Задача детекции границы между человеческим и машинно-генерируемым контентом

Курсовая работа

Студент: Пшеничников Глеб Викторович, 317 группа
Научный руководитель: Грабовой Андрей Валериевич

Актуальность задачи

Развитие больших языковых моделей (LLM) привело к большому количеству искусственно созданных текстов, мало отличимых от написанных человеком. В связи с чем увеличивается значимость проблем:

- Стремительный рост академического плагиата
- Дезинформация пользователей, использующих языковые модели для написания статей, образовательных работ, так как сгенерированные текста могут содержать неточную, ложную информацию

В отдельную категорию задач, требующих решение, можно вынести анализ гибридных текстов, созданных человеком и искусственным интеллектом совместно. Эта задача занимает важное место, так как такие тексты отражают реальные сценарии использования генеративных моделей для дополнения, расширения статей, научных работ, образовательных текстов.

Постановка задачи

Основная задача:

- **Дано:** набор текстов $T = \{t^j\}_{j=1}^{M}$, где каждый текст $t^j = \{s_1^j, s_2^j, \ldots, s_{m_j}^j\}$ — последовательность предложений.
- **Найти:** номера i^j для $j = 1, \ldots, M$, такие что:
  - $\{s_1^j, \ldots, s_{i^j-1}^j\}$ — человеческий текст,
  - $\{s_{i^j}^j, \ldots, s_{m_j}^j\}$ — машинно-генерированный текст.
- **Критерий:** минимизация функции потерь Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{M} \sum_{j=1}^{M} |i_{\text{true}}^j - i_{\text{pred}}^j|$$

где i_{true}^j — истинная граница, i_{pred}^j — предсказанная граница.

Ключевые подзадачи:

- Подтвердить гипотезу о том, что перплексия резко меняется при смене типа текста (человек $\rightarrow$ ИИ).
- Создать статистические словари n-грамм (n=1,2,3)

Обзор существующих решений

Существующие методы:

- **DetectGPT** — использует кривизну логарифмических вероятностей
- **Binoculars** — использует "бинокулярное расхождение"между двумя разными LLM
- **DNA-GPT** — анализ расходящихся N-грамм

Недостатки:

- Требуют значительных вычислительных ресурсов
- Разнообразие способов генерации у языковых моделей растет

Преимущества предлагаемого метода:

- Возможность детекции фрагментов, созданных разными LLM
- Способен определять начало сгенерированного фрагмента даже у закрытых генеративных моделей с недоступными весами

О данных

Тренировочные данные были взяты с соревнования *SemEval-2024 Task 8 (подзадача C)* и *PAN 2025*.

Тестирование проводилось на выборках с сорвенований *SemEval-2024 Task 8*, *DAGPap24*, *COLING 2025*, *PAN 2024*, *PAN 2025*.

Предобработка данных:

- *SemEval-2024 Task 8*: преобразование меток токенов в номера предложений
- *DAGPap24*: выделены все фрагменты нужного формата и определены метки предложений перехода
- Остальные выборки: конкатенация человеческих и созданных при помощи LLM текстов и вычисление номера предложений перехода

Полученные выборки содержали текста длиной от 10-и до 60-и предложений.

Для создания **словарей вероятностей** использовались три генеративные модели: GPT-2, BART, BLOOM, OPT

Методология

1. **Создание словаря вероятностей**
2. **Вычисление перплексии предложений текстов**

$$PP_n(S) = \exp\left(-\frac{1}{N-n+1}\sum_{i=n}^{N}\log P(w_i|w_{i-n+1},\ldots,w_{i-1})\right)$$

где $S = \{w_1, w_2, \ldots, w_N\}$ — предложение, последовательность токенов

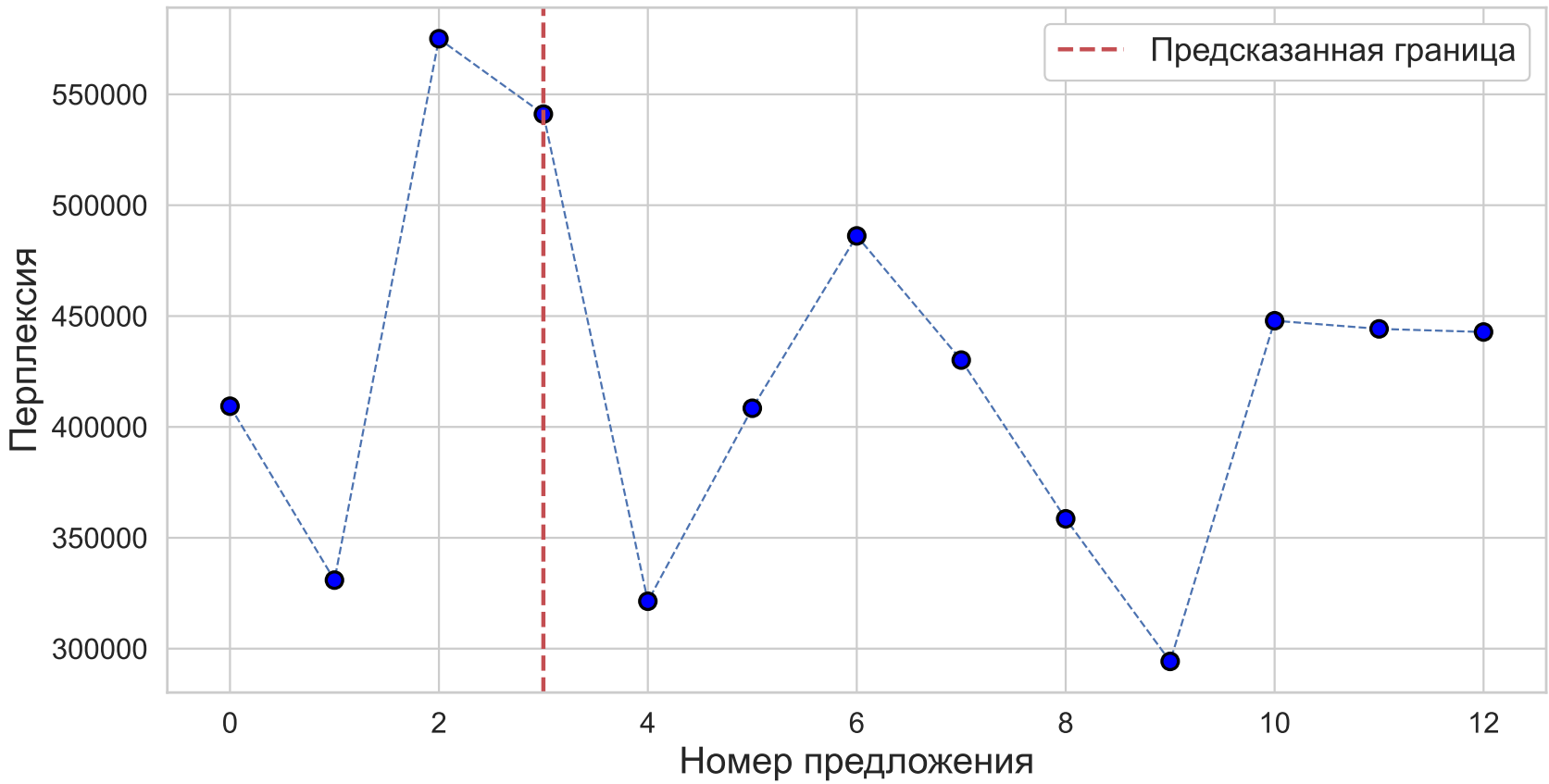
3. **Классификация каждого предложения**
4. **Определение оптимальной границы**

$$G_{weighted}(i) = \frac{|S_{left}|}{|S|}\cdot G(S_{left}) + \frac{|S_{right}|}{|S|}\cdot G(S_{right}) \to \min_i$$

где $|S|$ — общее количество предложений,
S_{left} (S_{right}) — набор предложений до (после) границы i,
$G(S_{\text{left}})$ и $G(S_{\text{right}})$ — индексы Джини для левой и правой частей соответственно.

Эксперименты

Рассмотрим, как в гибридных текстах изменяется перплексия на примере предложения, для которого метод точно предсказал границу:



Вывод: гипотеза подтверждается, по резкому уменьшению перплексии можно судить о границу перехода между частями текста.

Результаты тестирования

SemEval-2024 Task 8 (Subtask C):

- Средняя длина текста – 12 предложений
- Средняя ошибка: 1.03 предложения
- Классификация: $Precision = 91.1\%$, $Recall = 93.2\%$

DAGPap24:

- Средняя длина текста – 59 предложений
- Средняя ошибка: 10 предложений
- Классификация: $Precision = 82.7\%$, $Recall = 99.4\%$

COLING 2025:

- Средняя длина текста – 31 предложений
- Средняя ошибка: 5.5 предложений
- Классификация: $Precision = 82.1\%$, $Recall = 93.6\%$

PAN 2025:

- Средняя длина текста – 39 предложений
- Средняя ошибка: 10 предложений
- Классификация: $Precision = 75.9\%$, $Recall = 75.8\%$

Результаты для PAN 2024

Рассмотрим величину ошибки модели в предложениях при анализе гибридных текстов, созданных при помощи разных LLM.

Модель	**Средняя длина**	**Средняя ошибка**	**Полнота (%)**	**Точность (%)**
Text-Bison-002	45.7	9	70	80.4
Alpaca-7B	31.7	10.9	94.8	65.5
GPT2	**43.4**	**9.9**	**80.7**	**77.2**
Mistral-7B	48.7	10.7	70.9	78.1
Llama-2-7B	41.7	9	75.2	78.5
GPT-3.5	**38.5**	**7.1**	**69.7**	**81.4**
GPT-4	**44**	**8.7**	**68.5**	**80.1**
BLOOMZ-7B1	38.9	11.3	89	70.8
Qwen1.5-72B	40.5	8.5	73	78.9
Llama-2-70B	45.9	9.7	72.8	78.8
Mixtral-8x7B	46.5	9.8	70.5	78.8

Вывод: метод показывает хорошее качества для разных LLM

Заключение

Основные достижения:

- Реализована модель, определяющая номер предложения, в котором текст, написанный человеком, сменяется текстом, сгенерированным искусственным интеллектом
- Метод основан на анализе перплексии текста с использованием статистических словарей
- Подтверждена гипотеза, что перплексия напрямую зависит от типа написания текста

Конкретные результаты:

- В текстах, содержащих не более 12-и предложений, алгоритм ошибается не более, чем на 1 предложение; содержащих не более 60-и предложений – не более чем на 10 предложений
- Классификатор достигает в среднем 80% точности и 80% полноты в обнаружении сгенерированного предложения

Результат, выносимый на защиту

Предложен: метод определения границы перехода между человеческим и машинно-генерированным текстом на уровне предложений, основанный на анализе перплексии через статистические словари вероятностей n-грамм, бинарной классификации предложений и оптимизации индекса Джини

Обеспечивающий:

- Высокую точность для текстов длиной до 60 предложений
- Универсальность с разными LLM (работает для 13 моделей на словарях от 4 генераторов)

Ключевые особенности:

- Использование статистических словарей вероятностей n-грамм (без доступа к весам моделей)
- Комбинация перплексии, классификации предложений (Random Forest) и оптимизации индекса Джини
- Подтверждение гипотезы о резком изменении перплексии на границе перехода

Литература

Основные работы по перплексии и детекции машинного текста:

- G. M. Gritsaia, I. A. Khabutdinova, and A. V. Grabovoya, *Stack More LLM's: Efficient Detection of Machine-Generated Texts via Perplexity Approximation*, arXiv preprint (2023).
- P. Dmitrii, L. Mestetsky, and A. Grabovoy, *N-Gram Perplexity-Based AI-Generated Text Detection*, arXiv preprint (2024).
- G. M. Gritsay, A. V. Grabovoy, A. S. Kildyakov et al., *Artificially generated text fragments search in academic documents*, Dokl. Math. 108, S434–S442 (2023).
- G. Gritsai, I. Khabutdinov, and A. Grabovoy, *Multihead span-based detector for AI-generated fragments in scientific papers*, Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP), 2024.
- Ch. Vasilatos, M. Alam, T. Rahwan et al., *HowkGPT: Investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis*, arXiv:2305.18226 (2023).

СПАСИБО
ЗА ВНИМАНИЕ!