NLP Tasks

Natural Language Processing

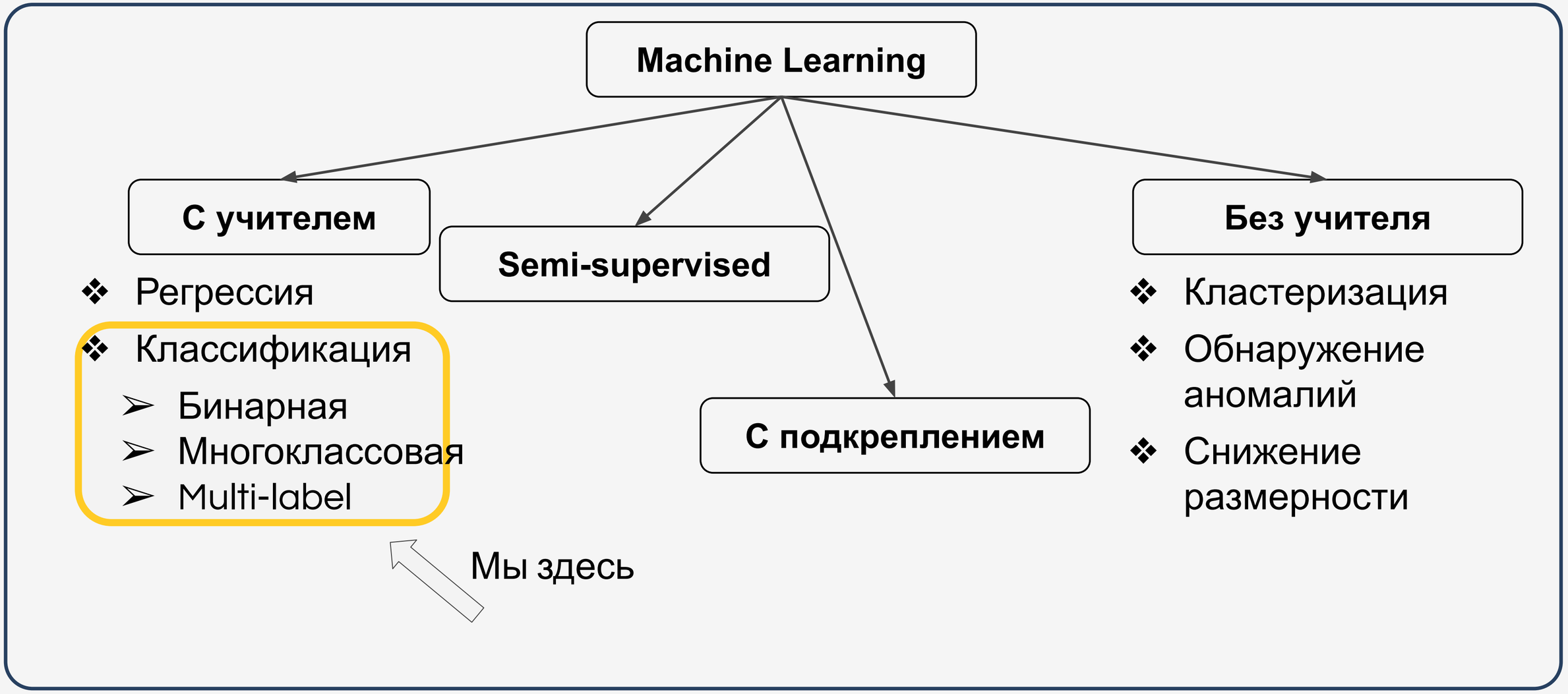
Ника Зыкова, 2025/12/11

План на сегодня

- Классификация задач:
  - По типу входных и выходных данных;
  - По уровню языка;
  - По используемым методам;
- Классические NLP задачи:
  - Постановка;
  - Входные/выходные данные;
  - Метрики;
  - Методы.

https://inixindojogja.co.id/5-tools-ai-llm-open-source-yang-mengubah-cara-manusia-berinteraksi-dengan-teknologi/

ML-задачи



NLP-задачи

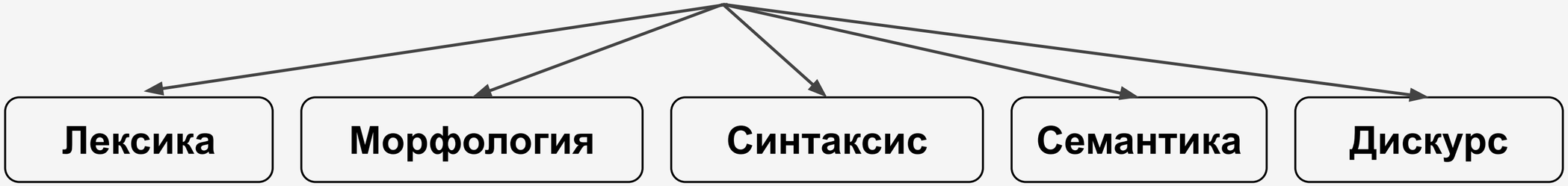
	Token-level	Span-level	Sequence-level
каша	NOUN	Clause 1	Negative
была	VERB		
вкусная	ADJ		
но	CONJ	Clause 2	
оформление	NOUN		
ужасное	ADJ		
вернусь	VERB	Clause 3	
еще	ADV		

NLP-задачи

Seq2Seq | **Auto-regressive** | **Structured prediction**

каша	the porridge	
была	was	если
вкусная	delicious	исправят
но	but	подачу
оформление	the decoration	в
ужасное	was terrible	следующий
вернусь	i will be back	раз
еще	again	

По уровню языка



- Лемматизация
- Части речи
- Морфологические характеристики

- Анализ тональности
- NER
- WSI / WSD

- Кореференция
- Анафора
- Диалоги

Почти все задачи тут

Используемые подходы

- Rule-based: лексиконы, правиловые парсеры,
- Статистичексие методы: n-gram, марковские цепи, CRF;
- Классический ML: статистические вектора + логрег / случайный лес etc, kernel methods;
- Нейросетевые: RNN, LSTM, CNN, word2vec и похожие эмбеддинги;
- Transformer-based: BERT & Co, своя модель на каждую задачу;
- LLMs: одна большая языковая модель с широким набором навыком на множество задач.

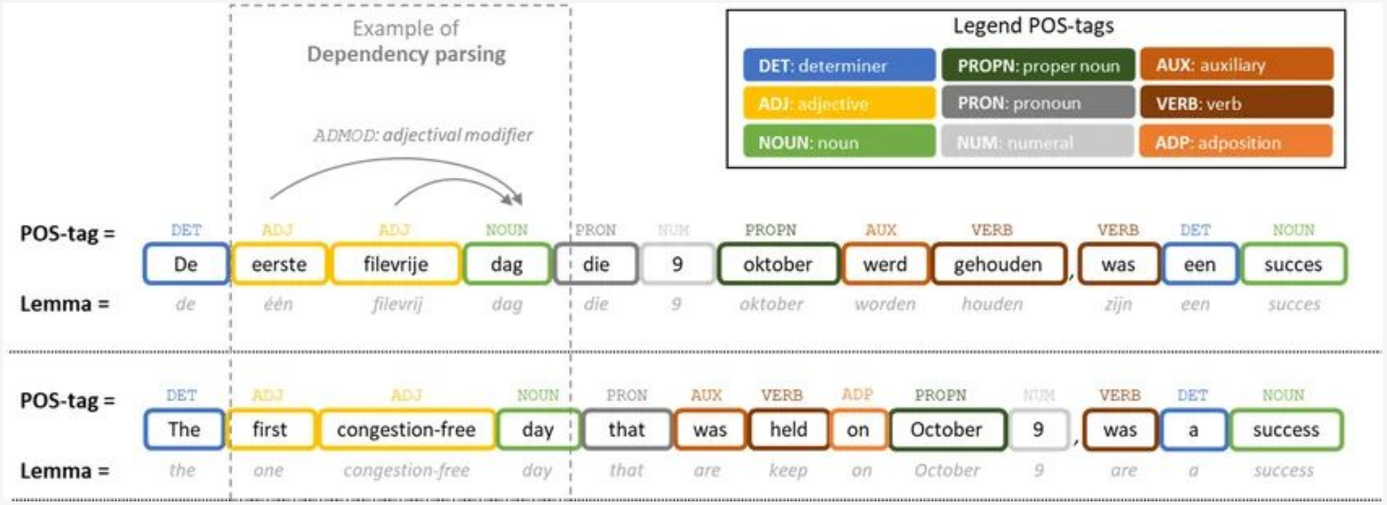
https://inixindojogja.co.id/5-tools-ai-llm-open-source-yang-mengubah-cara-manusia-berinteraksi-dengan-teknologi/

Задачи: морфология

Здесь несколько основных задач:

- POS-tagging;
- Лемматизация;
- Морфологический парсинг;
- Синтаксис (не морфология, но чаще всего вместе с ней).

К каким типам задач они относятся?



https://www.researchgate.net/figure/Example-of-part-of-speech-POS-tagging-and-lemmatization-for-two-example-sentences-The_fig1_337460636

Задачи: семантика

- Word Sense Disambiguation / Induction - контекстное снятие омонимии (мы либо знаем о наборах значений у слов, либо нет);
- (Aspect-Based) Sentiment Analysis: определение эмоциональной окраски (относительно конкретного аспекта / в целом);
- Named Entity Recognition: выделение именованных сущностей;
- Natural Language Inference: определение связаны ли два текста (связаны / противоречат / нейтральны);
- Abstractive / Extractive Question Answering: ответ на вопрос (генеративно / через извлечение релевантного куска текста);
- Машинный перевод;
- Суммаризация;
- Разрешение анафоры / кореференции.

https://www.researchgate.net/figure/Example-of-part-of-speech-POS-tagging-and-lemmatization-for-two-example-sentences-The_fig1_337460636

NER

Почему это сложно?

- Особенности орфографии: с большой буквы пишутся не только ИС;
- Многозначность: одна и та же ИС может имет разный смысл;
- Необходимость учитывать контекст;
- Большой уровень разнообразия;
- Сложность получения чистых и разнообразных данных для обучения;
- Вложенность ИС как отдельная сложность и задача

B - beginning - 1 токен сущности

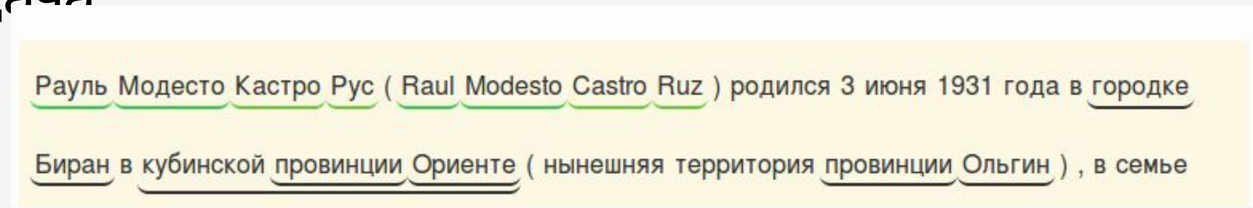
I - inside - не 1/последний токен сущности

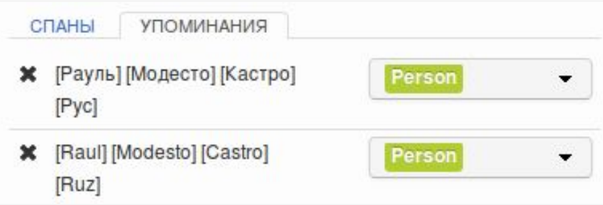
O - out - не сущность

E - ending - последний токен сущности

S - single - сущность из одного токена

Рауль	Модесто	Кастро	Рус	родился	в	городке	Биран	в	1931	году	.
B-PER	I-PER	I-PER	E-PER	OUT	OUT	OUT	S-LOC	OUT	OUT	B-DATE	E-DATE

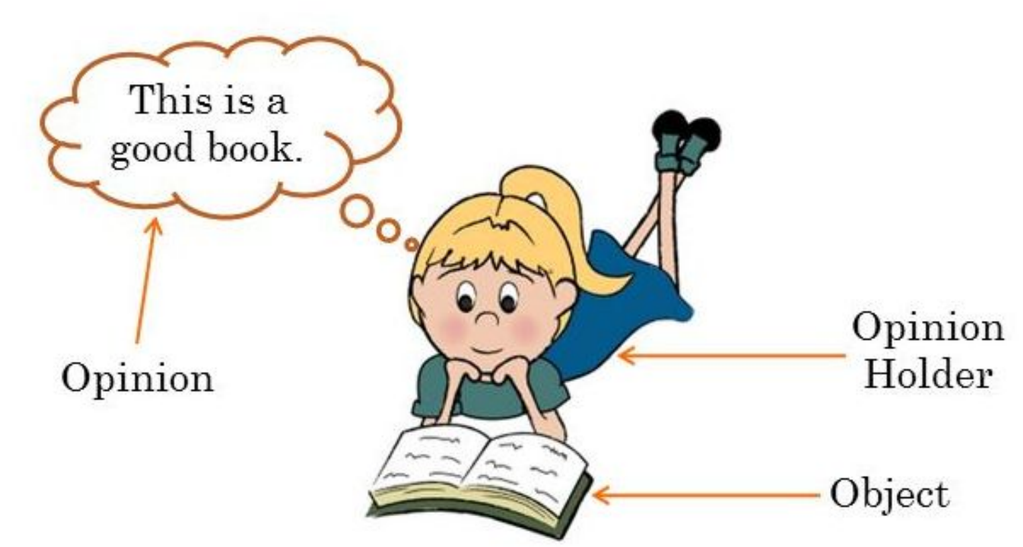


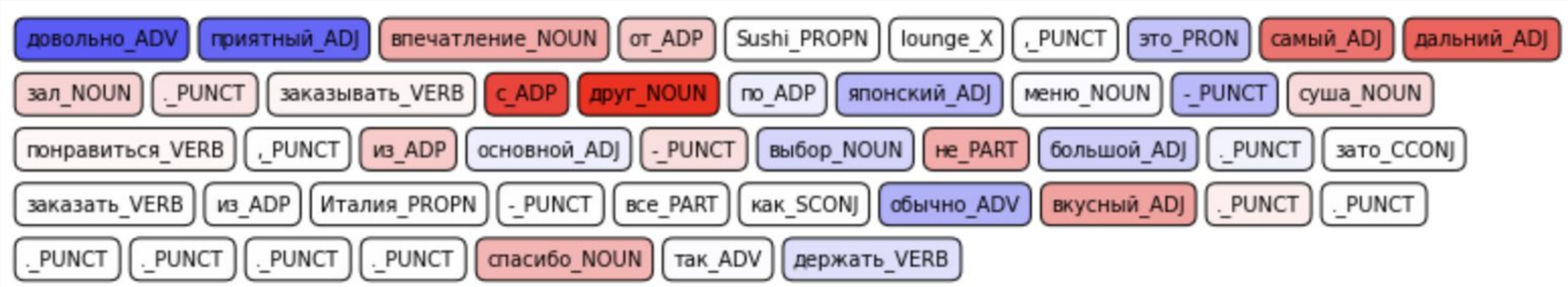


(AB)SA

Три близкие друг к другу задачи:

- Sentiment analysis: выявление / классификация оценочных суждений об объекте
- Opinion mining;
- Subjectivity detection: выявление субъективной информации (факты vs. мнения/эмоции).

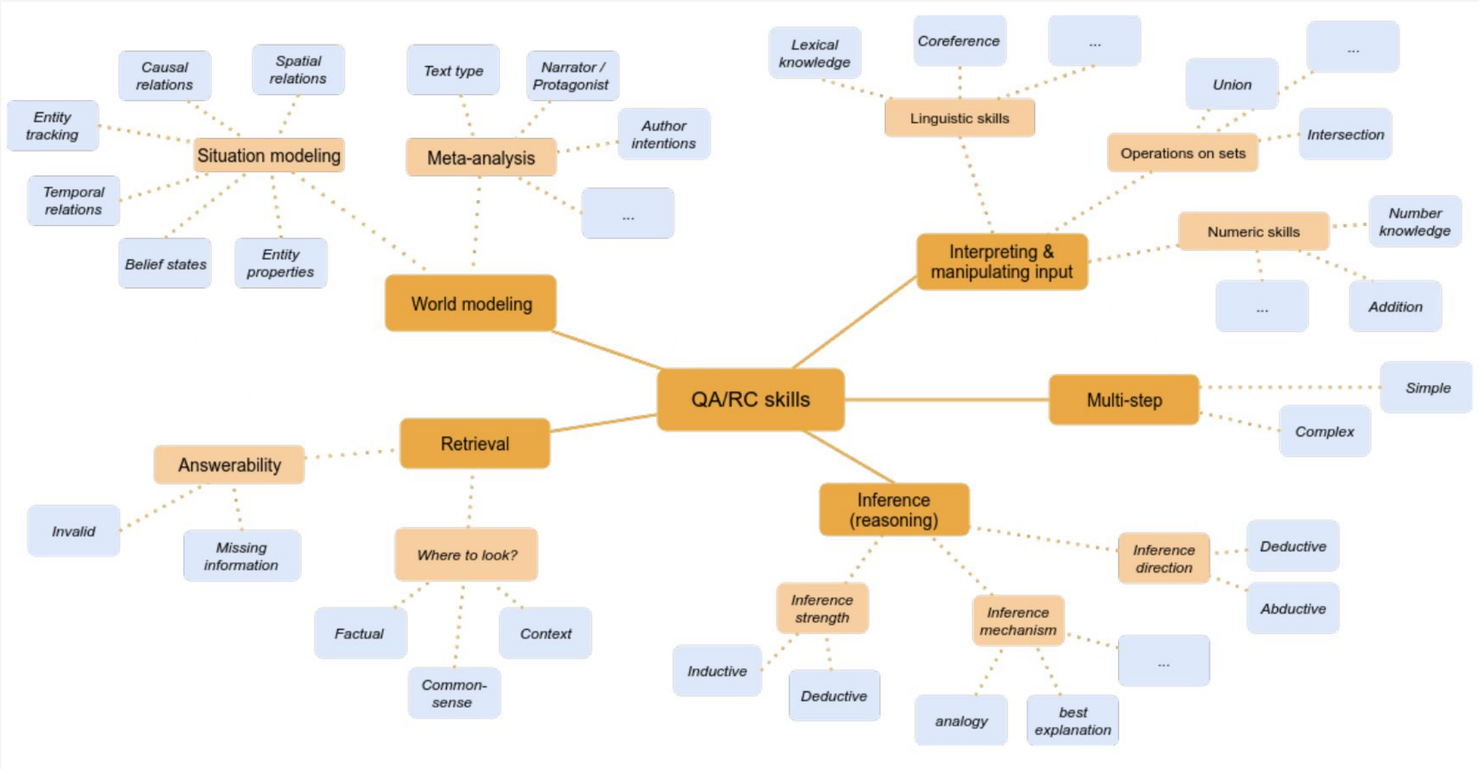




QA

Почему это сложно:

- В вопросе могут присутствовать множественные сущности или связи;
- Время и другие обстоятельства могут быть не выражены эксплицитно;
- Могут требоваться дополнительные рассуждения по данным.



QA: метрики

Зачем свои метрики: правильный ответ не всегда дословно совпадает с эталоном.

Метрики:

- Полное совпадение
- BLEU, NIST (на основе BLEU, но каждой н-грамме присваивается вес в зависимости от встречаемости в языке), ROUGE (считаем не только precision, но и остальные две метрики), METEOR (учитываются не только точные совпадения слов, но и наличие однокоренных слов и синонимов);

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

- BERTScore: используем специального Берта, который обучен оценивать ответы (близко к NLI);
- LLM as a Judge: проверка сторонней, более "сильной" моделью (чаще всего LLM)
- Human feedback (на практике редко реализуемо).

Ключевые слова

Сложности:

- Мы знаем как минимум четыре способа понимать, что такое ключевые слова
- Разные люди по-разному выделят слова даже в рамках одного определения
- Нужно понимание предметной области для разметки
- Может ли быть текст из одних ключевых слов? Вообще без ключевых слов? Какая плотность должна быть у разметки?