ОБНАРУЖЕНИЕ ИСКУССТВЕННЫХ РУССКИХ ТЕКСТОВ С ПОМОЩЬЮ АЛГОРИТМА DetectGPT

С.А. Есаян
Российско-Армянский университет
Центр передовых программных технологий
yesayan.svetlana@student.rau.am

АННОТАЦИЯ

В данной работе был адаптирован к русскому языку алгоритм обнаружения искусственных текстов DetectGPT и была изучена его эффективность на разных наборах данных. В частности, рассмотрена возможность применения этого метода для выявления фрагментов, сгенерированных с помощью ChatGPT, в дипломных работах студентов.

Ключевые слова: ChatGPT, DetectGPT, искусственные тексты

Введение

В последнее время появилось много моделей, генерирующих искусственные тексты, что привело к распространению ложных новостей, генерированию фальшивых отзывов о продуктах, генерированию дипломных работ и т.д.. Различение сгенерированных и естественных текстов стало серьезной проблемой и разработка эффективных методов для различия искусственных и естественных текстов является актуальной задачей. В нашей работе предполагается, что тексты могут быть двух видов: естественные, то есть написанные человеком, и искусственные - сгенерированные с помощью языковых моделей.

Есть много различных методов обнаружения искусственных текстов [1-5]. В статье [1] используют метод, который работает полагаясь на топологические особенности, получаемые от слоев attention [6] из архитектуры трансформеров. DetectGPT [3] — еще один метод обнаружения сгенерированных текстов. Основная гипотеза метода заключается в том, что незначительные изменения, внесенные в сгенерированный моделью текст, обычно имеют более низкую логарифмическую вероятность, чем исходный образец, в то время как незначительные изменения текста, написанного человеком, могут иметь как высокую, так и низкую логарифмическую вероятность. В основе этого метода лежат 2 языковые модели: одна для генерации частично изменённых версий исходного текста, а другая для определения правдоподобия.

DetectGPT был тестирован на английских и немецких текстах из датасета WMT16 [7] и результаты на обоих языках по метрике Auroc[1] были высокими (>0.8). Это может говорить о том, что метод устойчив для использования на других языках, отличных от английского. В данной работе было принято решение проверить эффективность DetectGPT в задаче обнаружения искусственных участков текста, полученных с помощью ChatGPT[2], в русскоязычных дипломных работах студентов. Был разработан тестовый набор из текстов дипломных работ и проводилось исследование наиболее эффективной конфигурации языковых моделей для применения подхода DetectGPT на

[1] AUROC вычисляется как площадь под кривой ROC. Кривая ROC показывает компромисс между истинно положительной (TPR) и ложно положительной (FPR) частотой в зависимости от порога решения.Наихудший показатель AUROC равен 0.,5, а наилучший - 1.0.

[2] “Introducing ChatGPT.” *OpenAI*, 30 November 2022, https://openai.com/blog/chatgpt. Accessed 18 March 2023.

текстах русского языка. В результате экспериментов были получены положительные первичные результаты для пары XLM-RoBERTa [8] и RuGPT3-large[3], где первая модель использовалась для модификации текста, а вторая — для вычисления правдоподобия.

Учитывая небольшой размер собранного тестового набора, для получения более достоверной оценки эффективности метода дополнительно протестировали его на датасете RuATD[4], который использовался в открытом соревновании по обнаружению машинно-сгенерированных текстов в рамках научной конференции Диалог-2022[5]. На этих данных метод оказался гораздо менее эффективным, и был проведен дополнительный анализ этих данных для выяснения причин таких результатов.

Описание DetectGPT

Основополагающая гипотеза метода DetectGPT состоит в том, что небольшие модификации искусственного экземпляра в среднем имеют более низкое правдоподобие по сравнению с исходным вариантом, а в случае естественного текста правдоподобие его модификаций может быть как ниже, так и выше. Для вычисления правдоподобия предлагается использование больших языковых моделей. Другими словами, если имеем текст X и его слегка изменённые версии x, то для большой языковой модели $p\theta$ в среднем $log\ p\theta(X) - log\ p\theta(x)$ должно быть относительно большим для искусственных образцов, по сравнению с естественными. Алгоритм, используемый на практике для реализации и применения метода, описан на Схеме 1.

Алгоритм DetectGPT

Input: Текст X, языковая модель P_θ, модель для генерации модифицированных версий текста q, количество модифицированных версий K, порог ε

$x_i \sim q(\cdot|x),\ i = 1, 2, \ldots, K$

$$\mu = \frac{1}{k} * \sum_{i=1}^{K} log\ p\theta(x_i)$$

$$d = log\ p\theta(X) - \mu$$

$$\delta^2 = \frac{1}{K-1} \sum_{i=1}^{k} (log\ p\theta(x_i) - \mu)^2$$

if $\frac{d}{\sqrt{\delta^2}} > \varepsilon$ **then**

    return true // текст искусственный

else

    return false // текст естественный

Схема 1. *Псевдокод реализации алгоритма DetectGPT.*

Разработка тестового набора для русского языка

Для создания тестового набора использовались фрагменты текстов ВКР, предоставленные вузом РАНХиГС. Из 30 документов было получено 200 фрагментов, в среднем состоящих из 168 слов. Половина была сохранена в исходном виде, как естественные тексты. Остальные 100 были использованы для генерации искусственных текстов с помощью чат-бота ChatGPT: были взяты первые 30 слов из каждого фрагмента и использовали их в качестве подсказок.

Оценка эффективности модели

[3] www.huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2 Accessed 18 March 2023.
[4] www.kaggle.com/c/ruatd-2022-bi Accessed 18 March 2023.
[5] www.dialog-21.ru Accessed 18 March 2023.

В методе DetectGPT используются 2 языковые модели. Первая модель использовалась для получения незначительно изменённых текстов из первоначального фрагмента. В исходной статье эмпирическим способом показано, что 100 таких текстов достаточно для хороших результатов. Для получения изменённых текстов случайным образом маскировалось 15% слов в фрагментах и maskfill[6] модели заполняли эти пропуски. Для русского языка в рамках этой работы были рассмотрены следующие maskfill модели: *XLM-R, mBert, RuBert*.

Вторая модель использовалась для получения логарифмических вероятностей изначального фрагмента и 100 текстов, полученных из него. Были выбраны следующие языковые модели: *RuGPT2-large, RuGPT3-small, RuGPT3-medium, RuGPT3-large, XLM-R, mBert, RuBer*t.

Для экспериментов была выбрана метрика Accuracy, так как датасет балансированный: 100 естественных и 100 искусственных фрагментов.

Таблица 1. *Результаты метода DetectGPT при использовании разных конфигураций языковых моделей на собранном датасете. Первое число - accuracy score, второе число - выбранный порог.*

	RuGPT2- large	RuGPT3-small	RuGPT3-medium	RuGPT3-large	*Avg.*
XLM-R	0.93 (-3.53)	0.94 (-3.45)	0.91 (-3.07)	**0.95** (-4.6)	*0.932*
mBert	0.92 (-4.7)	0.93 (-4.98)	0.915 (-4.48)	0.92 (-5.47)	*0.921*
RuBert	0.91 (-3.84)	0.945 (-3.31)	0.935 (-4.1)	0.925 (-5.23)	*0.928*
Avg.	*0.92*	*0.938*	*0.92*	*0.932*	

Таблица 2. *Результаты метода DetectGPT при использовании разных конфигураций языковых моделей на собранном датасете. Первое число - accuracy score, второе число - выбранный порог.*

	XLM-R	mBert	RuBert	*Avg.*
XLM-R	0.505 (1.99)	0.49 (0.01)	0.5 (1.99)	*0.498*
mBert	0.5 (0.06)	0.5 (0.0)	0.505 (0.1)	*0.502*
RuBert	0.495 (0.4)	0.49 (0.64)	**0.53** (0.59)	*0.505*
Avg.	*0.5*	*0.493*	*0.512*	

Из таблиц 1 и 2 видно, что использование GPT моделей даёт заметно более хорошие результаты. Модели на основе Bert для получения логарифмических оценок текста работают на уровне рандома (accuracy ~0.5), и их использование в алгоритме DetectGPT нецелесообразно. Наилучший результат получился при комбинации XLM-R + RuGPT3-large (accuracy 0.95).

Оценка модели на датасете RuATD

Был проведён эксперимент метода DetectGPT на тестовом датасете RuATD, для его более точной оценки, так как предыдущий эксперимент проводился на небольшом наборе данных. Количество текстов в тестовом датасете RuATD — 64,533, в среднем состоящих из 31 слов. Была выбрана наилучшая комбинация предыдущего эксперимента XLM-R + RuGPT3 и порог -4.6. Так как результат был плохим (accuracy - 0.499), было принято решение получить гистограмму оценок текстов, до классифицировании по порогу, и так как оценки были около нуля, посчитали accuracy для -0.05, -0.1, -0.2 порогов также.

Таблица 3. *Результаты DetectGPT (с XLM-R и RuGPT3-large) на датасете RuATD .*

Threshold	Accuracy
-4.6	0.499

[6] www.huggingface.co/tasks/fill-mask Accessed 18 March 2023.

-0.05	0.489
-0.1	0.489
-0.2	0.491

Анализ полученных результатов

Результаты на двух датасетах были достаточно разными. На нашем датасете метод DetectGPT(XLM-R + RuGPT3large) дал точность 0.95, а на датасете RuATD 0.49-0.5. Для анализа полученных результатов было проведено сравнение датасетов путем измерения длины текстов в каждом из них. Результаты данного сравнения представлены на Рис. 1.

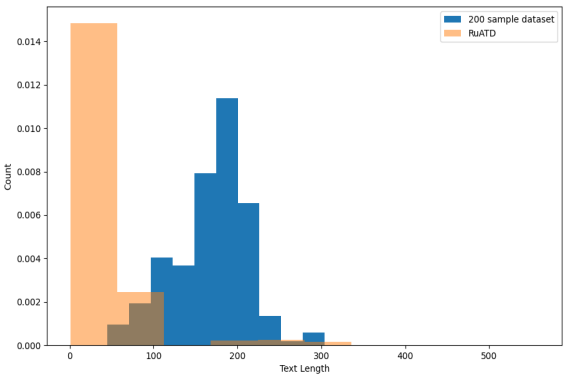

Рис. 1. *Гистограмма описывает количество слов в текстах в 2 датасетах: RuATD, 200 sample dataset*

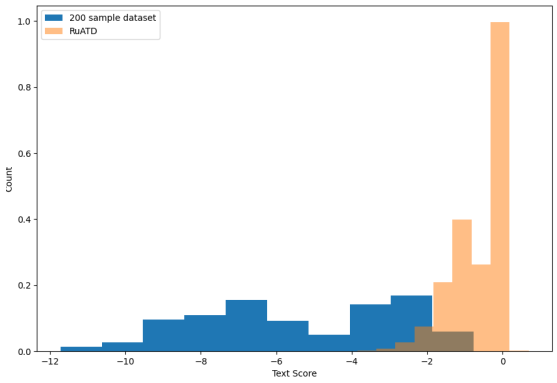

Рис. 2. *Гистограмма описывает оценки текстов модели RuGPT3-large на 2 датасетах: RuATD, 200 sample dataset*

Из Рис. 1. можно сделать вывод, что тексты RuATD намного короче, чем тексты нами собранного датасета. Также для анализа были построены гистограммы оценок текстов модели RuGPT3large для 2 датасетов (Рис. 2.). В Рис. 2. видно, что большинство текстов в RuATD получили оценки ближе к 0 (> 50,000), что усложняет выбор порога для классификации. Это, скорее всего, связано с тем, что тексты в RuATD короткие, в среднем состоят из 31 слов, и полученные из них 100 модифицированные версии не дают достаточно информации для правильной работы DetectGPT.

Заключение

В ходе выполнения данной работы было выявлено, что метод DetectGPT можно использовать для обнаружения искусственных фрагментов в ВКР. Для этого был собран небольшой датасет из фрагментов ВКР, предоставленные вузом РАНХиГС, а искусственные фрагменты генерировались с помощью ChatGPT. Метод DetectGPT основывается на 2 языковых моделях, поэтому для оценки его эффективности на собранном датасете были выбраны разные комбинации языковых моделей. Лучший результат получился в комбинации XLM-R и RuGPT3-large. Учитывая небольшой размер датасета, для получения более достоверной оценки эффективности метода дополнительно протестировали его на датасете RuATD, результаты которого были плохими, что может быть связано с тем, что тексты в этом датасете короткие и их оценки в основном ближе к 0, в отличие от текстов, которые были в нашем датасете.

ЛИТЕРАТУРА

1. Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. (2020). Automatic Detection of Machine Generated Text: A Critical Survey. *International Committee on Computational Linguistics*, 2296-2309.

2. Kushnareva, L., Cherniavskii, D., Mikhailov, V., Artemova, E., Barannikov, S., Bernstein, A., Piontkovskaya, I., Piontkovski, D., & Burnaev, E. (2021). Artificial Text Detection via Examining the Topology of Attention Maps. 635–649.
3. Mitchell, E., Lee, Y., Khazatsky, A., D. Manning, C., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature.
4. Gehrmann, S., Strobelt, H., & M. Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text - Sebastian Gehrmann Harvard SEAS. *Association for Computational Linguistics*, 111–116.
5. Nguyen-Son, H.-Q., Phuong Thao, T., Hidano, S., Gupta, I., & Kiyomoto, S. (2021). Machine Translated Text Detection Through Text Similarity with Round-Trip Translation. *Association for Computational Linguistics*, 5792–5797.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., N. Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.
7. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., & Turchi, M. (2016). Findings of the 2016 Conference on Machine Translation. 131–198.
8. Conneau, A., & Lample, G. (2019). Cross-lingual Language Model Pretraining. *33rd Conference on Neural Information Processing Systems*.

ARTIFICIAL TEXT DETECTION FOR RUSSIAN TEXTS USING DetectGPT METHOD

S. H. Yesayan
Russian - Armenian University
Center of Advanced Software Technologies
yesayan.svetlana@student.rau.am

ABSTRACT

In the course of the experiments conducted in this work, we studied the effectiveness of the DetectGPT algorithm in detecting artificial texts in Russian, as well as the possibility of using this method to identify ChatGPT generated fragments in the graduation theses.

Keywords: ChatGPT, DetectGPT, artificial texts

ՌՈՒՍԵՐԵՆ ԱՐՀԵՍՏԱԿԱՆ ՏԵՔՍՏԵՐԻ ՀԱՅՏՆԱԲԵՐՈՒՄԸ **DetectGPT** ԱԼԳՈՐԻԹՄՈՎ

Ս. Հ. Եսայան
Հայ - Ռուսական համալսարան
Առաջադեմ ծրագրային տեխնոլոգիաների կենտրոն
yesayan.svetlana@student.rau.am

ԱՄՓՈՓՈՒՄ

Աշխատանքում ուսումնասիրվել է DetectGPT ալգորիթմի էֆֆեկտիվությունը ռուսերեն արհեստական տեքստերի հայտնաբերման խնդրում: Մասնավորապես, դիտարկվել է ալգորիթմի կիրառման հնարավորությունը ռուսերեն ավարտական դիպլոմային աշխատանքներում ChatGPT-ի օգնությամբ գեներացված արհեստական պարբերությունների հայտնաբերման նպատակով:

Հիմնաբառեր: ChatGPT, DetectGPT, արհեստական տեքստեր