Распознавание галлюцинаций языковых моделей

Ксения Евгеньевна Петрушина

Московский физико-технический институт,
Сколковский институт науки и технологий

Научный руководитель: А. И. Панченко

2023

Цель исследования

Цель

Распознавание галлюцинаций языковых моделей в задачах машинного перевода, перефразирования и определения значения слова.

Задача

Исследовать методы распознавания галлюцинаций в сгенерированном тексте и провести сравнительный анализ методов.

Постановка задачи

Дана выборка

$$\mathcal{D} = \{(\text{src}_i, \text{hyp}_i, \text{label}_i)\}_{i=1}^{N},$$

где src_i – входные данные языковой модели, hyp_i – ответ языковой модели, $\text{label}_i \in \{0, 1\}$ – целевая переменная, обозначающая является ли ответ модели галлюцинацией. Необходимо построить модель классификации

$$f(\text{src}_i, \text{hyp}_i) = p_i \in [0, 1]$$

Метрикой качества является

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} (\mathbb{I}[f(\text{src}_i, \text{hyp}_i) \geq thr] = \text{label}_i)$$

Список литературы

1. Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks.
   URL:
   https://dl.acm.org/doi/pdf/10.5555/3495724.3496517
2. David Dale et al. 2023. HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation.
   URL: https://arxiv.org/pdf/2305.11746.pdf
3. Albert Q. Jiang et al. 2023. Mistral 7B.
   URL: https://arxiv.org/pdf/2310.06825.pdf

Результаты экспериментов

Method	Accuracy
Sentence transformer	**0.696**
BERTScore f1	0.656
Mistral-7B	0.640

Таблица: Paraphrase generation task

Method	Accuracy
LaBSE-en-ru	0.786
BLASER 2.0-QE	**0.802**
Mistral-7B	0.684

Таблица: Machine translation task

Method	Accuracy
Token-RAG + BERTScore f1	0.598
Wiktionary definition + E5	**0.654**

Таблица: Definition modeling task

Заключение

- Были рассмотрены различные подходы к задаче распознавания галлюцинаций языковых моделей
- Основной фокус исследования направлен на измерение похожести предложений