multiplication of a query matrix \mathbf{Q} (purple) and a key matrix \mathbf{K}^T (orange), followed by a softmax function and a value matrix \mathbf{V} (blue). The result is a 2x2 matrix.

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 $y[t]$

Модели внимания: аннотирование 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.

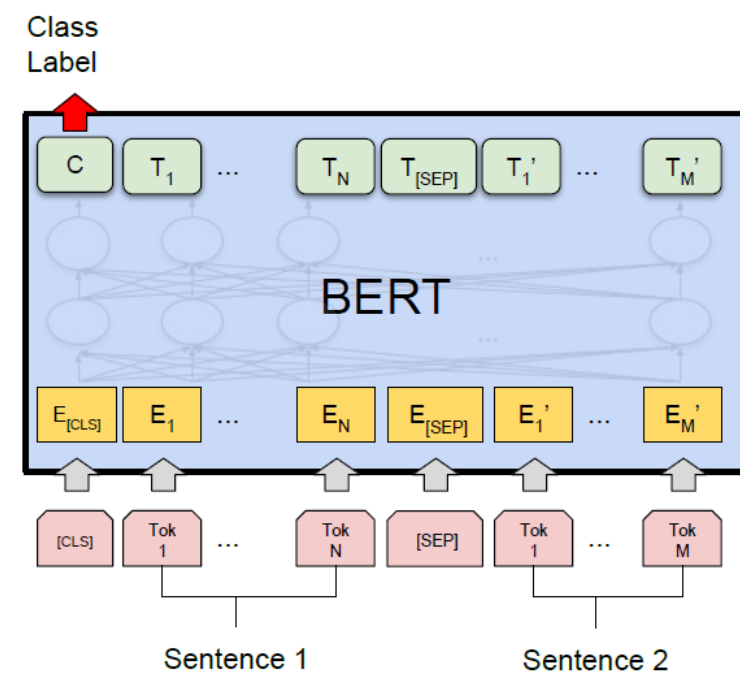


A giraffe standing in a forest with trees in the background.

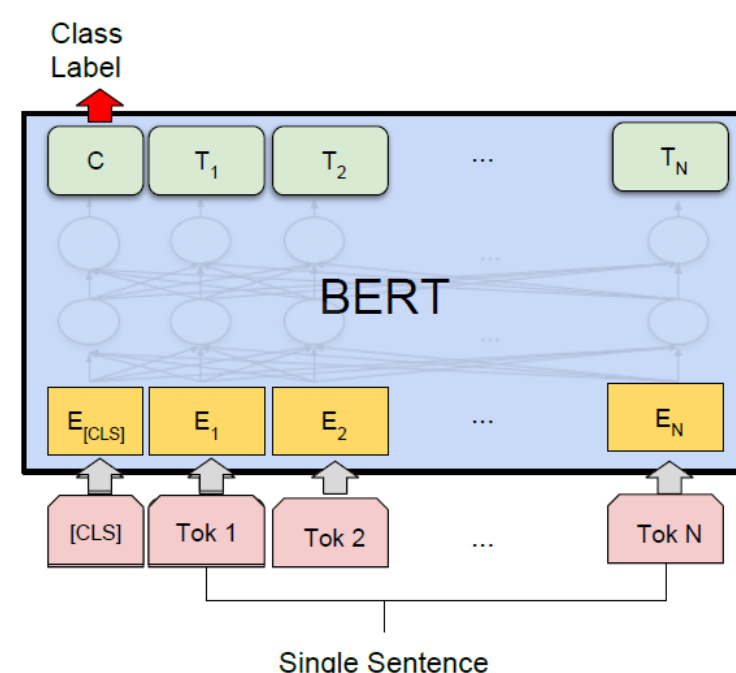
Интерпретация: на какие области модель обращает внимание, генерируя подчёркнутое слово в описании изображения

Нейросетевые модели языка

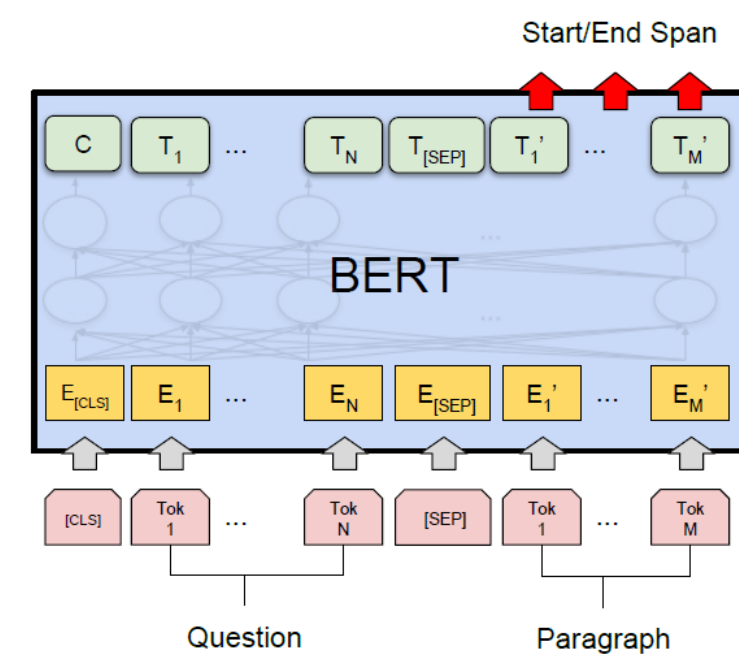
- Обучаются по терабайтам текстов, «они видели в языке всё»
- Способны генерировать фейковый текст, не отличимый от реального
- Мультиязычны: обучаются на десятках языков
- Мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



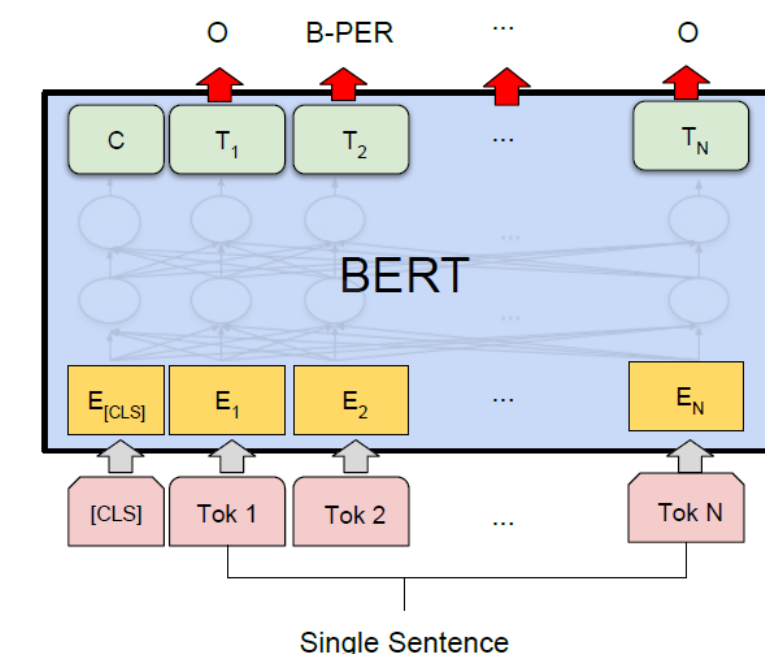
(a) Sentence Pair Classification Tasks:
MNL, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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



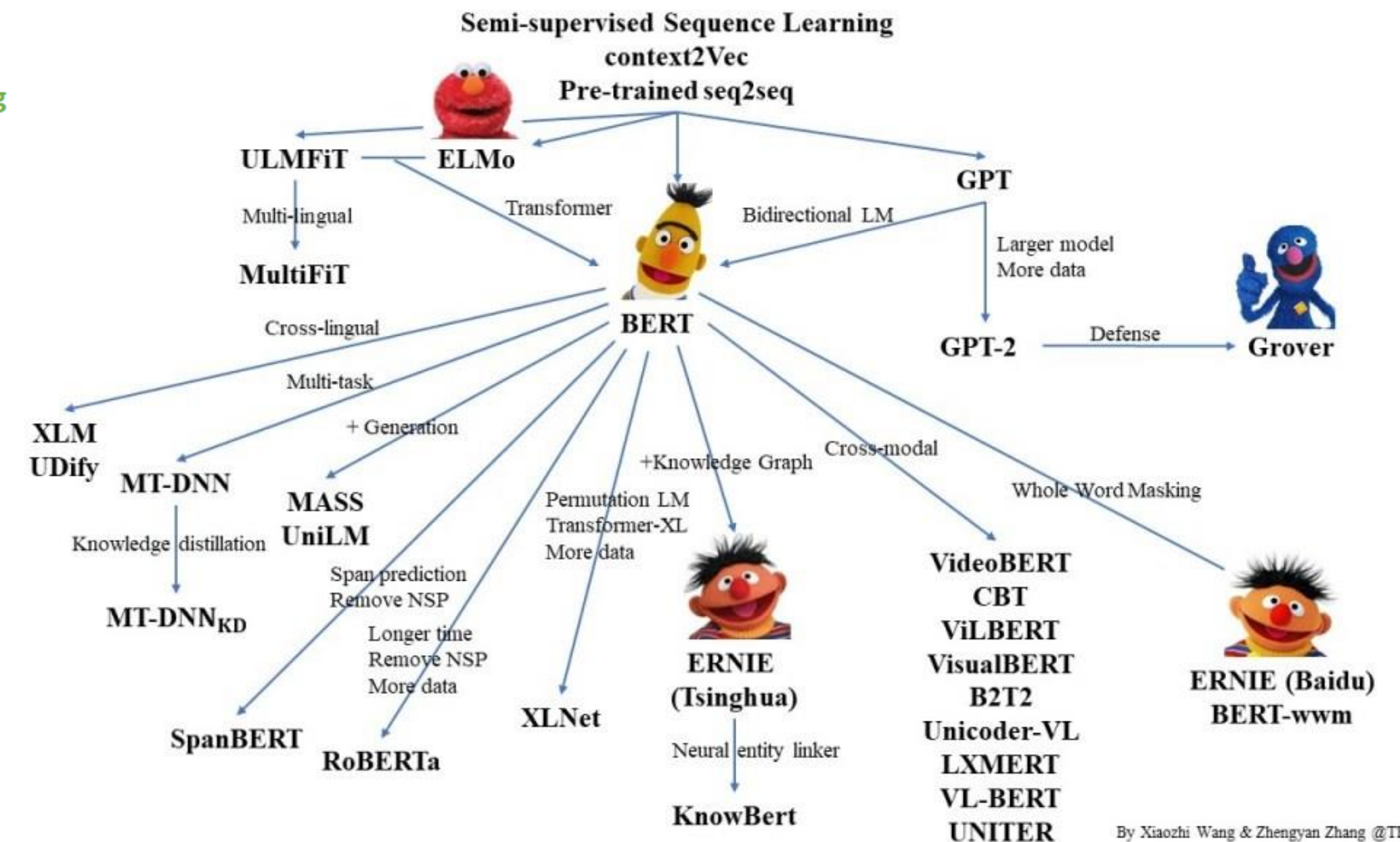
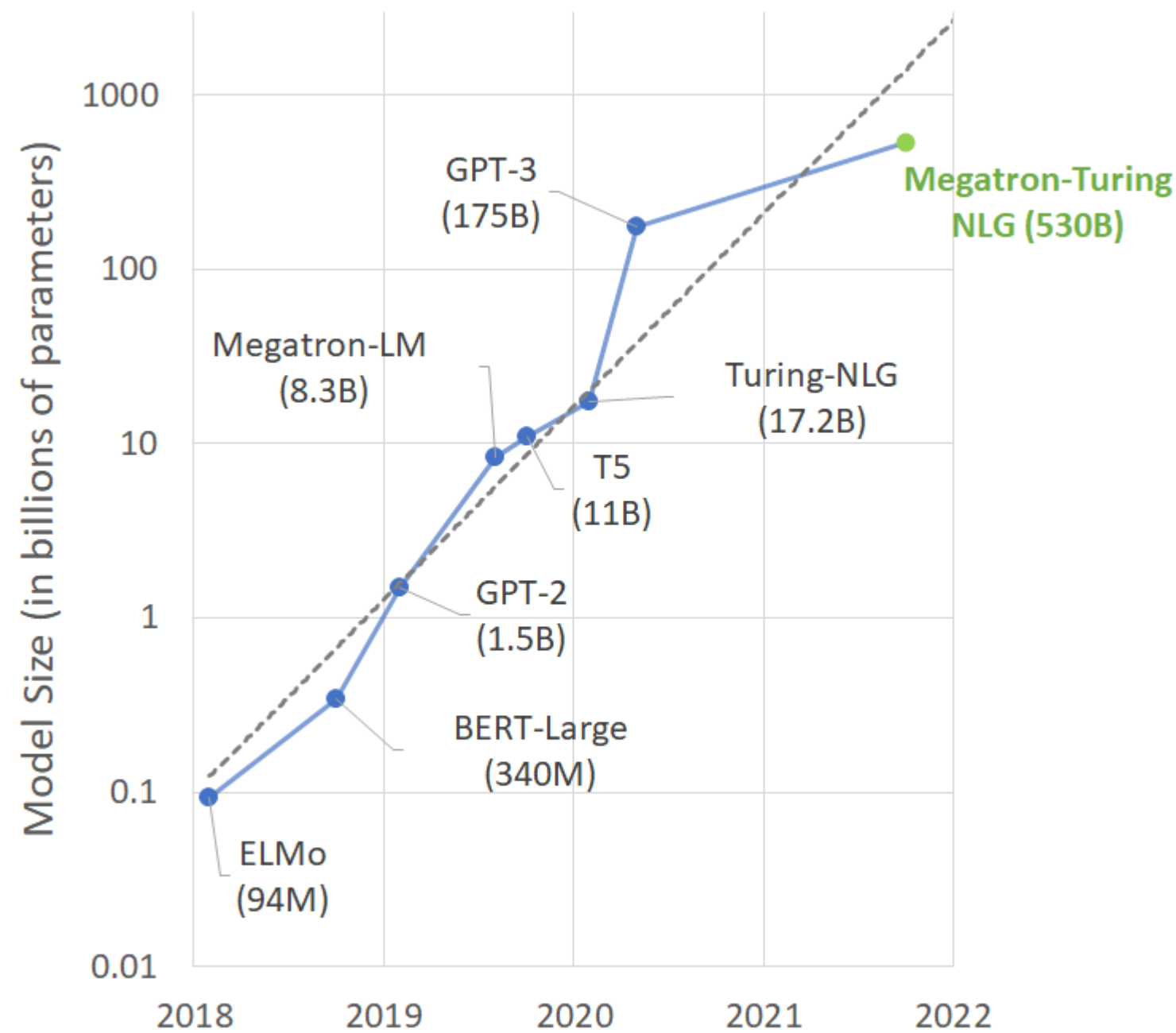
(c) Question Answering Tasks:
SQuAD v1.1



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

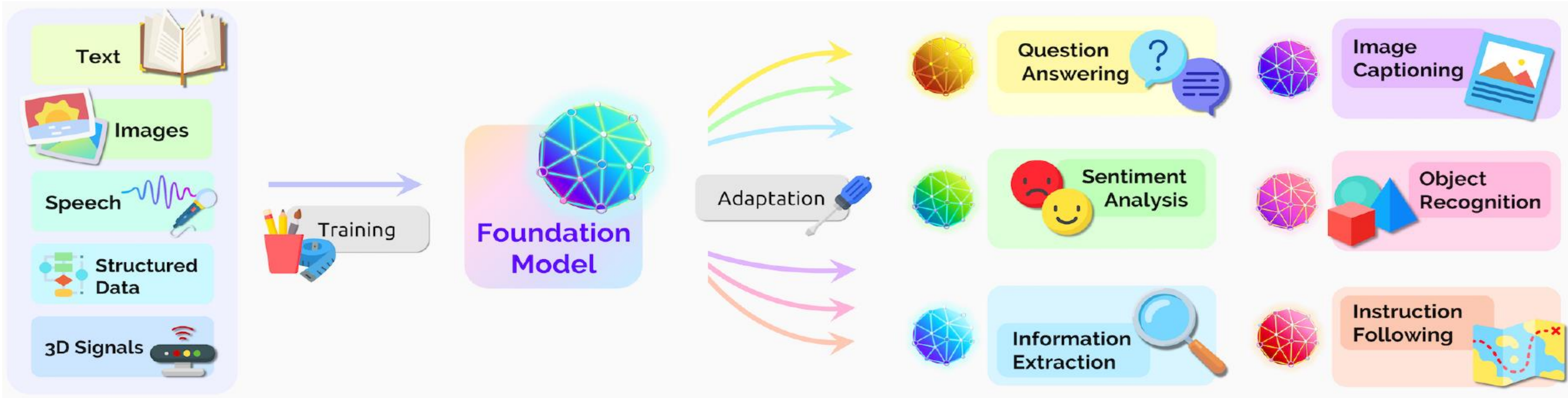
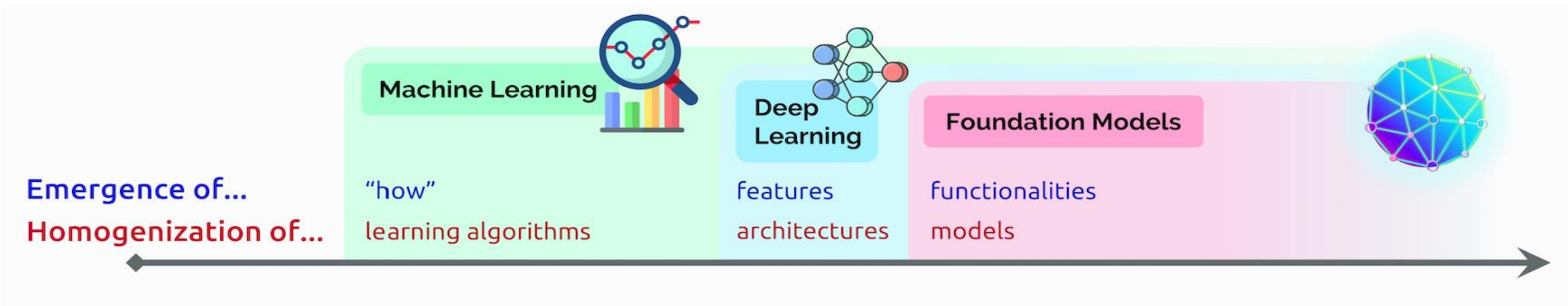
Нейросетевые модели языка

Рост числа параметров нейросетевых трансформерных моделей языка



От трансформеров к Foundation Models

Мультимодальное семантическое векторное пространство



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: ₹100М русский язык + ₹100М 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.

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

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

Сравнение разметки, сгенерированной алгоритмом, с разметкой 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 безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

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

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

связь РПОВТОР
РПОВТОР РЛИШН ПРОБЛЕМА
РПОВТОР РПОВТОР РПОВТОР
РЛИШН
РПОВТОР
РПОВТОР
РПОВТОР
РПОВТОР ГОДНОР ГОДНОР ГОДНОР
ГВИДОВР РПОВТОР
РПОВТОР РПОВТОР
РПОВТОР РПОВТОР
РПОВТОР ГВИДОВР РПОВТОР
РПОВТОР
РПОВТОР

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 безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

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

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

РПОВТОР T1
РПОВТОР T1
РПОВТОР T2 РПОВТОР T1
ПРОБЛЕМА РПОВТОР T2
РЛИШН
ПРИМЕР РПОВТОР T3
РТАВТ T4 РПОВТОР T1 РГ
РПОВТОР T1
РТАВТ T4
РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
РТАВТ T4 РПОВТОР T1
ПОЯСНЕНИЕ
РПОВТОР T1
РПОВТОР T1

Примеры задач разметки текстов в NLP/ML

- распознавание онимов (named entity recognition, NER)
- распознавание частей речи (part of speech tagging, POS)
- выделение тональности номинатива (sentiment analysis, SA)
- выделение синтаксических связей (syntax parsing)
- выделение семантических ролей (semantic role labeling, SRL)
- выделение текстовых полей данных (slot filling)
- выделение полей в библиографических записях
- сегментация научных или юридических текстов
- разрешение анафоры, кореферентности, эллипсиса

Пример: разметка онимов (NER)

О́ним (имя собственное) служит для выделения именованного объекта среди других объектов, его индивидуализации и идентификации

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

Примеры категорий:

- персона, организация, локация, время
- ссылка на нормативно-правовой акт
- заболевание, симптом, препарат
- биологический вид
- астрономический объект

Legend: Person p, Organization o, Other z, Location l, Date d

Shinzō Abe is a Japanese politician serving as the 63rd and current Prime Minister of Japan and Leader of the Liberal Democratic Party (LDP) since 2012, previously being the 57th officeholder from 2006 to 2007. He is the third-longest serving Prime Minister in post-war Japan.[1]

Abe comes from a politically prominent family and was first elected Prime Minister by a special session of the National Diet in September 2006. Then aged 52, he became Japan's youngest post-war Prime Minister and the first to have been born after World War II. Abe resigned on 12 September 2007 for health reasons. He was replaced by Yasuo Fukuda, the first in a

Пример: разметка семантических ролей (SRL)

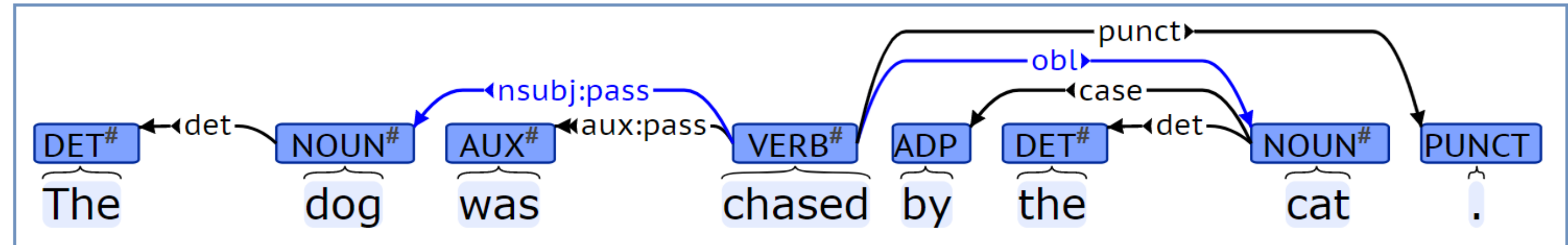
Задача: найти в предложении *актанты* — именные группы, обозначающие участников ситуации и их *семантические роли*.

Примеры семантических ролей:

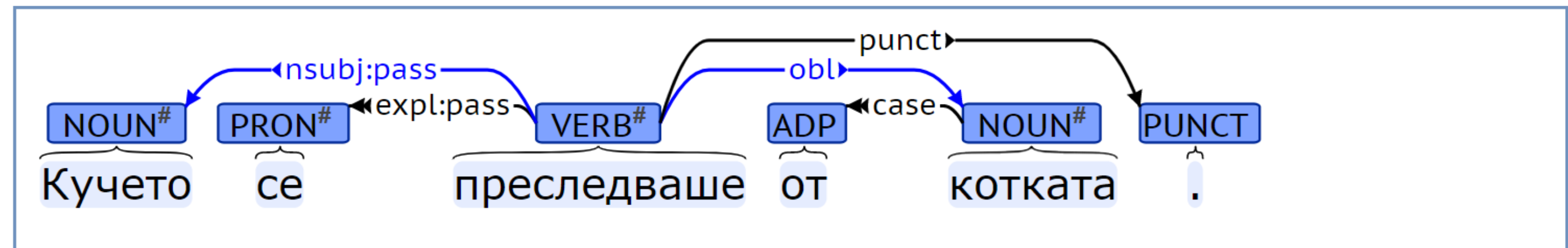
- **агенс:** одушевлённый инициатор и контролёр действия
- **пациенс:** участник, на которого направлено действие
- **бенефактив:** участник, получающий пользу или вред
- **адресат:** получатель сообщения (может быть бенефактивом)
- **инструмент:** посредством чего осуществляется действие
- **экспериенцер:** носитель чувств и восприятий
- **стимул:** источник восприятий
- **источник:** исходный пункт движения
- **цель:** конечный пункт движения

Пример: частеречная и синтаксическая разметка

английский:



болгарский:



теги
частей
речи

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

Выделение и тегирование фрагментов текста

Нотация BIOES (begin-inside-outside-end-single) для выделения начала и конца фрагмента

Для задачи распознавания именованных сущностей:

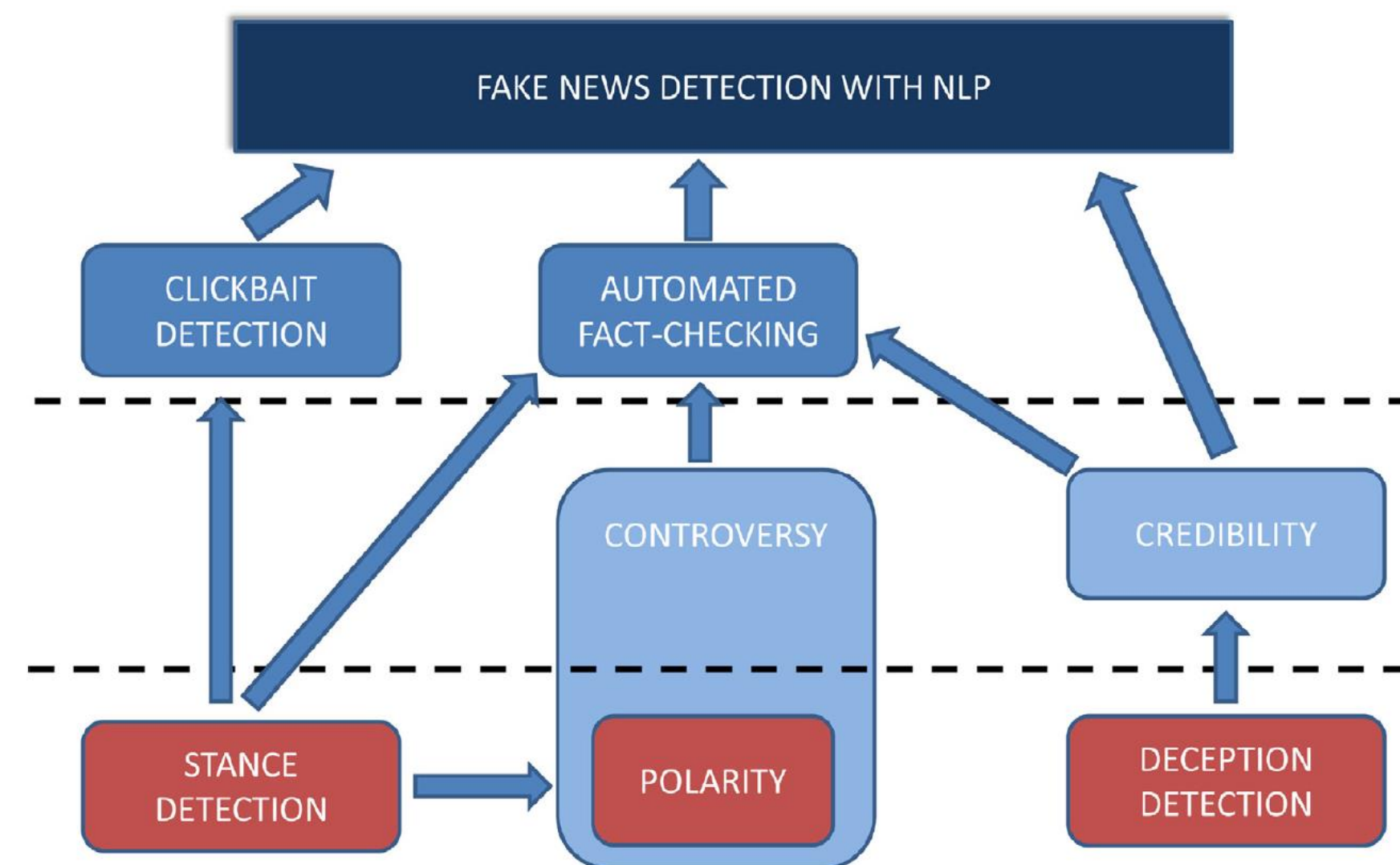
B-PER I-PER I-PER I-PER E-PER OUT OUT S-LOC
Карл Фридрих Иероним фон Мюнхгаузен родился в Боденвердере

Для задачи определения семантических ролей:

B_ACT **I_ACT** **I_ACT** **O** **B_NUM_PER** **O** **B_LOC** **I_LOC**
Book **a** **table** **for** **3** **in** **Domino's** **pizza**

Область исследований «Fake News Detection»

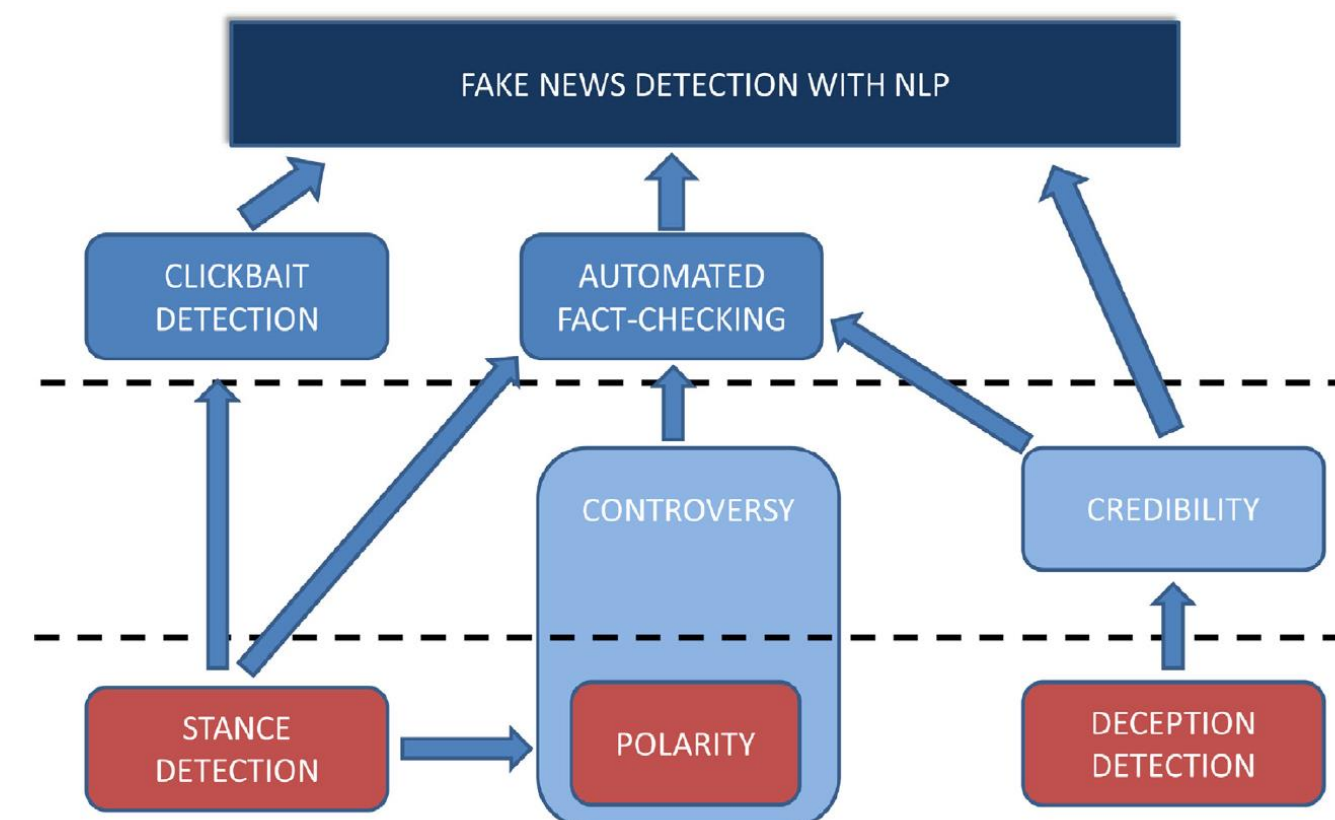
1. Deception Detection
выявление обмана в тексте новости
2. Automated Fact-Checking
автоматическая проверка фактов
3. Stance Detection
выявление позиции за/против запроса (claim)
4. Controversy Detection
выявление и кластеризация разногласий
5. Polarization Detection
классификация позиций по многим темам
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.*

Чего-то не хватает...

1. **Fake News** – не единственный и не самый сильный инструмент политики постправды.
2. **Пропаганда** использует не только фейки, но и подтасовку фактов, замалчивание, манипулятивные воздействия и т.д.
3. **Информационные войны** нацелены на разрушение социокультурного кода — образов, идей, ценностей, установок.
 - Как распознавать манипулятивные воздействия и идеологические атаки?
 - Как находить разногласия и замалчивание?
 - 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 дискурса и система подзадач ML/NLP для его детекции

воздействия → **фейки** → **пропаганда** → **инфо-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 конструкторов картины мира: ценностей, идеологем, мифологем
8.  оценивание возможных психо-эмоциональных реакций реципиента
9.  выявление целевых аудиторий воздействия
10.  **оценивание и предсказание скорости распространения (virality prediction)**
11.  **оценивание достоверности источников (credibility scores)**
12.  детекция деструктивных воздействий (угроз, провокаций, вербовки, экстремизма)

Четыре основных типа подзадач ML/NLP

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 идеологем, ценностей, элементов социокультурного кода
- выявление психо-эмоциональных реакций и целевых аудиторий
- выявление мнений, тональных оценочных суждений

4. Кластеризация или тематическое моделирование

- кластеризация мнений по заданной теме (controversy detection)
- выявление поляризации общественного мнения (polarization detection)

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

... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ... *(Kiev opinion)*

... По словам Захарченко, Киев встретит свой "ужасный конец" ... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду ... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как в республиках, так и в России ... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места ... *(Moscow opinion)*

Subject

Object

Agent

Locative

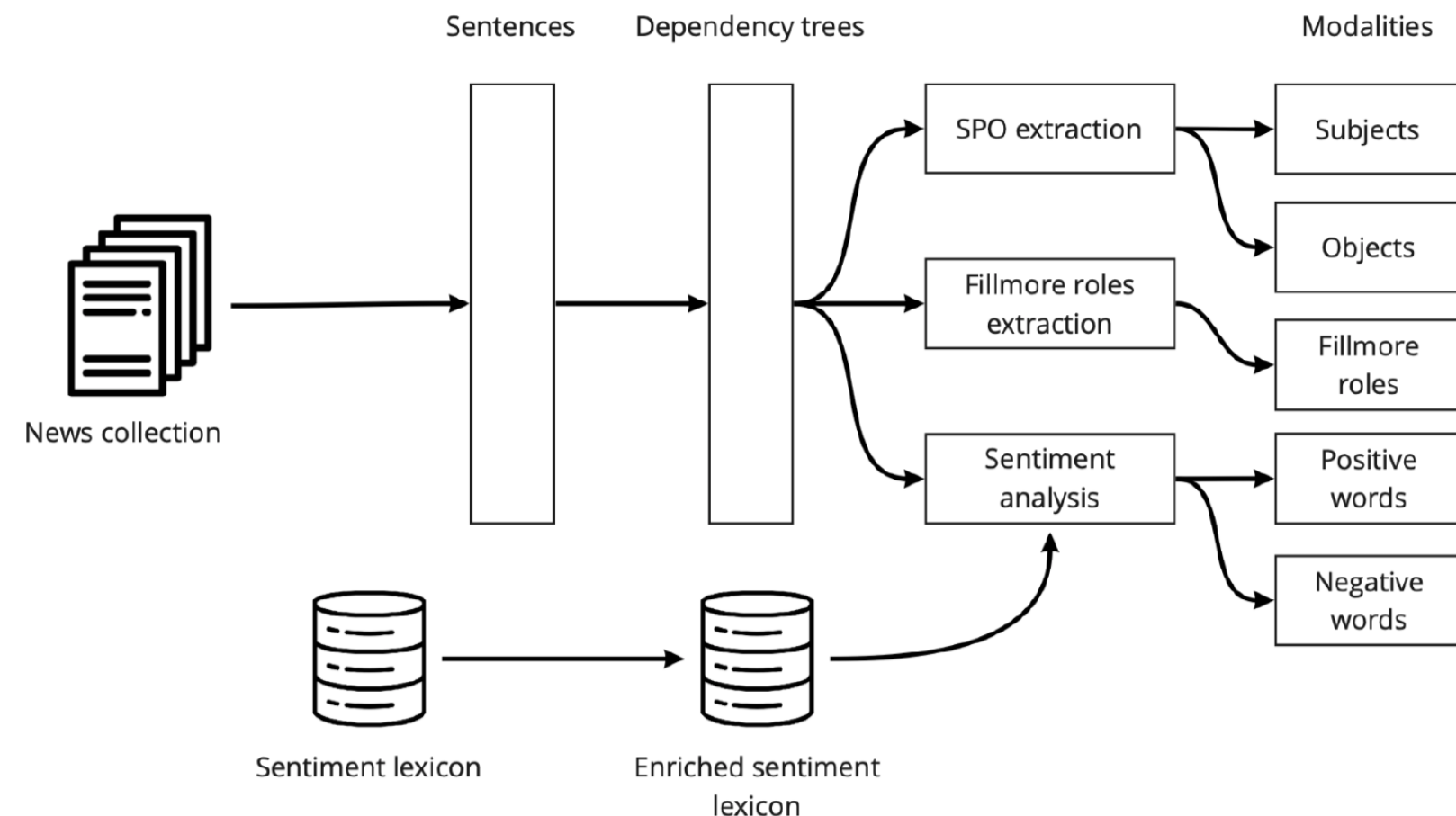
Negative lexicon

Dependent word

«Порошенко», «Россия», «Украина» встречаются одинаково часто, однако:

- «Порошенко» — субъект в первом тексте и объект во втором;
- «Россия» — агент в первом тексте и локация во втором;
- негативная тональность: «Россия», «Кремль» в 1-м, «Киев», «Украина» во 2-м

Задача выделения мнений в теме или событии



Modalities	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
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	<i>Pr</i>	<i>Rec</i>	<i>F1</i>
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 триплетных фактов (SPO), номинативов, их семантических ролей по Филлмору и их тональных окрасок. Все они используются в модели тематической векторизации как модальности.

Feldman D. G., Sadekova T. R., Vorontsov K. V. [Combining Facts, Semantic Roles and Sentiment Lexicon in A Generative Model for Opinion Mining](#). Dialogue 2020.

Обобщение разметки: путь к стандартизации

Пик научной фантастики (и советской, и западной) пришелся на 1960–1970-е годы. Однако в 1970-х годах этот жанр начал постепенно затухать и сходить на нет, уже в 1980-х на Западе начинает набирать силу жанр фэнтези. Конечно же, это неслучайно. Именно 1960-е годы стали пиком научно-технического прогресса в XX веке. К тому времени закончилась первая половина XX столетия, за эти полсотни лет было изобретено столько, что все казалось возможным, верилось, что прогресс будет нарастать по экспоненте: **1960-е — это мир безудержного социального и культурно-технического оптимизма**. Человек полетел в космос, запустил искусственные спутники и задумался об освоении других планет.

Но этот порыв человечества в будущее создавал определенную угрозу для власти имущих как на Западе, так и в Советском Союзе. И уже в 1960-е годы перед сотрудниками Тавистокского института изучения человека в Великобритании (причем по иронии судьбы он располагается в графстве Девоншир, рядом с дартмурскими болотами, где разыгрывалась мрачная драма «Собаки Баскервильей» Конан Дойля) **была поставлена задача притормозить научно-технический прогресс путем внедрения определенных информационно-психологических и организационных моделей**. В частности, стартовала работа по созданию молодежных и женских субкультур и движений (именно в это время как по заказу появились The Beatles, The Rolling Stones, стал развиваться экологизм).

Одна из главных задач, поставленных перед Тавистокком, звучала так: to stamp out the cultural optimism of the 1960s (искоренить, вырубить, вытравить культурный оптимизм 1960-х годов). А **научная фантастика, особенно советская, безусловно, была оптимистической по своему настрою**.

Некоторые менее оптимистические ноты (не могу их назвать пессимистическими, но они выглядели более сложными, чем просто оптимизм) прослеживались у ряда писателей в соцлагере, в частности в книгах Станислава Лема (достаточно почитать его «Астронавтов» и «Магелланово облако»). Однако общий настрой советской фантастики до середины 1960-х годов был преимущественно оптимистичным — это видно и по творчеству братьев Стругацких, и по романам Ивана Ефремова.

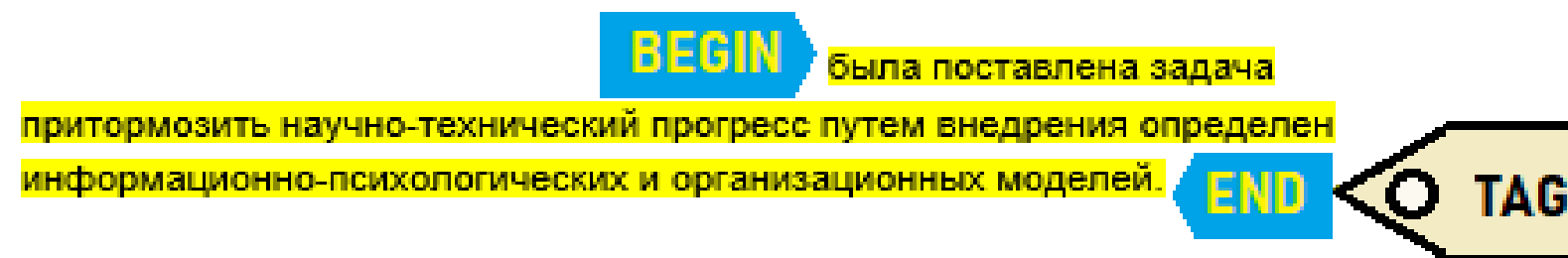
Первый доклад Римскому клубу (он создан в 1968 году) назывался «Пределы роста». В нем утверждалось, что человечество в своем индустриальном развитии достигло пределов, избыточно давит на природную среду, надо тормозить промышленно-экономическое развитие, перейдя к «нулевому росту». То есть 50 процентов всех средств должно идти на нейтрализацию негативных последствий, которые несет индустриальное развитие.

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

Методики оценивания: путь к стандартизации

- В основе методики — сравнение пар разметок текста: «модель \leftrightarrow эксперт», «эксперт-1 \leftrightarrow эксперт-2», на основе оптимального сопоставления их элементов
- Согласованность разметок (A,B) измеряется многими критериями, вычисляется их средневзвешенная согласованность **Consist(A,B)**
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по выборке **Consist(A,E)** разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по выборке **Consist(E1,E2)** разметок двух экспертов, E1 и E2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

Выводы о тенденциях

1. Предобученные модели внимания / трансформеры позволяют решать всё более трудные задачи NLP / NLU
2. В том числе, стоит модели для мониторинга и детекции угроз в медийном информационном пространстве
3. Разметка текстовых данных — магистральный путь формализации гуманитарных знаний во многих областях
4. Идём к стандартизации моделей, разметок и оценивания

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

k.v.vorontsov@phystech.edu

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