

Секционное научно-методическое заседание
«Межотраслевые вопросы стандартизации искусственного интеллекта»
Подкомитета 02 «Данные» (ПК02) Технического комитета по стандартизации
«Искусственный интеллект» (ТК164)

Задачи понимания естественного языка: на пути к стандартизации разметки и оценивания моделей

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

и.о. 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

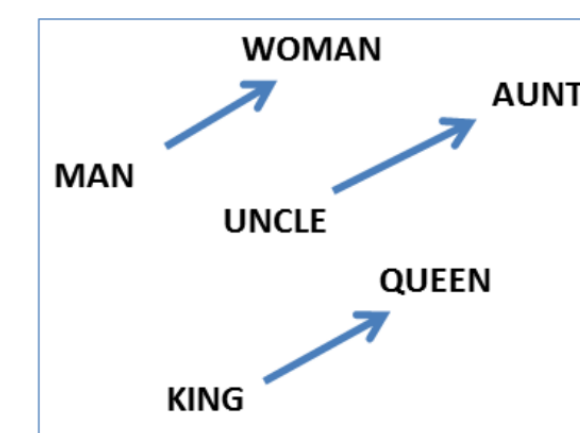
Декомпозиция задач по уровням «пирамиды NLP»

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



Модели векторизации слов (word embeddings)

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



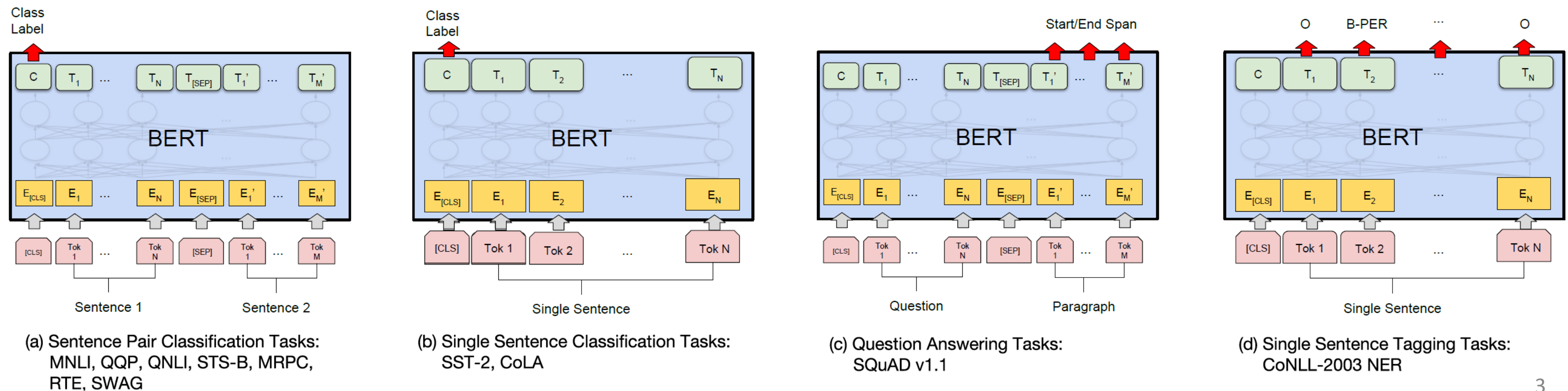
Нейросетевые модели контекстной векторизации

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

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

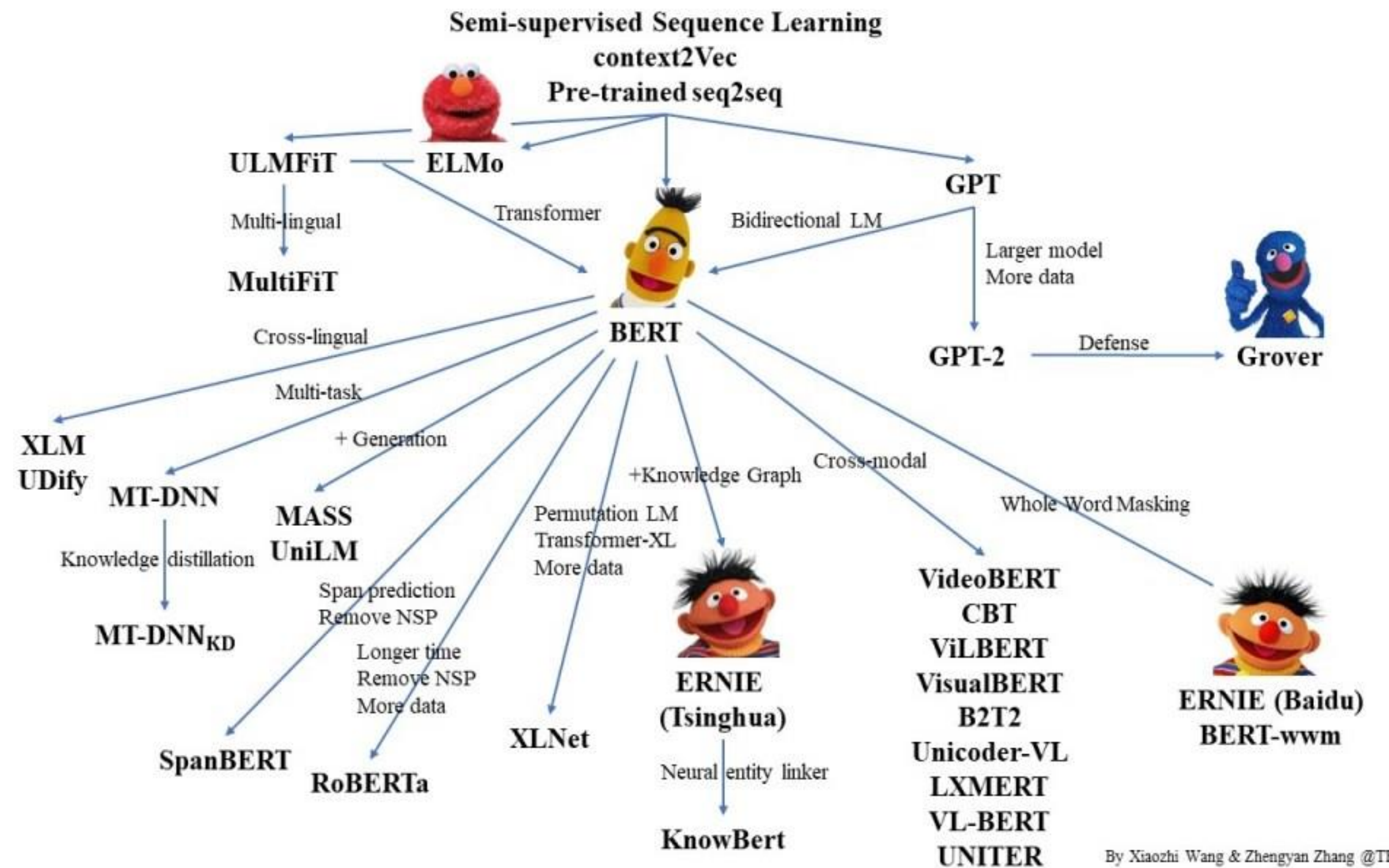
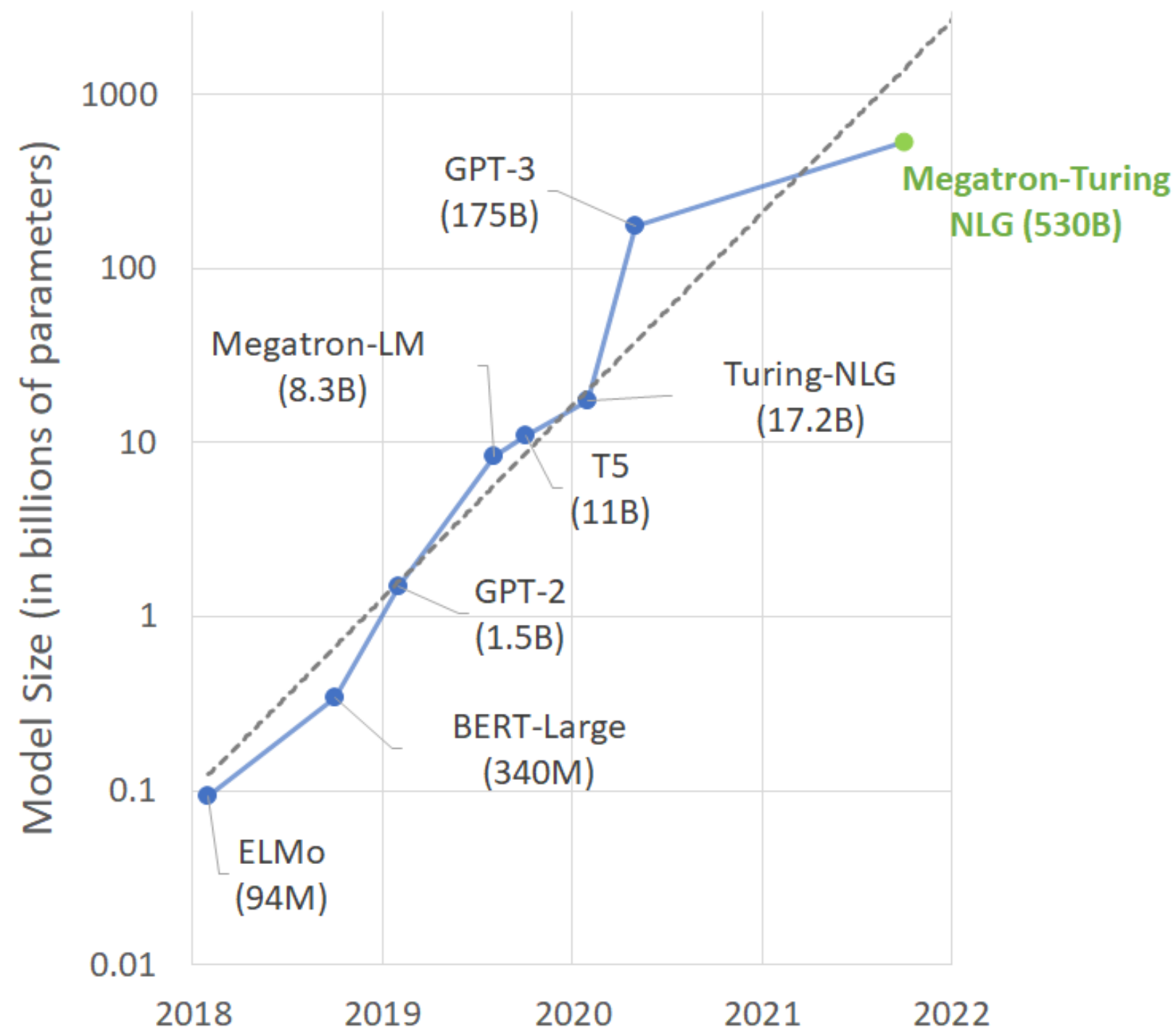
Нейросетевые модели языка

- 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 + дообучения на небольшой выборке



Нейросетевые модели языка

Рост числа параметров нейросетевых трансформерных моделей языка



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 Т1
РПОВТОР Т1
РПОВТОР Т2 РПОВТОР Т1
ПРОБЛЕМА РПОВТОР Т2
РЛИШН
ПРИМЕР РПОВТОР Т3
РТАВТ Т4 РПОВТОР Т1 РГ
РПОВТОР Т1
РТАВТ Т4
РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
РТАВТ Т4 РПОВТОР Т1
ПОЯСНЕНИЕ
РПОВТОР Т1
РПОВТОР Т1

Примеры задач разметки текстов в NLP/ML

- распознавание онимов (named entity recognition, NER)
- распознавание частей речи (part of speech tagging, POS)
- выделение тональности онима (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
- историческое событие, артефакт, документ
- товар, изделие, предмет искусства
- заболевание, симптом, препарат
- астрономический объект, телескоп

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

Выделение и тегирование фрагментов текста

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

Для задачи распознавания именованных сущностей:

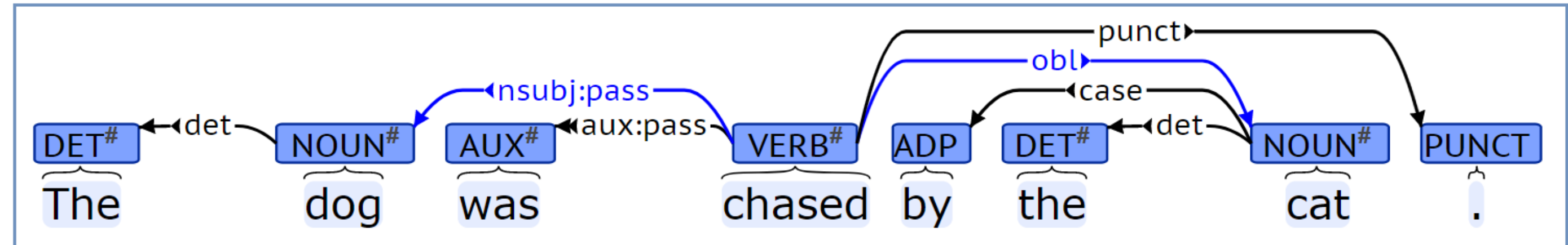
B-PER I-PER I-PER I-PER E-PER OUT OUT S-LOC
Карл Фридрих Иероним фон Мюнхгаузен родился в Боденвердере

Для задачи определения семантических ролей:

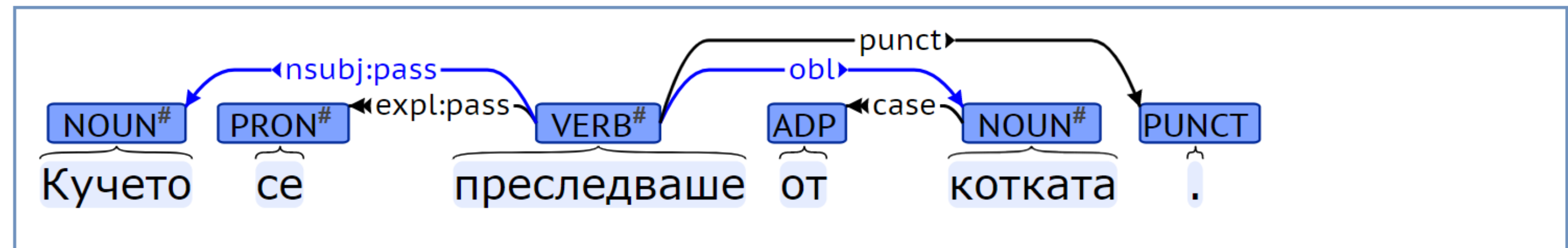
B_ACT **I_ACT** **I_ACT** **O** **B_NUM_PER** **O** **B_LOC** **I_LOC**
Book **a** **table** **for** **3** **in** **Domino's** **pizza**

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

SuperGLUE: набор тестовых задач понимания ЕЯ

BoolQ: **Boolean Questions.**

Каждый пример - короткий текст и вопрос с ответом «Да-Нет»

CB: **Commitment Bank**

Трех-классовая оценка, следует ли вывод из текста, противоречит тексту или нейтрален

COPA: **Choice of Plausible Alternatives**

Выбор причинно-следственно связанных предложений из нескольких альтернатив

MultiRC: **Multi-Sentence Reading Comprehension**

Выбор правильных ответов на вопросы по тексту из предложенных вариантов

ReCoRD: **Reading Comprehension with Commonsense Reasoning Dataset**

Предсказание пропущенных слов в предложении на основании текста и общего здравого смысла

RTE: **Recognizing Textual Entailment**

Оценка связи (лексической, логической и т.д.) двух коротких текстов

WiC: **Word in Context**

Оценка омонимов. Использование слов в одном или разных смыслах в двух предложениях.

WSC: **Winograd Schema Challenge**

Разрешение кореференции: определение, относятся ли указанные слова к одному объекту или нет

Russian SuperGLUE (<https://russiansuperglue.com>)

LiDiRus: **LiDiRus диагностика**

Диагностика

RCB: **Russian Commitment Bank**

Логический вывод

PARus: **Choice of Plausible Alternatives for Russian language**

Здравый смысл

MuSeRC: **Russian Multi-Sentence Reading Comprehension**

Машинное чтение

TERRa: **Textual Entailment Recognition for Russian**

Логический вывод

RUSSE: **Russian Words in Context (based on RUSSE)**

Здравый смысл

RWSD: **The Winograd Schema Challenge (Russian)**

Причинно-след. связь

DaNetQA: **Yes/no Question Answering Dataset for the Russian**

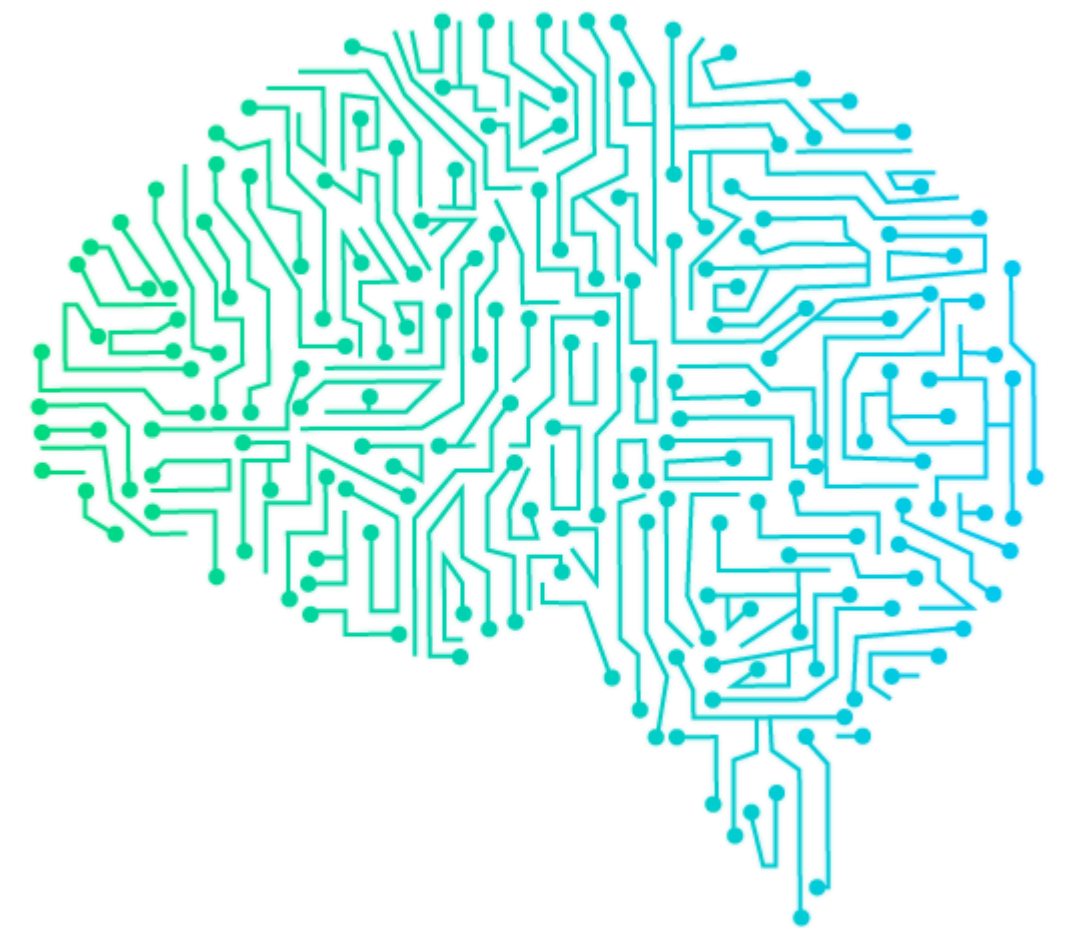
Знание

RuCoS: **Russian Reading Comprehension with Commonsense Reasoning**

Машинное чтение

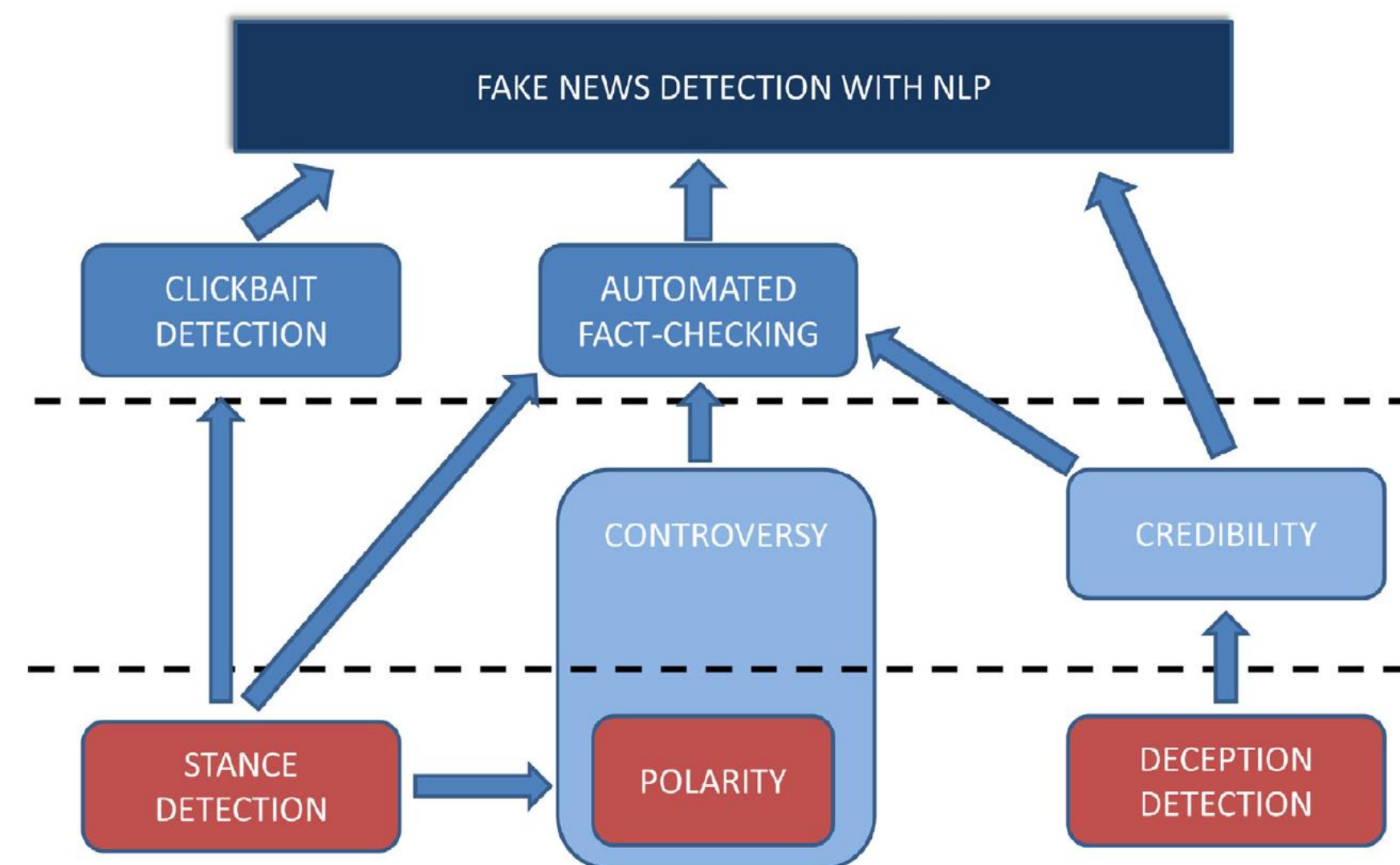


**russian
superglue**



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.*

1. Deception Detection (выявление обмана)

- **История:** более 50 лет исследований в психологии и криминологии
- **Задача** классификации текста на два класса: *обман / не обман*
- **Обучающие выборки:**
 - Контролируемый эксперимент: люди *врут / не врут* на заданную тему
 - Материалы судебных заседаний (датасет DECOUR)
 - Отзывы на товары/услуги, проверяемые с помощью краудсорсинга
- **Признаки** – лингвистические маркеры (Linguistic-Based Cues, LBC)
- **Критерии:** Ассурасу или F-мера 70–92% в зависимости от задачи
- На небольших датасетах классический ML лучше и проще DL
- Проблема переноса моделей на другие датасеты

Типы лингвистических маркеров

Манипулятивные и суггестивные приёмы

- многословие: плеоназмы, лишние слова, тавтологии, расщепления сказуемого
- избыточные повторы слов и фраз
- повышенная когнитивная сложность текста, перегруженные синтаксические конструкции
- повышенная экспрессивность, преобладание негативной тональности
- категоричность, психологическое давление

Уход от личной ответственности

- безличные глаголы, глаголы абстрактной семантики, модальные глаголы, объективация
- неконкретность, уклончивость, безличность, неопределённость высказываний

Подача информации

- оторванность от контекста: пониженная детализация места, времени, событий
- упрощение, пониженное лексическое разнообразие, лексическая недостаточность
- замалчивание фактов, сообщение ложных сведений (fact-checking, см. далее)

2. Automated Fact-Checking (проверка фактов)

- **История:** ручной fact-checking давно используется в журналистике
- **Задача** классификации текста целиком, по порядковой шкале: *True, Mostly True, Half True, Mostly False, False*
- **Обучающие выборки:**
 - Платформы для проверки фактов: Politifact, FullFact, FactCheck и др.
 - Соревнования: CLEF-2018,19,20,21, FEVER, SemEval (Rumour-Eval)
 - Датасеты: NELA-GT-2018,19, FakeNewsNet, Snopes и др.
- **Вспомогательная задача:** стоит ли отправлять текст на проверку?
Три класса: *Non-Factual Sentence, Unimportant, Check-Worthy*
(пример: ClaimBuster, <https://idir.uta.edu/claimbuster>, 2015)

3. Stance Detection (выявление позиции)

- **История:** задача textual entailment (текстового следования) – классификация пар текстов «текст $t \Rightarrow$ гипотеза h » на три класса: « h следует из t », « h противоречит t », « h не относится к t »
- **Задача:** классификация текста h относительно запроса (claim) t : *agree, disagree, discusses (позиция не высказана), unrelated*
- **Обучающие выборки:**
 - SNLI: 570K пар предложений: entail, contradict, independent
 - Датасеты: Emergent, SemEval-2016 6A(stance), FakeNewsChallenge FNC-1
- **Критерии:** F1-мера до 97% на новостях; Accuracy до 68% на Twitter

4. Controversy / 5. Polarization Detection

Две специальные разновидности задачи Stance Detection

- **Controversy Detection** (выявление полемики, разногласий):
 - кластеризация мнений без учителя
 - выделение сообществ сторонников каждого мнения в социальной сети
 - количественное оценивание объёма и динамики сообществ
- **Polarization Detection** (выявление поляризованности общества):
 - выявление разногласий по совокупности запросов или тем
- **Обучающие выборки:**
 - Датасеты социальных сетей, обычно Twitter
 - Википедия
- **Критерии:** Accuracy 73–83% (на Википедии, методом kNN)

6. Clickbait Detection (обнаружение кликбейта)













- **История:** задача появилась в 2016 году. Обнаружение заголовков или ссылок-приманок, не соответствующих сути контента
- **Задача:** классификация пары «заголовок, текст» на два класса
Задача аналогична Textual Entailment и Stance Detection
- **Признаки:** гиперболизация, противоречия, web-трафик
- **Обучающие выборки:**
 - Датасеты: Webis-Clickbait 2017 (32К заголовков) и др.
 - Соревнование: Clickbait challenge 2017
- **Критерии:** F1-мера до 68%; Ассурасу до 86%

7. Credibility Scores (Оценивание надёжности)

- **История:** старая задача в социологии, психологии, маркетинге
- **Задача:** оценить уровень доверия (credibility, trustworthiness) для источника (СМИ, блогера, пользователя) или отдельной новости
- **Признаки:**
 - распространение ненадёжного контента (spam, deception, fake и др.)
 - вероятность быть ботом (по диспропорции рассылок и качеству контента)
 - стиль контента, геолокация и образовательный уровень читателей
- **Обучающие выборки:**
 - много несопоставимых датасетов, отсутствует «золотой стандарт»
- **Критерии:** AUC до 89%; ассигасу до 81%; MSE до 0.33
 - много критериев, не хватает методологического единства

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 гиперболлизация
- 1.7 Моделирование негативного сценария
- 1.8 Вкрапление депрессивов

2. Позитивизация

- 2.1 Эвфемизация
- 2.2 Лозунговые слова и словосочетания
- 2.3 Позитивирующая гиперболлизация

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

Примеры приёмов манипулирования

Лозунговые слова

Также мэр **Владивостока Шестаков** рассказал, что в 2022 году **власти займутся улучшением работы общественного транспорта.**

Навешивание ярлыков

11 октября 2021 года Талибы назвали **провалом политику НАТО в Афганистане**

Эвфемизация

Кардашьян приложила немало усилий, чтобы помочь **супругу** в том числе и с его **психическими проблемами**, однако со временем ей стало это в тягость.

Негативирующая гиперболизация

Сергей Нетесов в беседе с «Известиями» указал на то, что **Россия** **стоит на пороге самой мощной волны пандемии за всё время существования COVID-19.**

Задача выявления поляризации мнений в теме

... 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

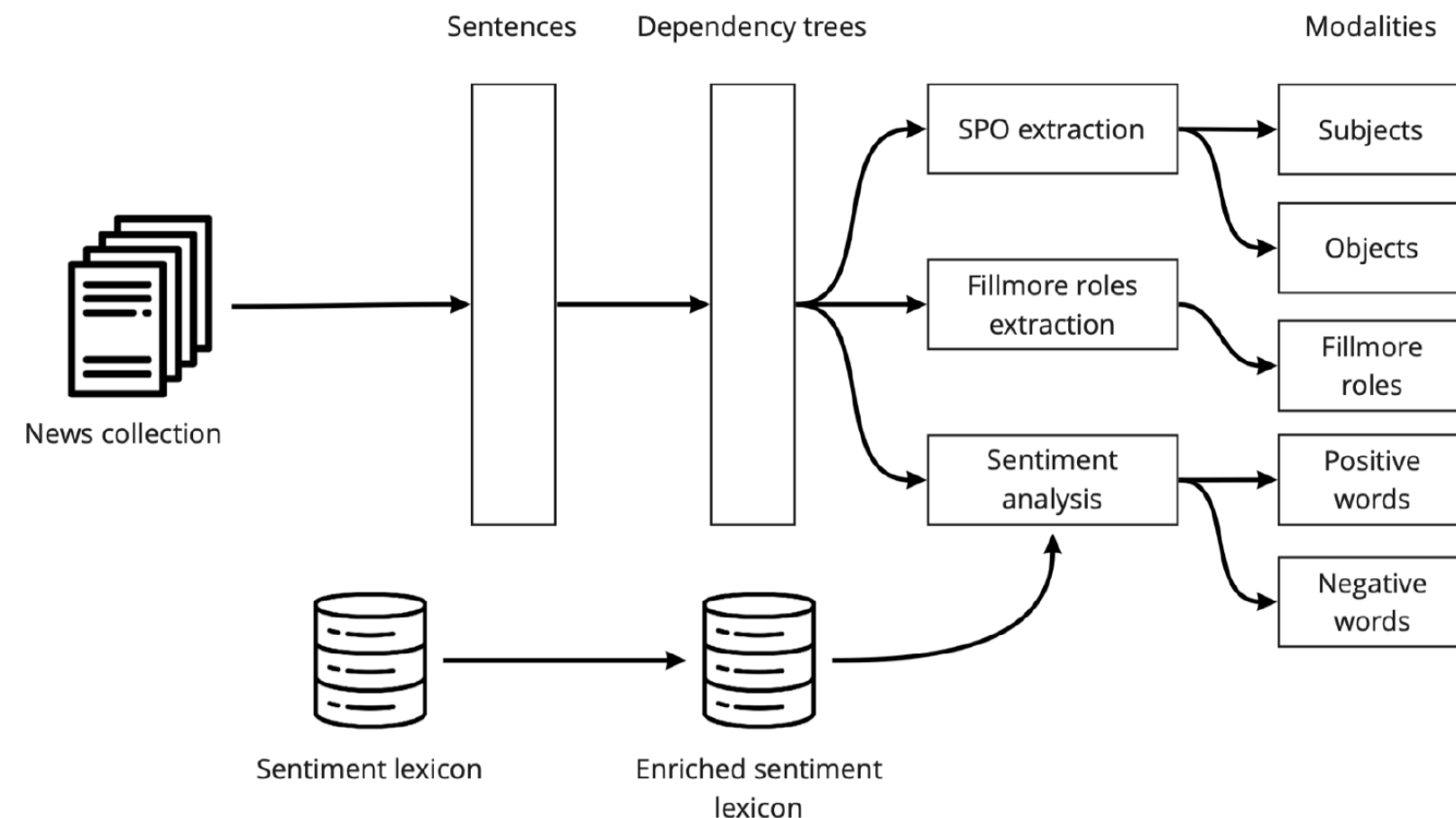
Negative lexicon

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 выявления поляризации мнений в теме



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 сходиться 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 Баскервилей» 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.

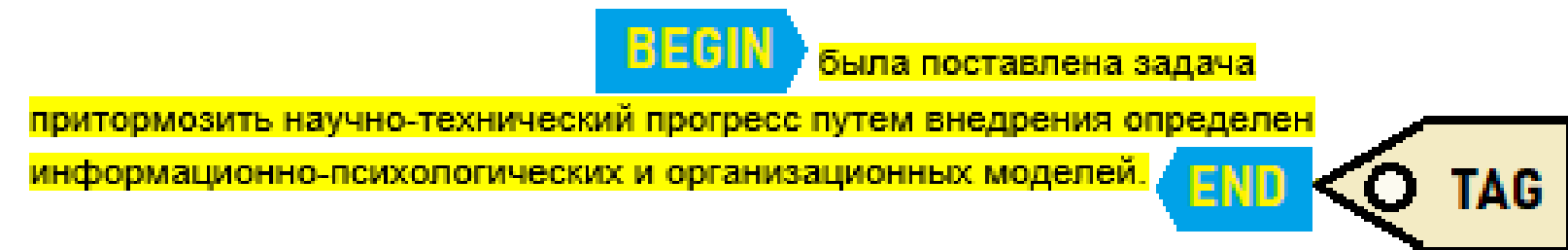
Первый доклад Римскому клубу (он создан в 1968 году) назывался «Пределы роста». В нем утверждалось, что человечество в своем индустриальном развитии достигло пределов, избыточно давит на природную среду, надо тормозить промышленно-экономическое развитие, перейдя к «нулевому росту». То есть 50 процентов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

Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, комментариев и тегов

Теги (классы) выбираются из словаря тегов

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



Ответ 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 $Con(A,B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по выборке $Con(A,E)$ разметки модели A и разметки 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

Базовые условия применения методики

1. Алгоритм (обученная модель) принимает на входе информацию об *объекте* и выдаёт выходную информацию об объекте, называемую *разметкой*.
2. Числовые критерии $M_1(X, Y) \dots M_N(X, Y)$ измеряют различные аспекты согласованности двух разметок X и Y одного и того же объекта.
3. Все критерии симметричны, $M_j(X, Y) = M_j(Y, X)$, принимают значения от 0 до 1 (или от 0% до 100%). Чем больше значение критерия, тем лучше согласованность. При полном совпадении разметок достигается максимальное значение 100%.
4. *Согласованность* $M(X, Y)$ разметок X и Y определяется как взвешенное среднее N критериев $M_j(X, Y)$ с неотрицательными весами w_j :

$$M(X, Y) = \frac{\sum_{j=1}^N w_j M_j(X, Y)}{\sum_{j=1}^N w_j}.$$

Веса являются параметрами методики и устанавливаются из соображений относительной значимости критериев. Возможно полное исключение некоторых критериев путём обнуления их весов, $w_j = 0$.

Базовые условия применения методики

5. Задана выборка объектов T . Для каждого объекта имеется множество разметок \mathcal{E} , выполненных экспертами, и разметка A , сгенерированная алгоритмом (обученной моделью). Экспертные разметки могут быть слабо согласованы друг с другом.
6. *Целью методики* является оценивание точности алгоритмической разметки A по отношению ко всем экспертным разметкам, по всем объектам выборки T .

Относительная точность разметки

Средняя Точность Алгоритмической Разметки показывает, насколько хорошо алгоритмические разметки согласуются с экспертными разметками:

$$\text{СТАР} = \text{avr}_T \left(\text{avr}_{\mathcal{E}} M(A, \mathcal{E}) \right).$$

Средняя Точность Экспертных Разметок показывает, насколько хорошо экспертные разметки согласуются друг с другом:

$$\text{СТЭР} = \text{avr}_T \left(\text{avr}_{\mathcal{E}, \mathcal{E}'} M(\mathcal{E}, \mathcal{E}') \right).$$

Относительная Точность Алгоритмической Разметки показывает, во сколько раз алгоритмическая разметка лучше согласуется с экспертными разметками, чем экспертные разметки согласуются друг с другом:

$$\text{ОТАР} = \frac{\text{СТАР}}{\text{СТЭР}} \times 100\%.$$

СТЭР является ориентировочным максимальным достижимым значением для СТАР.

Если $\text{ОТАР} \geq 100\%$, то можно утверждать, что алгоритмические разметки согласуются с экспертными разметками не хуже, чем экспертные разметки согласуются друг с другом; другими словами, что предсказательная модель работает не хуже экспертов.

Пример 1. Задача выявления манипуляций

Объектом является текстовый документ.

Разметкой текста называется последовательность $X = \{x_1, \dots, x_n\}$, каждый элемент которой x_i представляет собой тройку $x_i = (F_i, T_i, C_i)$ — « F_i фрагмент (начало B_i и конец E_i), T_i мишень манипуляции, C_i тип манипуляции».

Соответствие двух разметок $X = \{x_1, \dots, x_n\}$ и $Y = \{y_1, \dots, y_m\}$ — множество D пар элементов (x_i, y_k) , такое, что каждому x_i из X соответствует не более одного y_k и каждому y_k из Y соответствует не более одного x_i . Если для элемента x_i не найдено соответствия, « $x_i \rightarrow \emptyset$ ».

Соответствие двух разметок определяется по всем парам элементов (x_i, y_k) следующим образом. Вводится матрица потерь $L[i, k]$ размера $n \times m$:

$$L[i, k] = J_{ik} + [J_{ik} = 1] + 2[T_i \neq T_k] + [C_i \neq C_k],$$

где J_{ik} — расстояние Жаккара, определяемое для пары элементов (x_i, y_k) .

Соответствие D между разметками минимизирует сумму потерь (задача о назначениях):

$$Q(D) = \frac{1}{2} \sum_{(i,k) \in D} L[i, k] + \sum_i [x_i \rightarrow \emptyset] + \sum_k [y_k \rightarrow \emptyset] \rightarrow \min_D.$$

Пример 1. Задача выявления манипуляций

Критерий M1: точность и полнота поиска элементов разметки

Точность поиска (Precision) определяется как доля элементов разметки X , имеющих

сопоставленный элемент в разметке Y : $P = \frac{|D|}{n}$

Полнота поиска (Recall) определяется как доля фрагментов разметки Y , имеющих

сопоставленный фрагмент в разметке X : $R = \frac{|D|}{m}$

Агрегированный критерий (F_1 -мера) определяется как их гармоническое среднее:

$$M_1(X, Y) = \frac{2PR}{P + R} = \frac{2 |D|}{n + m} .$$

Критерий M2: определение мишени манипуляции

Доля сопоставленных элементов разметок X и Y , имеющих одинаковую мишень:

$$M_2(X, Y) = \frac{1}{|D|} \sum_{(i,k) \text{ из } D} [T(x_i) = T(y_k)] .$$

Пример 1. Задача выявления манипуляций

Критерий M3: определение типа манипуляции

Доля сопоставленных элементов разметок X и Y , имеющих одинаковый тип:

$$M_3(X, Y) = \frac{1}{|D|} \sum_{(i,k) \text{ из } D} [C(x_i) = C(y_k)].$$

Критерий M4: точность определения фрагмента манипуляции

Средняя точность совпадения фрагментов в сопоставленных элементах разметок X и Y .

Точность совпадения пары фрагментов (x_i, y_k) вычисляется как мера Жаккара – отношение числа слов в пересечении к числу слов в объединении двух фрагментов:

$$M_4(X, Y) = \frac{1}{|D|} \sum_{(i,k) \text{ из } D} \frac{|x_i \cap y_k|}{|x_i \cup y_k|}.$$

Использование критерия M_4 с нулевым весом $w_4 = 0$ означает, что при измерении точности разметки положение фрагментов игнорируется. Тем не менее, оно всё равно учитывается при поиске соответствий элементов и может влиять на критерий M_1 .

Пример 2. Задача выявления поляризации ОМ

Объектом является множество новостных сообщений, преимущественно относящихся к одному событию или теме.

Разметкой является набор меток $X = \{x_1, \dots, x_n\}$, каждый элемент которого x_i является меткой i -го сообщения.

Метки 1, 2, 3, и т.д. соответствуют кластерам — полюсам общественного мнения по данной теме.

Метка «0» означает, что сообщение является нейтральным, т.е. не пропагандирует никакое из полярных мнений.

Метка «-1» означает, что сообщение не релевантно, т.е. не относится к общей для всех сообщений теме.

Пример 2. Задача выявления поляризации ОМ

Выявление поляризации мнений является задачей кластеризации заданного множества новостных сообщений на кластеры-мнения.

Задача имеет следующие особенности.

1. Нерелевантные сообщения, если таковые имеются, образуют скорее шумовой фон, чем отдельный кластер, т.е. являются разрозненными и не обязательно схожими.
2. Кластер нейтральных сообщений может отсутствовать.
3. Число кластеров-мнений может быть произвольным, включая 0 и 1.
4. Разные разметки одного и того же объекта (множества новостных сообщений) могут содержать различное число кластеров.
5. Нумерация кластеров-мнений может быть произвольной и не обязана совпадать в разных разметках.

Пример 2. Задача выявления поляризации ОМ

Критерий M1: точность и полнота кластеризации мнений

Для сравнения разметки $X = \{x_1, \dots, x_n\}$ с «золотым стандартом» — экспертной разметкой $Y = \{y_1, \dots, y_n\}$, используются VCubed-версии точности и полноты поиска (точность — доля релевантных среди найденных, полнота — доля найденных среди релевантных).

Точность и полнота сначала определяются относительно каждого объекта x_i , затем усредняются по всем объектам с меткой мнения:

$$P = \operatorname{avr}_{x_i > 0} P_i; \quad P_i = \frac{\sum_k [x_k = x_i \text{ и } y_k = y_i]}{\sum_k [x_k = x_i]};$$
$$R = \operatorname{avr}_{y_i > 0} R_i; \quad R_i = \frac{\sum_k [x_k = x_i \text{ и } y_k = y_i]}{\sum_k [y_k = y_i]};$$

(если знаменатель дроби оказывается равным нулю, то считается, что дробь равна нулю).

Агрегированный критерий (F_1 -мера) определяется как гармоническое среднее:

$$M_1(X, Y) = \frac{2PR}{P + R}.$$

Пример 2. Задача выявления поляризации ОМ

Критерий M2: точность и полнота определения нейтральных сообщений

Точность и полнота относительно метки c :

$$P_c = \frac{\sum_k [x_k = y_k = c]}{\sum_k [x_k = c]} ; \quad R_c = \frac{\sum_k [x_k = y_k = c]}{\sum_k [y_k = c]} .$$

Агрегированный критерий (F_1 -мера) определения нейтральных сообщений:

$$M_2(X, Y) = 2P_0R_0 / (P_0 + R_0).$$

Критерий M3: точность и полнота определения нерелевантных сообщений

Агрегированный критерий (F_1 -мера) определения нерелевантных сообщений:

$$M_3(X, Y) = 2P_{-1}R_{-1} / (P_{-1} + R_{-1}).$$

Критерий M4: точность определения числа мнений

Обозначим через K_X и K_Y число различных мнений в разметках X и Y соответственно.

$$M_4(X, Y) = \frac{\min\{K_X, K_Y\}}{\max\{K_X, K_Y\}} .$$

Выводы

1. Предобученные модели внимания / трансформеры позволяют решать всё более трудные задачи NLP / NLU
2. Разметка текстовых данных — магистральный путь формализации знаний в гуманитарных исследованиях
3. Стандартизация разметки: фрагменты + теги + связи + ответы
4. Стандартизация оценивания: согласованность + СТАР/СТЭР

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

voron@mlsa-iai.ru

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