

Верификация искусственно сгенерированных текстовых фрагментов

Г. М. Грицай

Научный руководитель: к. ф.-м. н. А.В. Грабовой

Кафедра интеллектуальных систем ФПМИ МФТИ

Специализация: Интеллектуальный анализ данных

Направление: 09.04.01 Информатика и вычислительная техника

2025

Поиск сгенерированных текстовых фрагментов

Исследуется проблема верификации текстовых последовательностей.

Цель исследования —

построение методов поиска, верификации и интерпретации сгенерированных текстовых последовательностей.

Требуется предложить

Метод детектирования машинно-сгенерированных текстовых последовательностей, основанный на паттернах присущих искусственно созданным фрагментам, а также метод их интерпретации и обоснования.

Метод решения

Предлагаемый метод основан на контроле длины входной последовательности, множественном тестировании сегментов исходного текста, классификации и мультизадачной регуляризации.

Задача классификации текстовых последовательностей

Пусть задан \mathbf{W} — алфавит и множество документов:

$$\mathbb{D} = \{[t_j]_{j=1}^n \mid t_j \in \mathbf{W}, n \in \mathbb{N}\}.$$

Задана выборка из N документов:

$$\mathbf{D} = \bigcup_{i=1}^N D^i, D^i \in \mathbb{D}.$$

1. Детекция автора всего документа:

$$\phi : \mathbb{D} \rightarrow \mathbf{C},$$

где $\mathbf{C} = \{0, 1\}$ для бинарной постановки или $\mathbf{C} = \{0, \dots, k - 1\}$ для многоклассовой детекции и k языковых моделей-авторов.

Задача детекции фрагментов в текстовых последовательностях

2. Детекция фрагментов с генерацией:

Задано множество непересекающихся фрагментов документа:

$$\mathbf{T}^* = \{[t_{s_j}, t_{f_j}]_{j=1}^J \mid t_{s_j} = t_{f_{j-1}}, \quad s_j \in \mathbb{N}_0, \quad f_j \in \mathbb{N}\},$$

где t_{s_j} и t_{f_j} — стартовый и завершающий индекс j -ого фрагмента, J — количество фрагментов документа.

Представим модель в виде суперпозиции двух преобразований:

$$\phi = \mathbf{g} \circ \mathbf{f},$$

$$\mathbf{f} : \mathbb{D} \rightarrow \mathbf{T}^*, \quad \mathbf{g} : \mathbf{T}^* \rightarrow \mathbf{C},$$

$$\phi : \mathbb{D} \rightarrow \mathbf{T}, \quad \mathbf{T} = \{[t_{s_j}, t_{f_j}, c_j]_{j=1}^J \mid t_{s_j} = t_{f_{j-1}}, \quad s_j \in \mathbb{N}_0, \quad f_j \in \mathbb{N}, \quad c_j \in \mathbf{C}\},$$

где \mathbf{f} — разделитель текста на непересекающиеся фрагменты, \mathbf{g} — бинарная классификация каждого текстового фрагмента.

Бинарная классификация фрагментов

Минимизируем эмпирический риск в наборе данных \mathbf{D} :

$$\hat{g} = \operatorname{argmin}_{g \in \mathfrak{F}} \sum_{D^i \in \mathbf{D}} \sum_{x_j, c_j \in D^i} [g(t(x_j)) \neq c_j], \quad t: \mathbf{T}^* \rightarrow (V)^n,$$

где x_j фрагмент документа D^i , t - токенизатор, V - словарь всевозможных токенов предобученной модели, n - фикс. длина входного вектора, а \mathfrak{F} набор всех рассмотренных алгоритмов для классификации.

Функция потерь задачи классификации:

$$\mathcal{L}_{\text{BCE}}(g, \mathbf{D}) = -\frac{1}{|\mathbf{D}|} \sum_{D^i \in \mathbf{D}} \sum_{(x_j, c_j) \in D^i} [c_j \cdot \log(\hat{g}(t(x_j))) + (1 - c_j) \cdot \log(1 - \hat{g}(t(x_j)))],$$

Отслеживаемые метрики качества: *precision*, *recall*, *F₁-score*.

Постановка подхода мультизадачного обучения

Пусть M задачам классификации соответствует множество датасетов $\mathbb{D} = \{d_1, d_2, \dots, d_M\}$. Модель мультизадачного обучения (MTL) с сильным совместным использованием параметров (HPS) состоит из общей подсети h_{θ_s} с параметрами θ_s и T специфичных сетей под конкретную задачу $g_{\theta_1}, \dots, g_{\theta_T}$ с параметрами $\{\theta_i\}$, все параметры MTL: $\theta = \theta_s \cup \bigcup_{i \in [T]} \theta_i$. Обозначим L_1, L_2, \dots, L_T функции потерь каждой задачи. В подходе с MTL будем оптимизировать:

$$\mathcal{L}(\theta) = \sum_{x_j \in \mathbb{D}} \sum_{t \in [T]} L_t(g_{\theta_t} \circ h_{\theta_s}(x_j), c_t).$$

Определение. Эмпирическая сложность Радемахера.

Пусть $G := \{g : Z \rightarrow \mathbb{R}\}$ — класс функций, а $S := \{z_1, \dots, z_n\}$ — выборка из распределения P , тогда эмпирическая сложность Радемахера класса G определяется как:

$$\hat{\mathfrak{R}}_G(n) := \mathbb{E}_{\sigma} \left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right],$$

где σ_i — независимые случайные величины, равномерно распределённые на $\{\pm 1\}$.

Сложность мультизадачного и однозадачного подходов

Рассмотрим функциональные классы для трансформер-модели в задаче классификации:

$$\mathcal{F}_{\text{STL}} = \{x \mapsto w_{\text{head}}^\top \phi(x; w_{\text{enc}}) \mid w_{\text{enc}} \in \mathcal{W}_{\text{enc}}, w_{\text{head}} \in \mathcal{W}_{\text{head}}\},$$

$$\mathcal{F}_{\text{MTL}} = \left\{ (x \mapsto w_t^\top \phi(x; w_{\text{shared}}))_{t=1}^T \mid w_{\text{shared}} \in \mathcal{W}_{\text{shared}}, w_t \in \mathcal{W}_{\text{head}} \right\}.$$

Теорема (Грицай, 2025). Пусть для решения задачи классификации $f \in \mathcal{F}_{\text{STL}}$ и $g \in \mathcal{F}_{\text{MTL}}$, где MTL охватывает T задач, $T - 1$ из которых связаны с целевой. Объем выборки: nT для STL и n на задачу для MTL. Дополнительные ограничения:

$$\|w_{\text{enc}}\| \leq B_{\text{enc}}, \quad \|w_{\text{shared}}\| \leq B_{\text{shared}} \leq \frac{B_{\text{enc}}}{\sqrt{T}}, \quad \|w_t\| \leq B_{\text{head}}, \quad \forall t \in [T],$$

$$\|\phi(x; w)\|_2 \leq L \cdot \|w\| \cdot \|x\|_2, \quad \|x\|_2 \leq R \quad \forall x \in \mathcal{X}.$$

Тогда для каждой задачи t в MTL выполняется:

$$\hat{\mathfrak{R}}_{\text{MTL}}^{(1)}(n) \leq \hat{\mathfrak{R}}_{\text{STL}}^{(1)}(nT).$$

Если $B_{\text{shared}} < \frac{B_{\text{enc}}}{\sqrt{T}}$, то неравенство становится строгим.

Проблемы множественных сравнений

Ранее был получен классификатор \hat{g} , минимизирующий эмпирический риск.

Проверка гипотез:

$$H_0 : \hat{g}(\text{fragment}) = 0,$$

$$H_1 : \hat{g}(\text{fragment}) = 1.$$

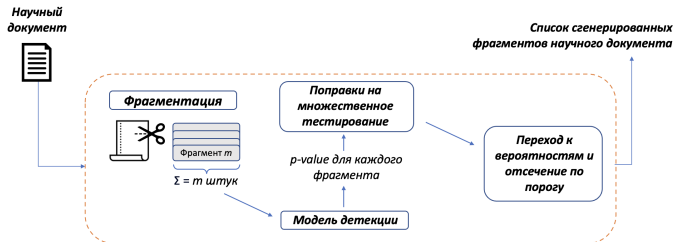
Оценка вероятности того, что хотя бы один из них будет неверным и контроль ошибок:

$$P(\text{false positive}) = 1 - (1 - \alpha)^m, \quad FWER = P(V > 0), \quad FDR = \mathbb{E}\left(\frac{V}{V + S}\right),$$

где V — число ложно положительных результатов, а S — число истинно положительных.

В текущей задаче используется метод контроля групповой вероятности ошибки:

$$p\text{-value} = 1 - \hat{g}(t(x_j))$$



Бинарная классификация на основе оценки перплексии

Документ задан последовательностью токенов $D^i = [t_j]_{j=1}^{|D^i|}$, где $t_j \in \mathbf{W}$, а $|D^i|$ — количество токенов в документе D^i .

$$PPL(D^i) = \exp \left(-\frac{1}{|D^i|} \sum_{j=1}^{|D^i|} \log P(t_j \mid t_1, t_2, \dots, t_{j-1}) \right),$$

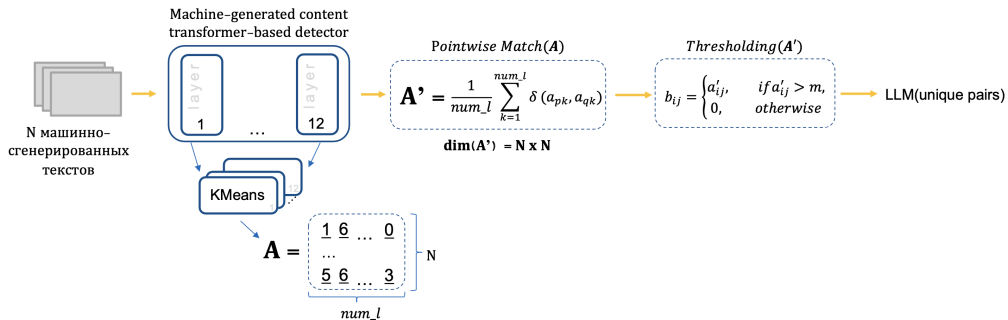
Гипотеза. Значение перплексии LLM может быть аппроксимировано статистической языковой моделью с помощью словаря N-грамм, составленному по выходам данной большой языковой модели.

$$PPL_{\text{approx}}(D^i) = \exp \left(-\frac{1}{K} \sum_{j=1}^K \log P(t_{\text{after n-gram}} \mid \text{n-gram}_j) \right),$$

где суммирование производится по количеству N-грамм входного текста, общее количество суммирований обозначается K .



Интерпретация сгенерированных текстовых фрагментов

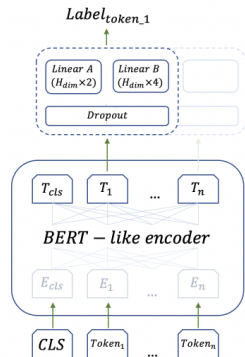
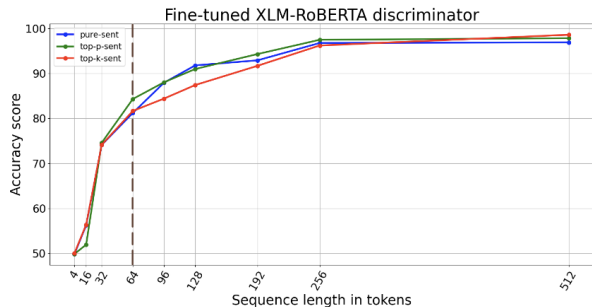


Архитектура настройки подхода, выделяющего пары текстов с совпадающими паттернами.

Этап проверки:

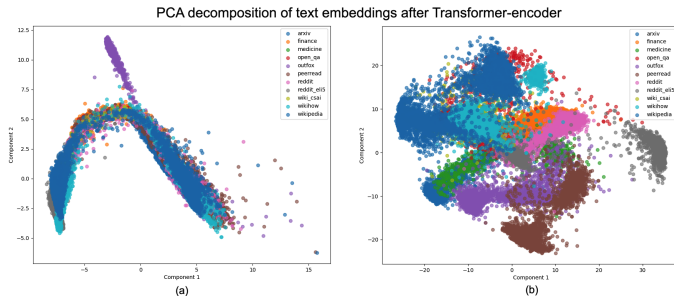
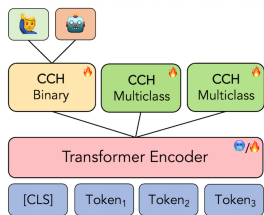
$$x_j \xrightarrow{\text{KMeans}(a_{\text{new},l})} \mathbf{c}_{\text{new}} \xrightarrow{s_j = \frac{1}{L} \sum_{l=1}^L \delta(c_{\text{new},l}, c_{j,l})} \{\mathbf{c}_j \in \text{Индекс} \mid s_j > m\} \xrightarrow{\text{LLM-признаки}} \text{Рел. признаки.}$$

Вычислительный эксперимент: архитектура



На рисунке (а) точность классификатора основанного на архитектуре трансформер возрастает с увеличением длины последовательностей, на (b) архитектура подхода детекции фрагментов с варьируемой длиной.

Вычислительный эксперимент: мультизадачное обучение



Архитектура MTL и разложение по двум главным компонентам текстов на основе векторного представления. На рисунке (a) структура векторного пространства для модели deberta-v3-base, настроенной в однозадачном режиме, на рисунке (b) - та же модель, но настроенная в режиме MTL.

Результаты вычислительного эксперимента

Язык	Эксперимент	F1-score	Precision	Recall
ru	базовое решение	0.955	0.958	0.955
	мультиязычное обучение	0.964	0.964	0.966
	перевод текстов 50%	0.966	0.968	0.966
	парафраз предложений 100%	0.968	0.970	0.968
en	базовое решение	0.796	0.855	0.802
	мультиязычное обучение	0.823	0.867	0.828
	перевод текстов 50%	0.825	0.868	0.830
	парафраз предложений 100%	0.822	0.866	0.827

Эксперимент с детекцией фрагментов фиксированной длины.

Модель	F1-score
TF-IDF + LogReg	60.93
DeBERTa v3 base	78.52
MTL	83.07

Эксперимент с детекцией при помощи мультизадачного обучения.

Модель	F1-score
DistilBERT	0.84
Mistral w. QLoRA	0.91
XLNet	0.95
SciBERT	0.96

Эксперимент с детекцией фрагментов варьируемой длины.

Модель	F1-score	Время (с)
TD-IDF	0.90	0.36
DetectGPT	0.37	471
Binoculars	0.92	236
KenLM + ARPA	0.91	0.27

Эксперимент с детекцией при помощи статистических языковых моделей.

Выносятся на защиту

1. Предложены методы поиска и детектирования машинно-сгенерированных фрагментов в текстовых последовательностях, основанные на фиксированной и варьируемой фрагментации, множественном тестировании и классификации сегментов.
2. Выявлена зависимость качества классификации от длины входной последовательности в моделях классификации с архитектурой трансформер.
3. Показано, что мультизадачное обучение повышает обобщающую способность модели, формирует кластерную структуру и улучшает заданные метрики качества бинарных задач.
4. Описан подход формирования обоснований срабатывания модели детекции на основе фигурирующих паттернов текста.
5. Проведена серия вычислительных экспериментов, которые показывают применимость предложенных методов.

Список работ по теме диссертации

Публикации по итогам конференций, индексируемые в международных базах данных

1. Gritsay G., Grabovoy A., Chekhovich Y. Automatic Detection of Machine Generated Texts: Need More Tokens // 2022 Ivannikov Memorial Workshop (IVMEM). – IEEE, 2022.
2. Gritsay G., Grabovoy A., et al Automated Text Identification: Multilingual Transformer-based Models Approach // CEUR Workshop Proceedings of SEPLN, 2023.
3. Boeva G., Gritsai G., Grabovoy A., et al Team ap-team at PAN: LLM Adapters for Various Datasets // CEUR Workshop Proceedings of CLEF, 2024.
4. Gritsai G., Grabovoy A. Automated Text Identification on Languages of the Iberian Peninsula: LLM and BERT-based Models Aggregation // CEUR Workshop Proceedings of SEPLN, 2024.
5. Chekhovich Y., Grabovoy A., Gritsai G. Generative AI Models with Their Full Reveal // International Conference on Technology Enhanced Learning in Higher Education, 2024.
6. Gritsai G., Grabovoy A., Khabutdinov I. Multi-head Span-based Detector for AI-generated Fragments in Scientific Papers // Workshop on Scholarly Document Processing @ ACL, 2024.
7. Gritsai G., Voznyuk A., Khabutdinov I., Grabovoy A. Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking // Workshop on Detecting AI Generated Content @ COLING, 2025.
8. Gritsai G., Voznyuk A., Grabovoy A., Chekhovich Y. Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts // Workshop on Preventing and Detecting LLM Misinformation @ AAI, 2025.

Выступления с докладом

1. Автоматическая детекция машинно-сгенерированных текстов: нужно больше токенов, Международная конференция «Иванниковские чтения», 2022.
2. Многозадачное обучение для распознавания машинно-сген. текстов «65-я научная конференция МФТИ», 2023.
3. Automated Text Identification: Multilingual Transformer-based Models Approach, IberLEF @ SEPLN, 2023.
4. Внимание, документ подозрительный! Жизнь с машинной генерацией в научном сообществе, RuCode, 2024.
5. Multi-head Span-based Detector for AI-generated Fragments in Scientific Papers, SDP @ ACL, 2024.
6. LLM Adapters for Various Datasets, PAN @ CLEF, 2024.
7. Automated Text Identification on Languages of the Iberian Peninsula, IberLEF @ SEPLN, 2024.
8. AI Detection Powered by Domain-Aware Multi-Tasking, DetectGenAI @ COLING, 2025.
9. Выравнивание представлений в многозадачном обучении для сген. текстов «67-я научная конференция МФТИ», 2025.

Список работ по теме диссертации

Публикации в журналах из списка ВАК

1. Г. М. Грицай, А. В. Грабовой и др. Поиск искусственно сгенерированных текстовых фрагментов в научных документах // Докл. РАН. Матем., информ., проц. упр., 541, 2023.
2. Avetisyan K., Gritsay G., Grabovoy A. Cross-Lingual Plagiarism Detection: Two Are Better Than One // Programming and Computer Software, 2023.
3. Г. М. Грицай, И. А. Хабутдинов, А. В. Грабовой Stack More LLM's: Эффективное обнаружение машинно-сгенерированных текстов с помощью аппроксимации значений перплексии // Докл. РАН. Матем., информ., проц. упр., 520, 2024.

Программные модули разработанные в рамках диссертационной работы

1. Программная система для распознавания текстовых материалов, созданных при помощи искусственного интеллекта // Свидетельство №202561749, дата регистрации в Реестре государственных программ 26.03.2025.