

Сравнение эффективности графового подхода и разметки последовательностей для решения задачи извлечения мнений из новостных текстов

Соловьева Софья

Проектные задачи компьютерной лингвистики
Студеникина Ксения Андреевна

Задача

Что решаем

Извлечение
структурированных мнений
из новостных текстов.

Формат кортежа:

**(Источник, Объект,
Тональность, Выражение)**

Пример

Текст:

*«Административный суд
Кёльна снял ограничения на
реализацию альбома
Rammstein».*

Кортеж:

- **Источник:**
*«Административный суд
Кёльна» (ORGANIZATION).*
- **Объект:** *«Rammstein»
(ORGANIZATION).*
- **Тональность:** *POS
(позитивная).*
- **Выражение:** *«снял
ограничения».*

Актуальность

- Новостные тексты
содержат мнения
с имплицитными
связями (например,
через действия или
контекст).
- Традиционные методы
(например, для отзывов)
плохо работают из-за
сложности новостного
стиля.

Данные и их анализ




1. Корпус:

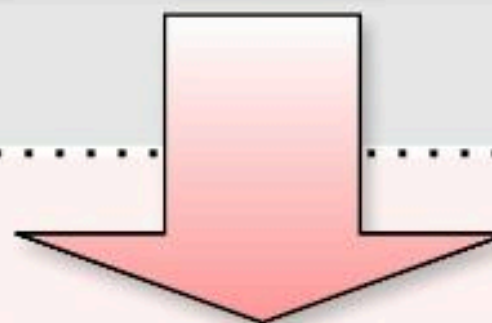
- RuOpinionNE-2024 (структурированные мнения из русскоязычных новостных текстов).
- Источник: комментарии ВКонтакте, новостные статьи.
- Формат данных: JSON с аннотацией кортежей мнений.

RuOpinionNE-2024

Входные данные задачи

(Текст с упомянутыми именованными сущностями)

На выборах Берсани  обошел мэра Флоренции
Маттео Ренци , считающегося одним из
наиболее заметных политиков "новой волны" в Италии 






Выходные данные

(кортежи мнений)

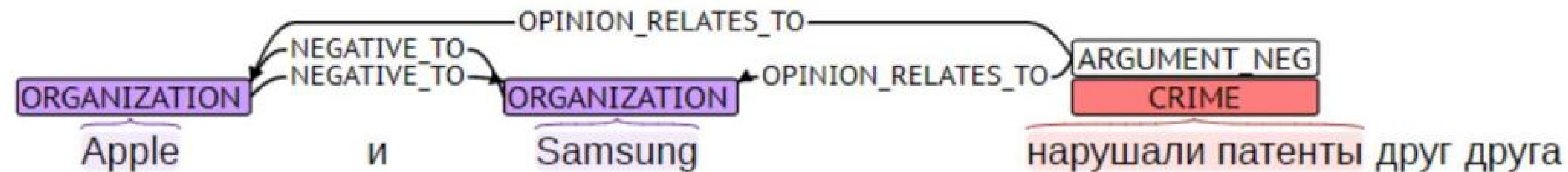
POSITIVE
заметных

Автор 

На выборах Берсани  обошел мэра Флоренции
Маттео Ренци , считающегося одним из
наиболее заметных политиков "новой волны" в Италии 

POSITIVE
обошел мэра Флоренции

Обзор предшествующих исследований



1. Графовый подход (Lin et al., 2022):

- **Суть:** Моделирование зависимостей между элементами мнения через синтаксический анализ.
- **Пример:**
 - **Текст:** «Apple и Samsung нарушали патенты друг друга».
 - **Граф:**
 - Ребро: Apple → Samsung (NEGATIVE_TO).
 - Аргумент: нарушали патенты (ARGUMENT_NEG).
- **Преимущества:**
 - Высокая точность для сложных связей.
 - Учет контекста через синтаксические деревья.
- **Недостатки:**
 - Требуется большого объема размеченных данных.
 - Сложность адаптации к другим языкам/доменам.

Обзор предшествующих исследований

2. Разметка последовательностей (Barnes et al., 2022):

- **Суть:** Последовательное извлечение элементов с помощью LLM и промптинга.
- **Пример:**
 - **Промпт:** «Найди источник, объект, выражение и тональность: [текст]».
 - **Ответ модели:** JSON-кортеж.
- **Преимущества:**
 - Простота реализации (например, через Hugging Face).
 - Эффективность при дисбалансе классов.
- **Недостатки:**
 - Низкая точность для фрагментированных выражений (например, «снял ограничения» + «на реализацию»).

3. Бейзлайны:

- Графовый (baseline_model): $F1 = 0.24$.
- LLM с промптингом (1-е место в RuOpinionNE-2024): $F1 = 0.41$.

Цель, гипотезы и метрики

1. Цель проекта

- **Сравнить эффективность двух подходов для извлечения мнений из новостных текстов:**
 - Графовый метод (синтаксические зависимости).
 - Разметка последовательностей (LLM + промптинг).

2. Гипотезы:

- **Гипотеза 1:**

Графовый подход обеспечит более высокую точность (SF1) за счет учета синтаксических связей между элементами мнения.

Обоснование: Новостные тексты часто содержат имплицитные отношения, которые лучше моделируются через графы.

- **Гипотеза 2:**

Разметка последовательностей будет эффективнее при дисбалансе классов (например, преобладании нейтральных мнений).

Обоснование: LLM могут использовать контекстные подсказки для генерации недостающих элементов.

Цель, гипотезы и метрики

Ключевая метрика: **SF1**

- **Что это?**

Метрика для оценки полноты структуры кортежей мнений:

- Учитывает точность извлечения источника, объекта, выражения и тональности.
- Наказывает за ошибки в тональности или пропуск элементов.

- **Почему SF1?**

- Отражение качества комплексного анализа (не только отдельных элементов, но их связей).
- Используется в соревнованиях (RuOpinionNE-2024, SemEval).

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

Выбор моделей

Для графового подхода

1. Синтаксический парсер:

- spaCy + русская модель (ru_core_news_lg):
 - Быстрая обработка, но ниже точность (UAS: ~78%).

2. Классификация связей:

- BERT-based модели:
 - DeepPavlov/rubert-base-cased для контекстуального анализа.

Для разметки последовательностей

1. Предобученные языковые модели:

- RuRoBERTa-large (ai-forever/ruRoberta-large):
 - 355M параметров.

2. Промптинг LLM:

- Llama-3-70B (через API Hugging Face):
 - Использовать шаблоны из RuOpinionNE-2024 (пример: «Извлеки мнения в JSON»).