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

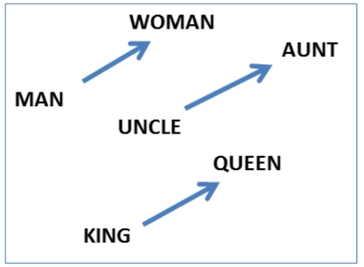
Эволюция подходов в обработке текстов

Декомпозиция задач по уровням «пирамиды NLP»

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

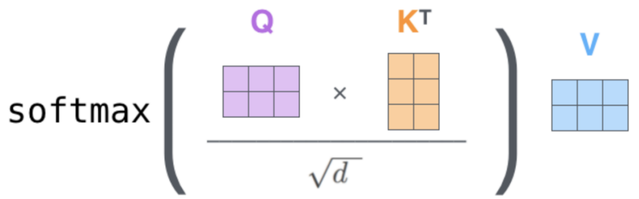
Модели векторизации слов (эмбедингов)

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

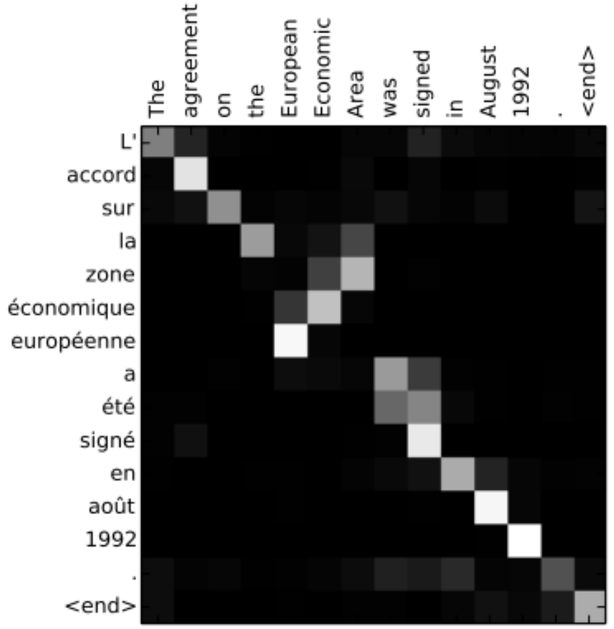


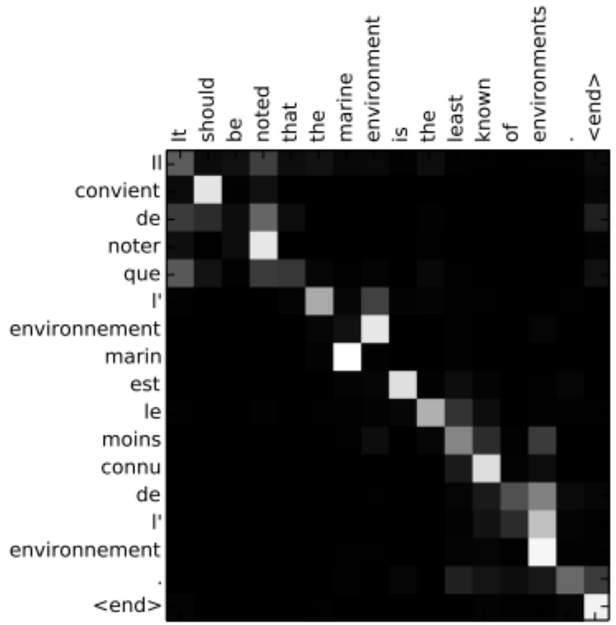
Нейросетевые модели контекстной векторизации

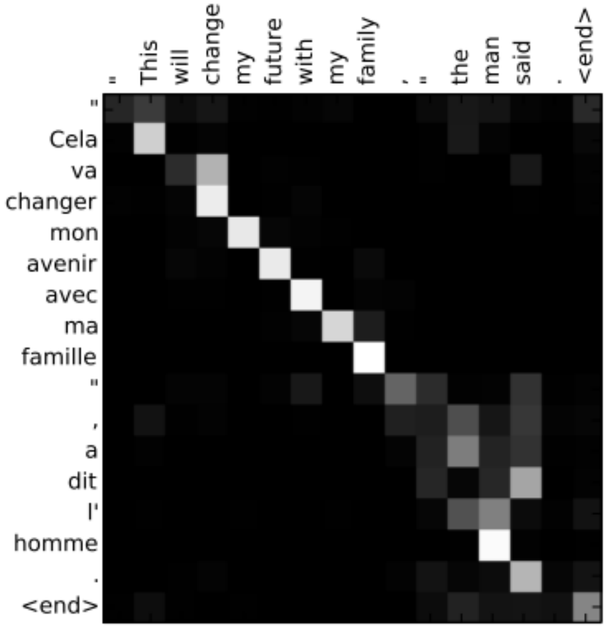
- рекуррентные нейронные сети: LSTM, GRU, …
- «end-to-end» модели внимания и трансформеры:
  машинный перевод [2017], BERT [2018], GPT-3 [2020], …



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

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.

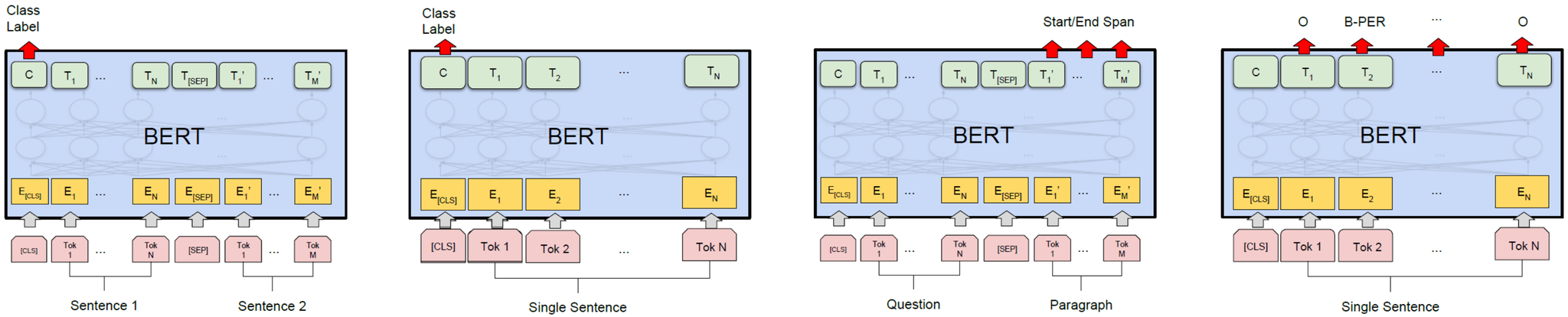
A giraffe standing in a forest with trees in the background.

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

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: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

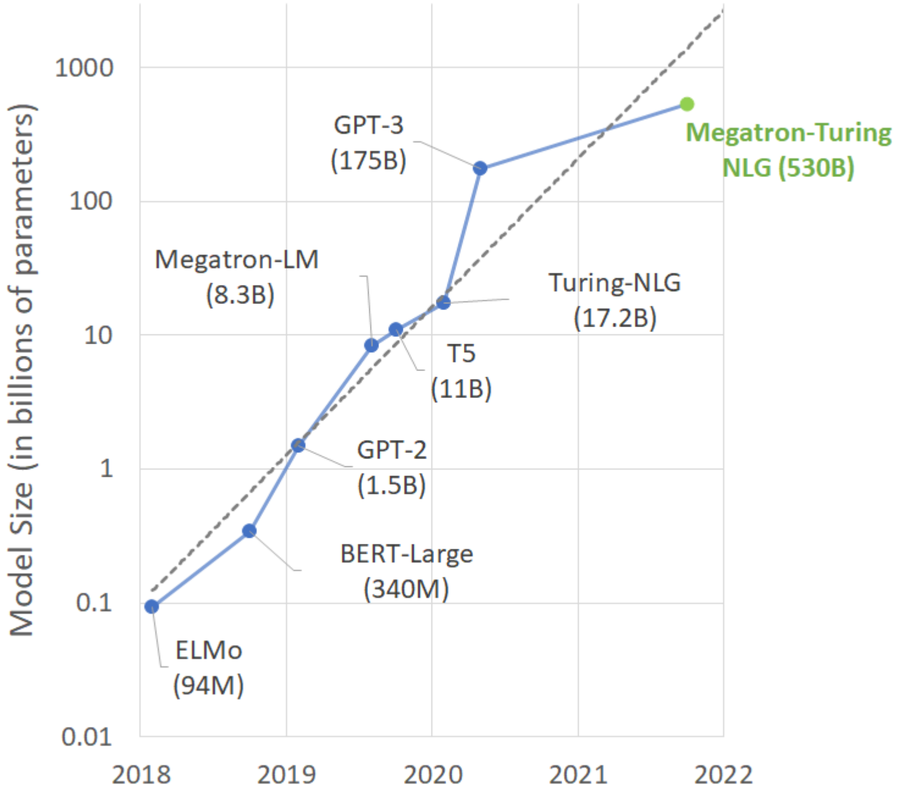
(b) Single Sentence Classification Tasks: SST-2, CoLA

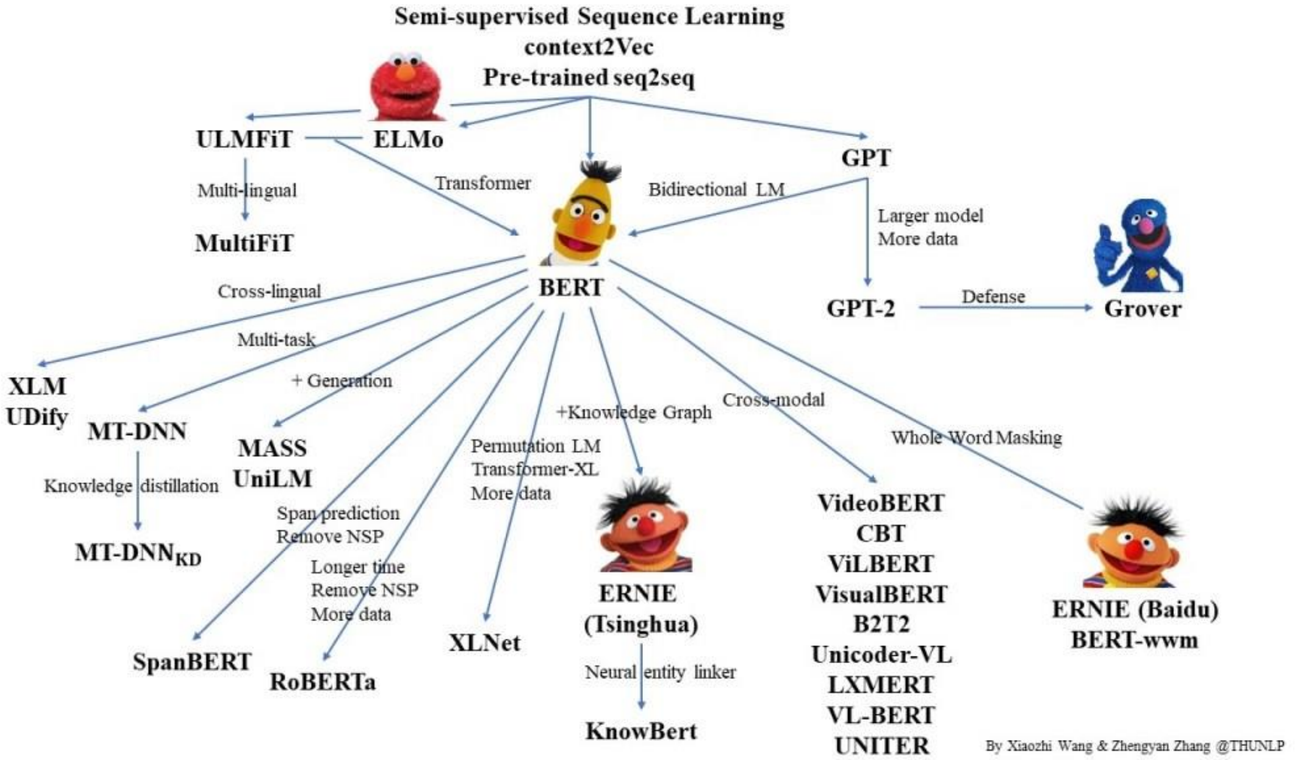
(c) Question Answering Tasks: SQuAD v1.1

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

Нейросетевые модели языка

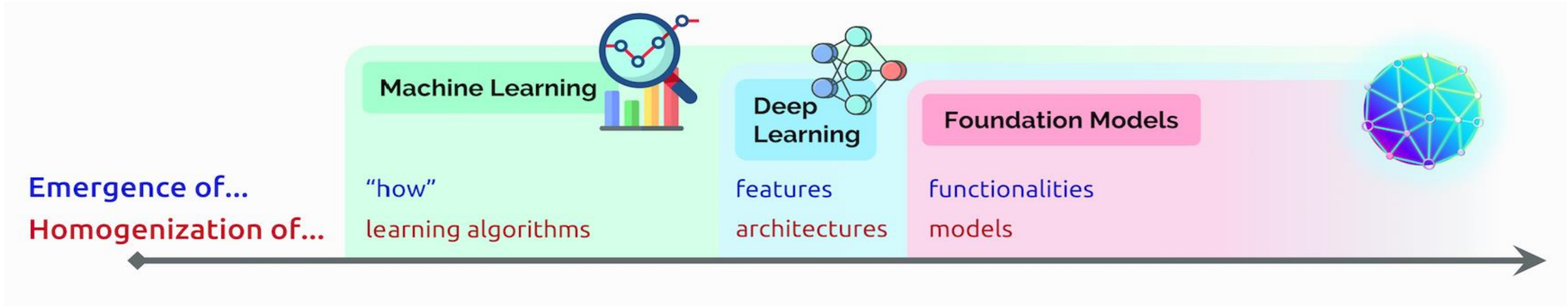
Рост числа параметров нейросетевых трансформерных моделей языка

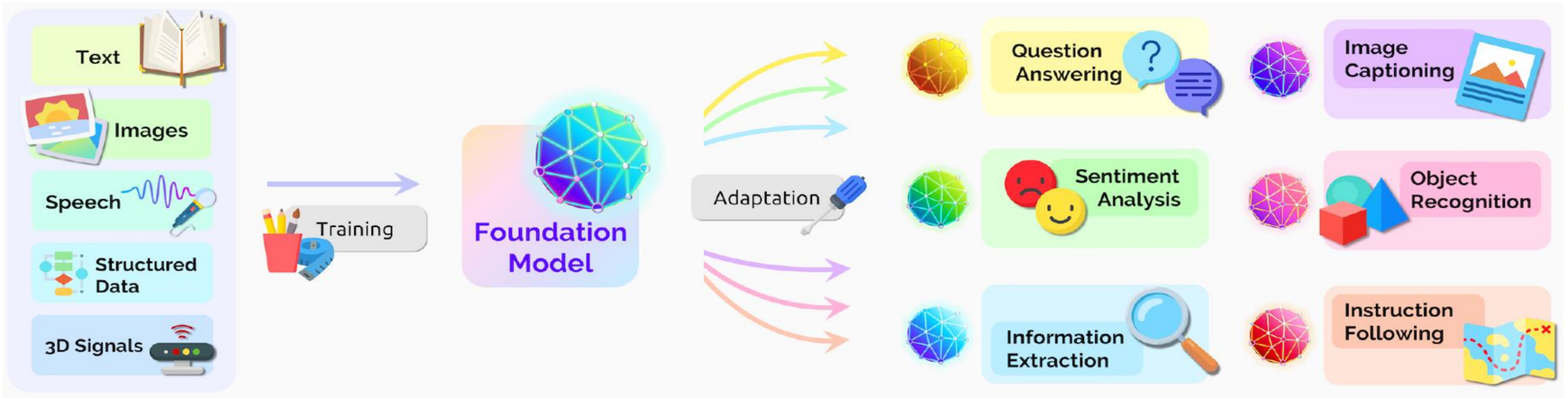




От трансформеров к Foundation Models

Мультимодальное семантическое векторное пространство





R.Bommasani et al. (CRFM, Stanford University). On the opportunities and risks of foundation models. 2021

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

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

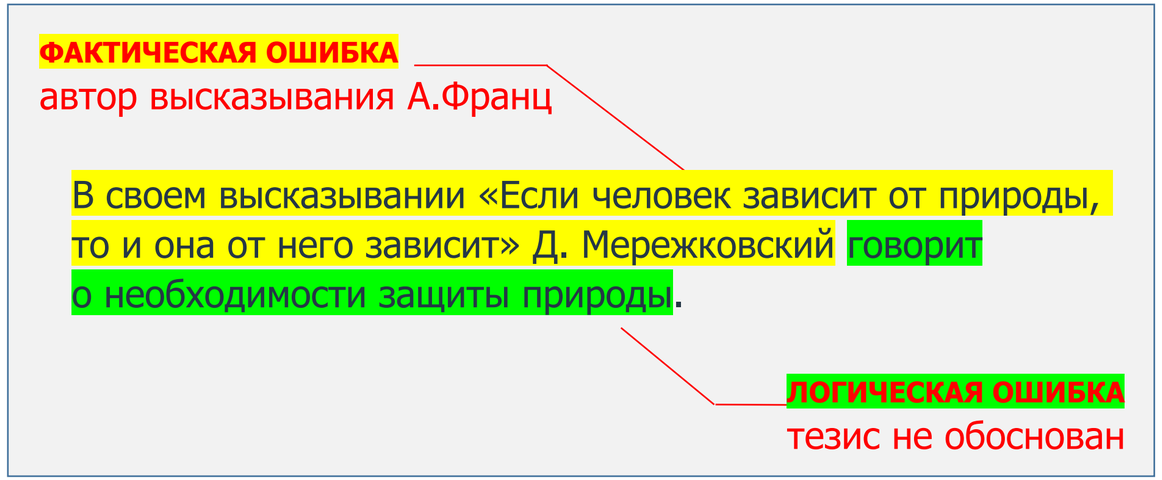
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

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.



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

Сравнение разметки, сгенерированной алгоритмом, с разметкой эксперта

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

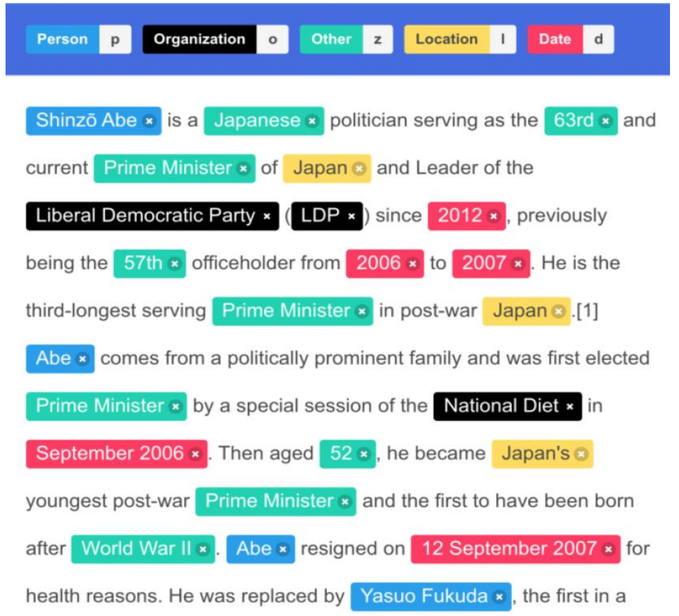
Пример: разметка онимов (NER)

Óним (имя собственное) служит для выделения именуемого 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



Пример: разметка семантических ролей (SRL)

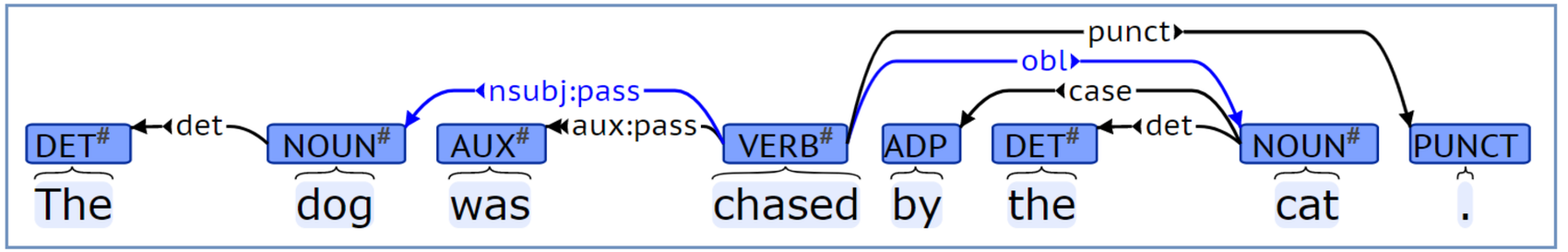
Задача: найти в предложении *актанты* — именные группы, обозначающие участников ситуации и их *семантические роли.*

Примеры семантических ролей:

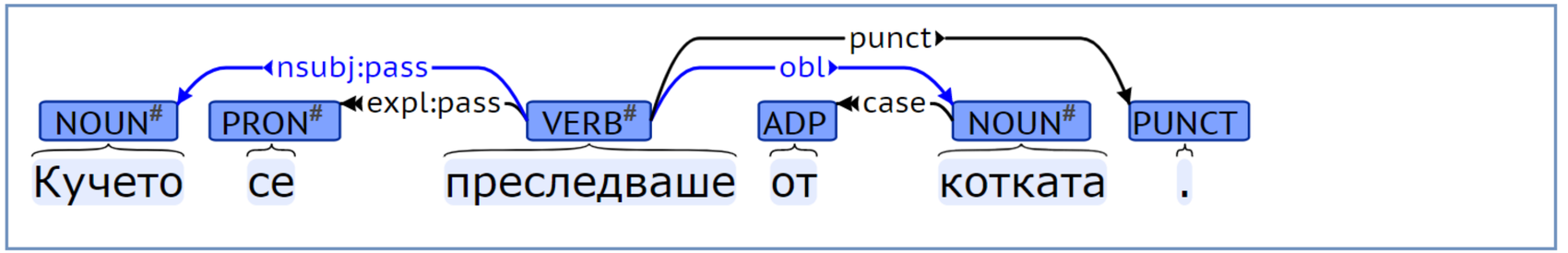
- **агенс**: одушевлённый инициатор и контролёр действия
- **пациенс**: участник, на которого направлено действие
- **бенефактив**: участник, получающий пользу или вред
- **адресат**: получатель сообщения (может быть бенефактивом)
- **инструмент**: посредством чего осуществляется действие
- **экспериенцер**: 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:

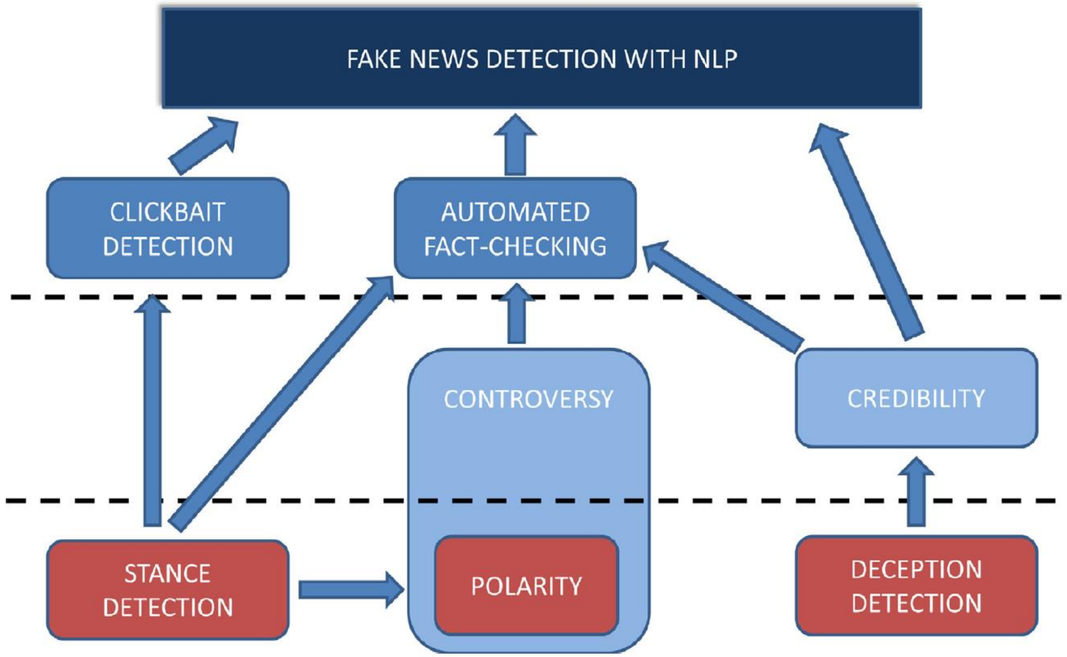
B-PER	I-PER	I-PER	I-PER	E-PER	OUT	OUT	S-LOC
Карл	Фридрих	Иероним	фон	Мюнхгаузен	родился	в	Боденвердере

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

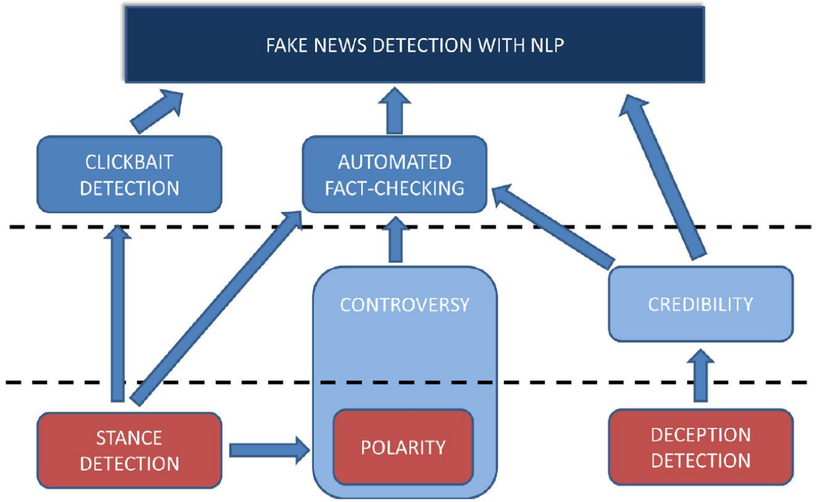


E.Saquete, D.Tomás, P.Moreda, P.Martínez-Barco, M.Palomar. **Fighting post-truth** using natural language processing: A review and open challenges. Expert Systems With Applications, Elsevier, 2020.

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 конструктов 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

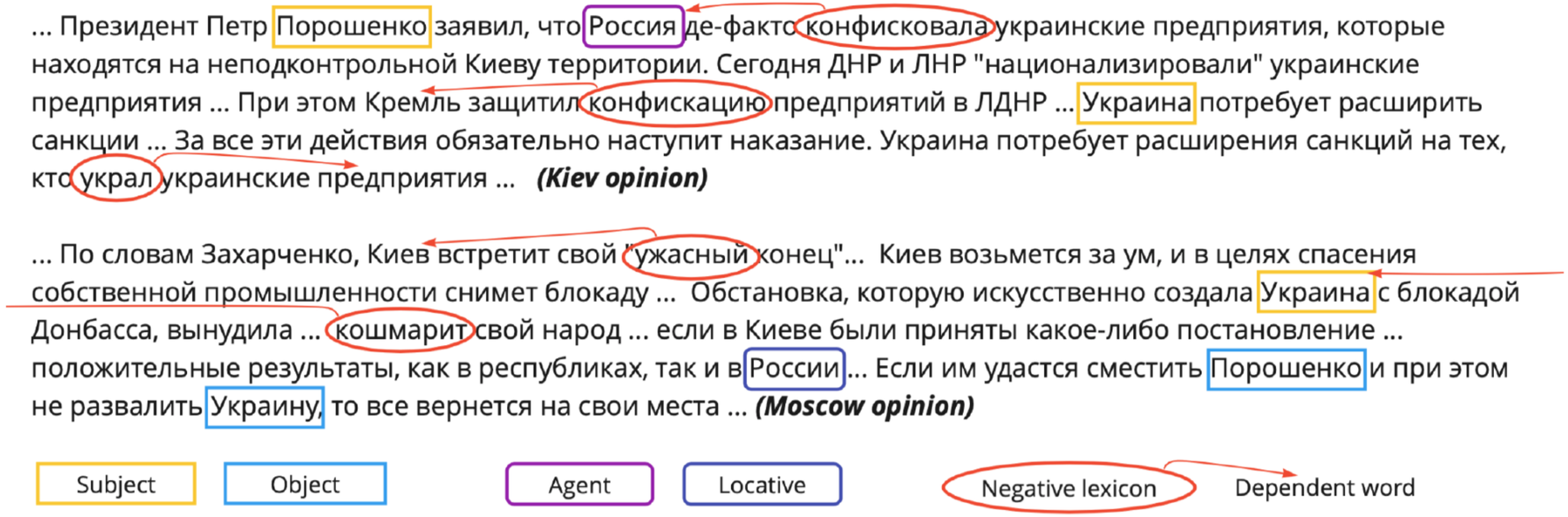
3. Деавторизация

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

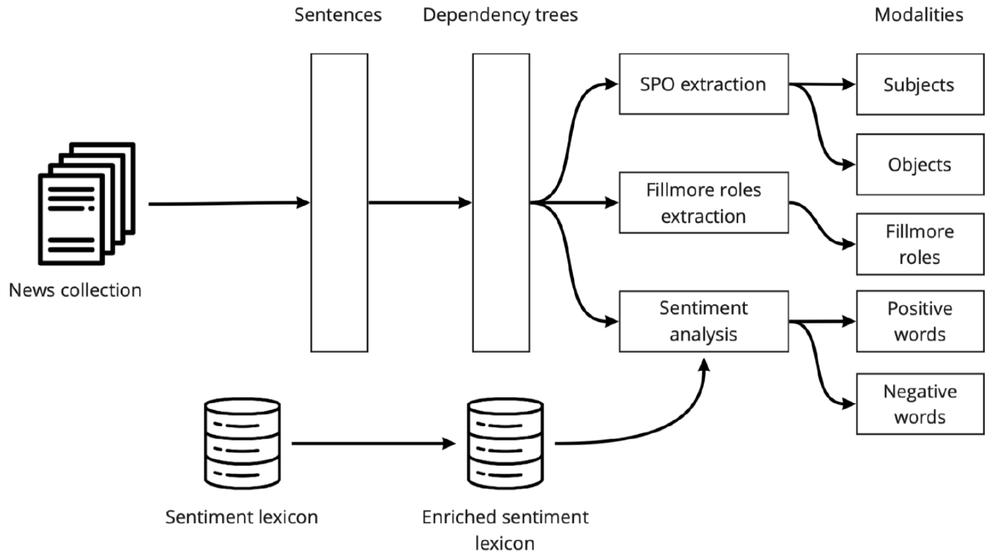
Задача выделения мнений в теме или событии



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

- «Порошенко» — субъект в первом тексте и объект во втором;
- «Россия» — агенс 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	*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

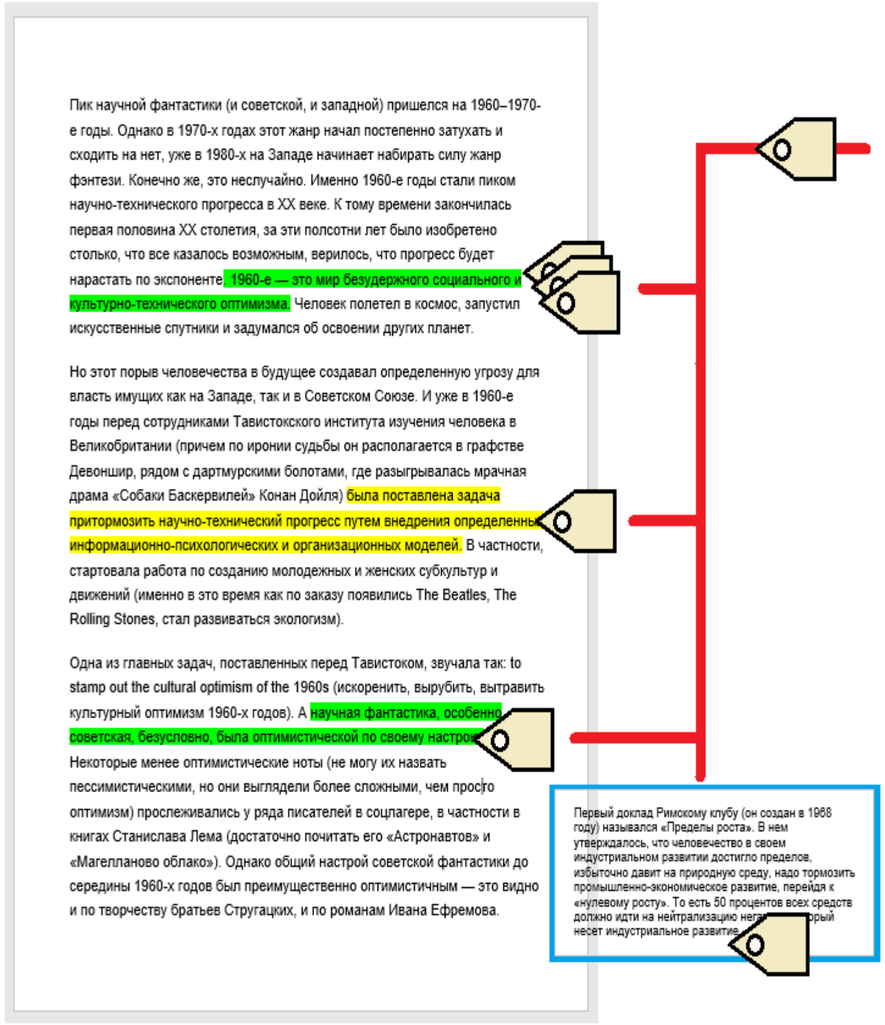
Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	**0.77**	**0.94**	**0.85**

Paris Trump

Мнение формализуется как устойчивое сочетание 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



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

Методики оценивания: путь к стандартизации

- В основе методики — сравнение пар разметок текста:
  «модель ↔ эксперт», «эксперт-1 ↔ 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, стоить 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