Поиск искусственно сгенерированных текстовых фрагментов в научных документах

Г. М. Грицай

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

Московский физико-технический институт

2024

Цель исследования

Цель

Предложить метод детектирования машинно-сгенерированных текстовых последовательностей, основанный на паттернах присущих искусственным фрагментам.

Задача

Повышение полноты распознавания искусственных текстовых последовательностей, используя семейство моделей глубокого обучения.

Метод решения

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

Постановка задачи детекции в текстовых последовательностях

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: $\mathbf{D} = \bigcup_{i=1}^{N} D^i, D^i \in \mathbb{D}$.
Множество непересекающихся фрагментов документа:

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

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:

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

$$\mathbf{f} : \mathbb{D} \to \mathbf{T^*}, \qquad \mathbf{g} : \mathbf{T^*} \to \mathbf{C},$$

$$\phi : \mathbb{D} \to \mathbb{T}, \qquad \mathbb{T} = \{[t_{s_j}, t_{f_j}, c_j]_{j=1}^{J} \quad | \quad t_{s_j} = t_{f_{j-1}}, \quad s_j \in \mathbb{N}_0, \quad f_j \in \mathbb{N}, \quad c_j \in \{0, 1\}\},$$

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

Проблемы множественных сравнений

Классификатор, минимизирующий эмпирический риск в наборе **D**:

$$\hat{g} = \underset{g \in \mathfrak{F}}{\operatorname{argmin}} \sum_{D^i \in \mathbf{D}} \sum_{x_j, c_j \in D^i} [g(t(x_j)) \neq c_j], \qquad t : \mathbf{T^*} \to (V)^n, \tag{1}$$

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.

Проверка гипотез:

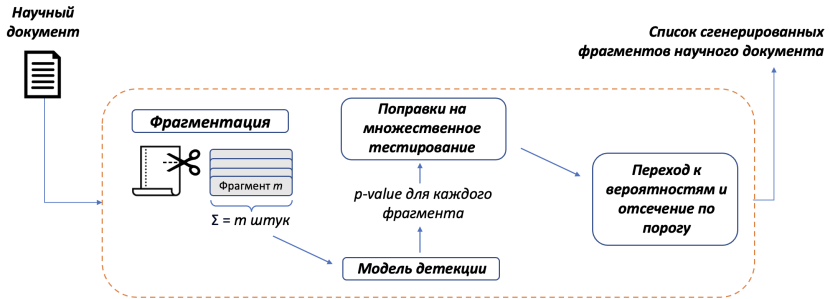
$$H_0 : \hat{g}(fragment) = 0,$$
$$H_1 : \hat{g}(fragment) = 1.$$

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

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

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

Предложенный алгоритм работы с документами



Полный цикл работы алгоритма детекции сгенерированных фрагментов в научных документах

В текущей задаче:

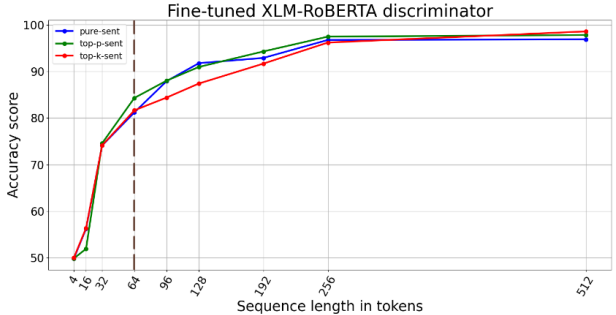
$$p - value = 1 - m(t(x_i)), \quad m : V \to P, \qquad \forall p \in P : \sum_{i=1}^{d} p_i = 1,$$

где m - предсказание предобученной модели, P - пространство векторов из $\mathbb{R}^d$.

Зависимость качества классификации от длины входа

$$\hat{h} = \underset{n \in \mathbb{N}}{\operatorname{argmin}} \sum_{x \in \mathbb{X}} [g(t_n(x)) \neq c], \qquad t_n : \mathbb{X} \to (V)^{\mathbf{n}}, \qquad c \in \mathbf{C}\{0, 1\}, \qquad (2)$$

где x - текстовая послед., X - множество всех текстовых фрагментов, V - словарь всевозможных токенов предобученной модели, n - варьируемая длина входного вектора.



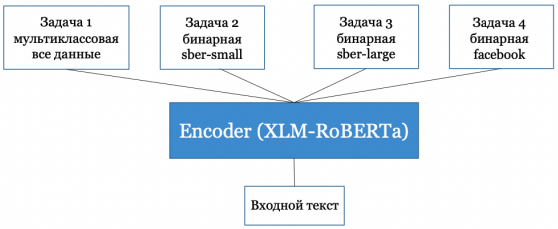
Точность классификатора основанного на архитектуре трансформер возрастает с увеличением длины последовательностей.

Подход многозадачного обучения

Пусть M задачам классификации соответсвует множество датасетов $\mathbb{D} = \{d_1, d_2, ..., d_M\}$. Задан ρ_i - вес сэмплирования $\forall d_i \in \mathbb{D}$:

$$\rho_i \propto |d_i|^{\alpha}, \qquad \alpha(e) = 1 - \frac{0.8(\epsilon - 1)}{\varepsilon - 1},$$

где ϵ - текущая эпоха, ε - общее количество эпох.



Архитектура и разложение по двум главным компонентам текстов на основе векторного представления после слоев энкодера при обучении модели в многозадачном режиме.

Набор данных для исследования

Структура собранного набор данных для русского и английского языков.

Часть датасета	Источник	Количество текстов	Средняя длина
Ru-human	MGTDR	31.500	3.810
	Yandex Q	11.700	1.950
	Rus. Essays	1.000	3.366
Ru-machine	MGTDR	31.500	3.665
	Alpaca	7.600	1.361
	Saiga	4.000	1.190
	Rus. Essays	1.000	3.414
En-human	DeepFake	31.500	2.315
	HC3	12.600	1.150
En-machine	DeepFake	31.500	2.398
	GPT4	12.600	1.560

Результаты вычислительного эксперимента

Язык	Эксперимент	F1-score	Precision	Recall
ru	ручные признаки	0.960	0.962	0.959
	мультиязычное обучение	0,964	0,964	0,966
	перевод текстов 25%	0,958	0,961	0,958
	перевод текстов 50%	0,966	0,968	0,966
	парафраз предложений 100%	**0,968**	**0,970**	**0,968**
	парафраз предложений 50%	0,964	0,965	0,963
Язык	**Эксперимент**	**F1-score**	**Precision**	**Recall**
en	ручные признаки	0.801	0.856	0.807
	мультиязычное обучение	0,823	0,867	0,828
	перевод текстов 25%	0,821	0,865	0,826
	перевод текстов 50%	**0,825**	**0,868**	**0,830**
	парафраз предложений 100%	0,822	0,866	0,827
	парафраз предложений 50%	0,816	0,862	0,817

Сводная таблица результатов вычислительного эксперимента для наборов данных на русском и английском языках.

Заключение

Сделано:

- Предложен подход детекции машинно-сгенерированных фрагментов в документах, основанний на фрагментации, множественном тестировании и классификации сегментов.
- Выявлена зависимость качества классификации от длины входной последовательности в моделях классификации с архитектурой трансформер.
- Показано, что многозадачное обучения повышает обобщающую способность модели и улучшает заданные метрики качества бинарных задач.

Планируется:

- Предложить подход детекции с фрагментацией варьируемой длины.
- Исследовать многозадачное обучение в пайплайне научных документов.

Список работ по теме НИР

Публикации в журналах

1. **Gritsay G., Grabovoy A., Chekhovich Y.** Automatic Detection of Machine Generated Texts: Need More Tokens // 2022 Ivannikov Memorial Workshop (IVMEM). – IEEE, 2022.
2. **Г. М. Грицай, А. В. Грабовой** Многозадачное обучение для распознавания машинно-сгенерированных текстов // 65-ая Всероссийская научная конференция МФТИ, 2023.
3. **Gritsay, G., Grabovoy, et all** Automated Text Identification: Multilingual Transformer-based Models Approach // CEUR Workshop Proceedings of SEPLN, 2023.
4. **Г. М. Грицай, А. В. Грабовой и др.** Поиск искусственно сгенерированных текстовых фрагментов в научных документах // Докл. РАН. Матем., информ., проц. упр., 541, 2023.
5. **Г. М. Грицай, А. В. Грабовой и др.** Генерация и поиск искусственно сгенерированных текстовых фрагментов в домене научных работ // Информатика и ее применения, 2024.
6. **Avetisyan K., Gritsay G., Grabovoy A.** Cross-Lingual Plagiarism Detection: Two Are Better Than One // Programming and Computer Software, 2023.

Выступления с докладом

1. Автоматическая детекция машинно-сгенерированных текстов: нужно больше токенов, Международная конференция «Иванниковские чтения», 2022.
2. Многозадачное обучение для распознавания машинно-сгенерированных текстов «65-я научная конференция МФТИ», 2023.
3. Automated Text Identification: Multilingual Transformer-based Models Approach, XXXIX International Congress of the Spanish Society for Natural Language Processing, 2023.