



ИЗВЛЕЧЕНИЕ МНЕНИЙ ИЗ НОВОСТНЫХ ТЕКСТОВ

*Курс «Проектные задачи
компьютерной лингвистики»
Преподаватель Студеникина
Ксения Андреевна*

СОРЕВНОВАНИЕ ПО ИЗВЛЕЧЕНИЮ МНЕНИЙ ИЗ НОВОСТНЫХ ТЕКСТОВ

RuOpinionNE-2024

Структурированный анализ тональности (structured sentiment analysis, SSA) на материале русскоязычных новостных текстов.

Dialogue Evaluation 2024

- Репозиторий на GitHub: <https://github.com/dialogue-evaluation/RuOpinionNE-2024>
- Страница на CodaLab: <https://codalab.lisn.upsaclay.fr/competitions/20244>

ПОСТАНОВКА ЗАДАЧИ

- **Входная информация:** предложение из новостного текста с именованными сущностями.
- **Выходная информация:** извлеченный кортеж мнений.

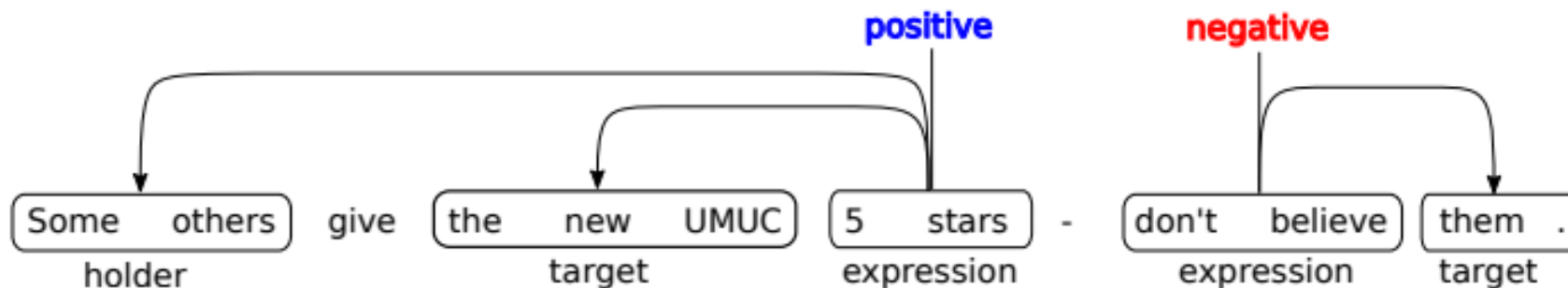
Для заданного текста, **кортеж мнения** представляет собой структуру, состоящую из последовательности четырех значений:

(Н, Т, Р, Е)

- Н (holder) – источник мнения: автор, именованная сущность (подстрока исходного текста);
- Т (target) – объект мнения: именованная сущность (подстрока исходного текста);
- Р (polarity) – тональность: позитивная/негативная;
- Е (expression) – языковое выражение: аргумент, на основании которого принята результирующая тональность (одна или несколько подстрок исходного текста);

ПОСТАНОВКА ЗАДАЧИ

- Предложения могут не содержать ни одного кортежа.
- Источниками и объектами мнения могут быть сущности следующих типов: персоны, организации, страны, города, регионы, профессии, национальности, идеология
- Местоимения в кортежи не включаются.
- Источник мнения также может быть пустым (NULL) для обозначения общего мнения или принимать значение AUTHOR для обозначения мнения автора.



СТРУКТУРИРОВАННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ

- Автоматический анализ тональности остается одной из наиболее востребованных задач в области понимания естественного языка.
- SSA представляет собой извлечение информации об отношениях в наиболее полном и однозначном виде, поскольку учитывает возможность существования разных мнений относительно разных упоминаемых в тексте сущностей.
- Качественный SSA стал возможным лишь относительно недавно благодаря развитию больших языковых моделей.
- Предыдущие работы были в основном посвящены извлечению мнений из отзывов и рецензий (Chen et al 2022; Lin et al. 2022; Morio et al. 2022; Zhang et al. 2022)
- RuOpinionNE-2024 затрагивает более сложный для извлечения мнений домен – новостные тексты.
- Качество SSA по-прежнему остается относительно невысоким, см. Barnes et al. 2022, в связи с чем существует необходимость развития подходов к решению этой задачи.

СОРЕВНОВАНИЕ 10 НА ВОРКШОПЕ SEMEVAL-2022

- Данные включали 5 датасетов на 7 языках (Barnes et al 2022)
- Баскский MultiBooked_{EU}, каталанский MultiBooked_{CA}: коллекция отзывов об отелях на сайте booking.com. Каждый набор данных содержит около 1500 предложений, что делает их самыми маленькими наборами данных в общей задаче. Однако эти предложения содержат больше размеченных кортежей, чем некоторые из более крупных корпусов.
- Английский Opener_{EN}, испанский Opener_{ES}: Opener содержит подборку отзывов об отелях на шести языках (de, en, es, fr, it, nl). Отзывы были собраны с разных сайтов бронирования в период с ноября 2012 по ноябрь 2013 года.
- Английский Multi-Perspective Question Answer corpus (MPQA) содержит новостные тексты со сложной системой разметки. Для выполнения задачи были сохранены только метки *agent*, *target*, *direct-subjective*, *polarity*. Это второй по величине набор данных с большим количеством источников, но относительно небольшим количеством объектов и выражений.
- Английский DS_{Unis} первоначально был опубликован как Darmstadt Service Review Corpus (DSRC). DSRC содержит обзоры онлайн-университетов и сервисов, размеченные по тональности на уровне предложений и выражений. Данные DSUnis содержат только обзоры университетов и не содержат разметки на уровне предложений.
- Норвежский NoReC_{Fine} является подвыборкой Norwegian Review Corpus. Корпус содержит разметку к более чем 11 тысячам предложений, взятых из профессиональных обзоров из различных областей, таких как кино, музыка, литература, товары и игры.

СОРЕВНОВАНИЕ 10 НА ВОРКШОПЕ SEMEVAL-2022

	sentences		holders			targets			expressions			polarity		
	#	avg.	#	avg.	max	#	avg.	max	#	avg.	max	+	neu	–
MultiBooked _{EU}	1,520	10.6	296	1.1	6	1,760	1.4	9	2,319	2.2	10	1,940	0	379
MultiBooked _{CA}	1,676	15.2	237	1.1	7	2,350	2.4	18	2,770	2.6	19	1,743	0	1,027
OpeNER _{EN}	2,492	14.8	413	1.0	3	3,843	1.8	21	4,149	2.4	21	2,981	0	1,168
OpeNER _{ES}	2,054	17.4	225	1.0	2	3,960	2.2	12	4,386	2.2	15	3,557	0	829
MPQA	10,048	23.3	2,265	2.7	40	2,437	6.3	50	2,794	2.0	14	1,082	465	1,059
DS _{Unis}	2,803	20.0	94	1.2	4	1,601	1.2	6	1,082	1.9	9	612	186	805
NoReC _{Fine}	11,437	16.9	1,128	1.0	12	8,923	2.0	35	11,115	5.0	40	7,547	0	3,557

Статистика наборов данных, включая количество предложений и среднюю длину (в токенах), а также среднюю и максимальную длину (в токенах) для источника, цели и выражения. Кроме того, учитывается распределение тональности – положительной, нейтральной и отрицательной – в каждом наборе данных.

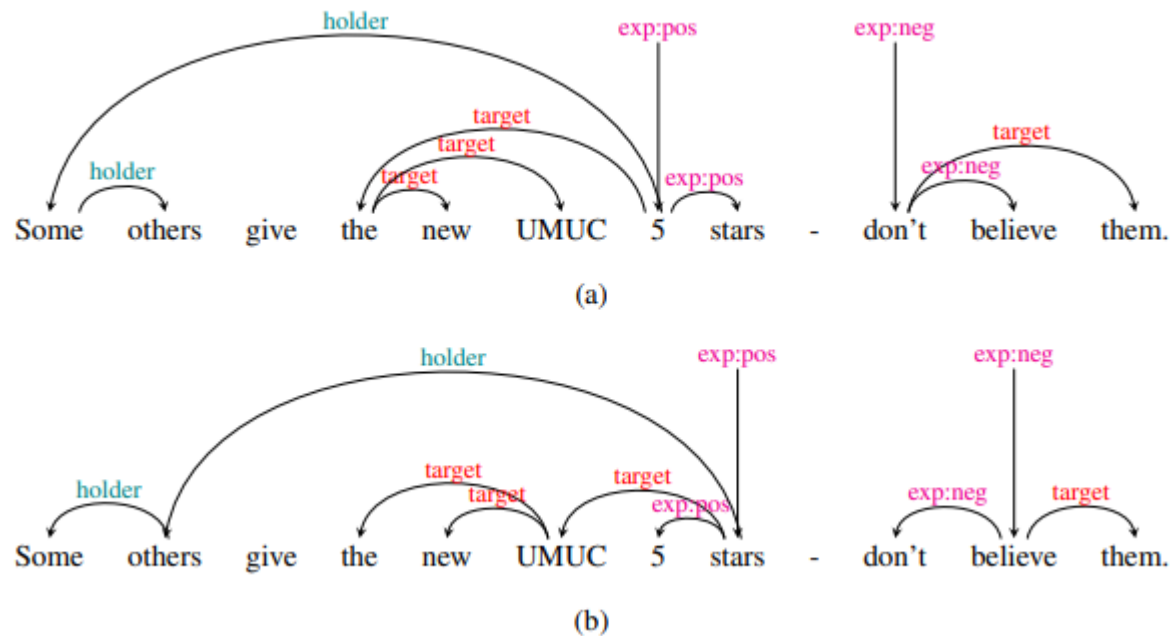
ПОДХОДЫ К СТРУКТУРИРОВАННОМУ АНАЛИЗУ ТОНАЛЬНОСТИ

Участникам соревнования предлагалось два подхода в качестве бейзлайна (Barnes et al 2022):

- Графовый подход (dependency graph)
- Разметка последовательности (sequence labeling)

ГРАФОВЫЙ ПОДХОД (DEPENDENCY GRAPH)

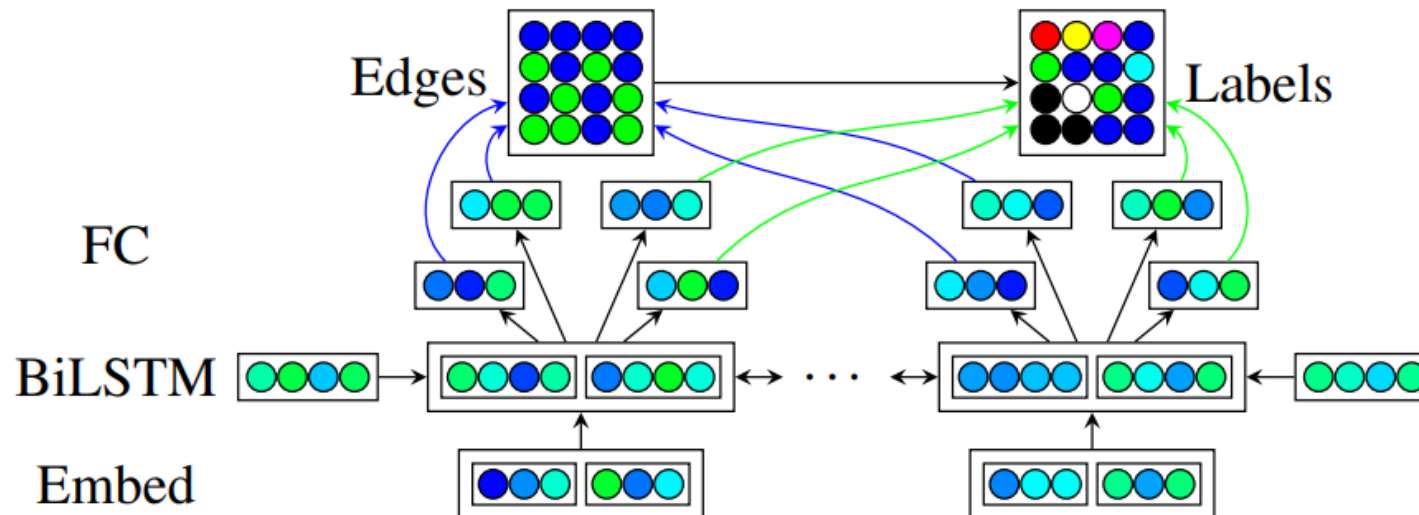
- Задача сводится к разметке направленного графа, который состоит из набора размеченных узлов и набора неразмеченных ребер, соединяющих пары узлов.
- Узлы могут охватывать несколько токенов и могут иметь несколько входящих ребер.



Два варианта парсинга графа:
(a) head-first, где первый токен любого интервала является вершиной,
(b) head-final, где последний токен является вершиной.

ГРАФОВЫЙ ПОДХОД (DEPENDENCY GRAPH)

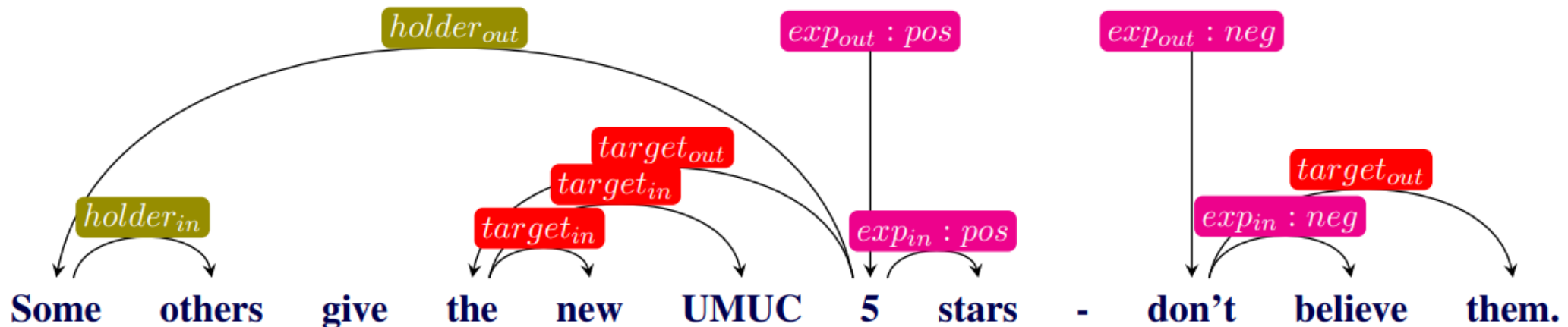
- Для каждого токена строится контекстуализированное векторное представление с помощью BiLSTM или BERT.
- Затем эти эмбединги обрабатываются полносвязными нейронными сетями. Используется отдельный классификатор, предсказывающий наличие ребра (Edges), и классификатор, предсказывающий метку ребра, при его наличии (Labels).



МОДИФИКАЦИЯ ГРАФОВОГО ПОДХОДА

Предлагается отдельно обучать представления для внешних и внутренних связей в графе (Lin et al. 2022)

- Например, между первым словом выражения e , содержащего мнение, и первым словом объекта этого мнения t существует внешняя связь (out-span), чему соответствует ребро ($head_e, dep_t$). Между первым словом e и остальными словами этого выражения существует внутренняя связь (in-span).
- Появляются отдельные метки ребер $target_{out}$ (внешняя связь) и $target_{in}$ (внутренняя связь).
- При представлении данных в таком виде SSA можно рассматривать как задачу классификации ребер графа на N классов, где N – число меток.



ГРАФОВЫЙ ПОДХОД (DEPENDENCY GRAPH)

- Преимущество состоит в относительно высоком качестве предсказания кортежей мнений и небольшом количестве обучаемых параметров по сравнению с большими языковыми моделями.
- Недостаток состоит в требовании достаточного количества обучающих данных и, потенциально, сильной зависимости от содержания этих данных, что может вылиться в плохую переносимость на другие домены (например, с новостей на отзывы о товарах).

РАЗМЕТКА ПОСЛЕДОВАТЕЛЬНОСТИ (SEQUENCE LABELING)

- Сначала обучались 3 BiLSTM для извлечения источников, объектов и выражений.
- Затем обучалась модель для построения отношений между элементами кортежа. Три векторных представления – полного текста, первого элемента (источника или объекта) и выражения – конкатенировались. Модель предсказывала, есть ли между ними отношение.
- Предсказания всех моделей объединяются для формирования полных графов тональности.
- Появление больших языковых моделей делает возможным применение данного подхода без дополнительного обучения.
- Написание подробной инструкции с большим количеством примеров позволяет наиболее точно передать задачу и избежать генерации подстрок, которые отсутствуют в исходном тексте.

МЕТРИКИ ДЛЯ ОЦЕНКИ

- Метрика определяет истинные положительные кортежи с одинаковой тональностью, взвешивая совпадение предсказанных (p – predicted) и истинных (g – gold) интервалов для каждого элемента кортежа: источник (h – holder), объект (t – target), выражение (e – expression). Значение усредняется (делится на 3) и обнуляется, если неправильно предсказана метка тональности POS/NEG (pol – polarity).

$$score(p, g) = \frac{\sum_{e \in \{s, t, h\}} \frac{|p_e \cap g_e|}{p_e}}{3} * 1\{p_{pol} = g_{pol}\}$$

- Значения источника, у которых нет связанных подстрок (NULL и AUTHOR), рассматриваются как отдельные специальные токены.
- Значения объекта и выражения никогда не бывают пустыми.

МЕТРИКИ ДЛЯ ОЦЕНКИ

- Для измерения **точности** количество правильно предсказанных токенов делится на общее количество предсказанных токенов.

$$precision = \frac{\sum_{p \in pred} \max_{g \in gold} score(p, g)}{|pred|}$$

- Для измерения **полноты** количество правильно предсказанных токенов делится на общее количество истинных токенов.

$$recall = \frac{\sum_{g \in gold} \max_{p \in pred} score(p, g)}{|gold|}$$

- Основной метрикой является **F₁-мера** – Sentiment Tuple F₁ / Sentiment Graph F₁.

$$SF_1 = \frac{2 * precision * recall}{precision + recall}$$

RUOPINIONNE-2024: ФОРМАТ ДАННЫХ

- Данные представлены в формате строк json.
- Каждая строка соответствует аннотированному предложению, представленному в виде словаря со следующими ключами и значениями:
- **sent_id**: уникальный числовой идентификатор предложения
- **text**: текст предложения без предобработки
- **opinions**: список всех мнений в предложении
- Каждое мнение (**opinions**) в предложении представляет собой словарь со следующими ключами и значениями:
- **Source**: список из строки и интервала для источника мнения
 - Для источника **NULL** интервал равен ["0:0"];
 - Для источника **AUTHOR** используется специальная метка ["NULL"]
- **Target**: список из строки и интервала для объекта мнения
- **Polar_expression**: список из строк и список из интервалов
- **Polarity**: метка тональности, которая может быть одной из следующих:
 - **NEG** – негативная,
 - **POS** – позитивная

ПРИМЕР С МНЕНИЕМ АВТОРА

```
{  
  "sent_id": 27,  
  "text": "Берлускони ушел в отставку немногим более года назад, после чего в стране с  
согласия парламентского большинства было составлено правительство технократов,  
которому было поручено проводить непопулярную экономическую политику.",  
  "opinions": [  
    {  
      "Source": ["AUTHOR"], ["NULL"],  
      "Target": ["правительство технократов"], ["128:153"],  
      "Polar_expression": ["непопулярную экономическую политику"], ["188:223"],  
      "Polarity": "NEG"  
    }  
  ]  
}
```

ПРИМЕР БЕЗ ИСТОЧНИКА

```
{  
  "sent_id": 1,  
  "text": "Вчера он уволил Азамата Сагитова, который возглавил башкирскую администрацию  
год назад после вынужденной отставки Радия Хабирова, сейчас занимающего пост  
заместителя начальника управления президента РФ по внутренней политике.",  
  "opinions": [  
    {  
      "Source": [["NULL"], ["0:0"]],  
      "Target": [["Азамата Сагитова"], ["16:32"]],  
      "Polar_expression": [["уволит"], ["9:15"]],  
      "Polarity": "NEG"  
    }  
  ]  
}
```

ПРИМЕР С ФРАГМЕНТИРОВАННЫМ ВЫРАЖЕНИЕМ

```
{  
  "sent_id": 58,  
  "text": "Административный суд Кёльна снял с последнего альбома немецкой индастриал-метал группы Rammstein «Liebe ist für alle da» все ограничения на реализацию.",  
  "opinions": [  
    {  
      "Source": [ [ "Административный суд Кёльна" ], [ "0:27" ] ],  
      "Target": [ [ "Rammstein" ], [ "87:96" ] ],  
      "Polar_expression": [ [ "снял", "ограничения на реализацию" ], [ "28:32", "125:150" ] ],  
      "Polarity": "POS"  
    }  
  ]  
}
```

БАЗОВАЯ МОДЕЛЬ

Решение организаторов baseline_model, $F_1 = 0.24$

- Модель Qwen 2.5 32B instruct:

Твоя задача состоит в том, чтобы проанализировать текст и извлечь из него выражения мнений, представленные в виде кортежа мнений, состоящих из 4 основных составляющих:

1. Источник мнения: автор, именованная сущность текста (подстрока исходного текста), либо NULL. Key = Source;
2. Объект мнения: именованная сущность в тексте (подстрока исходного текста). Key = Target;
3. Тональность: положительная/негативная (POS/NEG). Key = Polarity;
4. Языковое выражение: аргумент, на основании которого принята результирующая тональность (одна или несколько подстрок исходного текста). Key = Expression;

Если источник мнения отсутствует, то Source = NULL. Если источником мнения является автор, то Source = AUTHOR. В прочих случаях поле Source должно полностью совпадать с подстрокой исходного текста. Поля Target, Expression всегда совпадают с подстроками текста.

Ответ необходимо представить в виде json списка, каждый элемент которого является кортежем мнений. Каждый кортеж мнений это словарь, состоящий из четырех значений: Source, Target, Polarity, Expression.

БАЗОВАЯ МОДЕЛЬ

Для извлечённых Source, Target, Polarity, Expression должно быть справедливо утверждение: На основании выражения Expression можно сказать, что Source имеет Polarity отношение к Target.

Ниже представлены примеры выполнения задачи:

Текст

Премьер-министр Молдовы осудил террориста за бесчеловечные и жестокие действия.

Source: Премьер-министр Молдовы, Target: террориста, Polarity: NEG, Expression: бесчеловечные и жестокие действия

Текст

Знаменитая актриса продемонстрировала человечность и простоту, достойную уважения публики.

Ответ

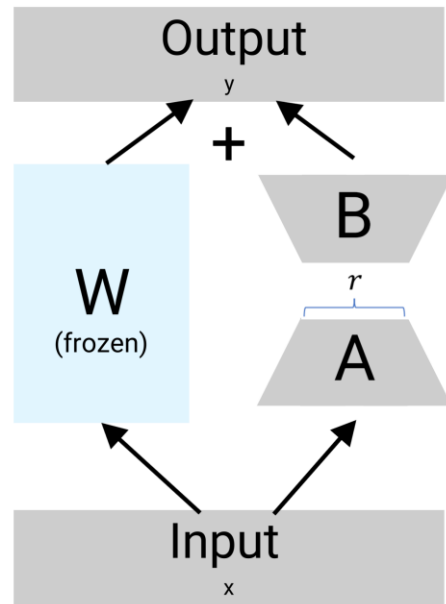
Source: AUTHOR, Target: актриса, Polarity: POS, Expression: продемонстрировала человечность и простоту, достойную уважения публики

Проанализируй таким же образом следующий текст.

Текст

1 МЕСТО В СОРЕВНОВАНИИ

- Решение VatolinAlexey, $F_1 = 0.41$
- Модель LLaMA-3.3-70B*
- Дообучение с помощью метод Low-Rank Adaptation (LoRA) на исходных данных без добавление промта



- Изменить принцип работы слоя W , скорректировав веса на ΔW
- Матрицу ΔW можем представить как произведение двух маленьких матриц A и B

*<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

3 МЕСТО В СОРЕВНОВАНИИ

- Решение msuai, $F_1 = 0.33$
- Модель Mistral Large 2* в режиме few-shot с 15 примерами
- В качестве примеров использовались наиболее близкие по смыслу предложения, сравнивались векторные представления модели ruBERT_{large-uncased}**

Ты эксперт в оценке тональности. Тебе нужно найти все негативные и позитивные отношения между сущностями в тексте и вывести их в следующем формате:

[источник отношения, объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Если источником отношения является автор, то пиши:

['AUTHOR', объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Если выраженного источника нет, то пиши:

['NULL', объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Допустимо вернуть пустой ответ: []

Не нужно давать пояснений к ответу.

Примеры <примеры>

Текст: <целевой текст>

Ответ:

*<https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>

**<https://huggingface.co/DeepPavlov/rubert-base-cased-sentence>

ЛИТЕРАТУРА

- Barnes, Jeremy et al. 2022. “SemEval 2022 Task 10: Structured Sentiment Analysis”. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval2022). Ed. by Guy Emerson et al. Seattle, United States: Association for Computational Linguistics, pp. 1280–1295.
- Chen, Cong et al. 2022. “MT-Speech at SemEval-2022 Task 10: Incorporating Data Augmentation and Auxiliary Task with Cross-Lingual Pretrained Language Model for Structured Sentiment Analysis”. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Ed. by Guy Emerson et al. Seattle, United States: Association for Computational Linguistics, pp. 1329–1335.
- Lin, Yangkun et al. 2022. “Zhixiaobao at semeval-2022 task 10: Approaching structured sentiment with graph parsing”. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 1343–1348.
- Morio, Gaku et al. 2022. “Hitachi at SemEval-2022 task 10: Comparing graph- and Seq2Seqbased models highlights difficulty in structured sentiment analysis”. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 1349–1359.
- Zhang, Qi et al. 2022. A KnowledgeEnhanced Adversarial Model for Cross-lingual Structured Sentiment Analysis. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.