Защита от атак с использованием омоглифов на модели NLP

Кузнецов А.А.[1,2], Михайлов А.А.[1,2]

[1]*Институт динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, 664033, Иркутск, ул. Лермонтова, д. 134, Российская Федерация*

[2]*Институт системного программирования им. В.П. Иванникова Российской академии наук, 109004, г. Москва, ул. Александра Солженицына, д. 25, Российская Федерация*

Аннотация

В статье рассматривается проблема атак на системы обработки естественного языка (NLP) с использованием визуально похожих символов — омоглифов. Омоглифы представляют собой графически идентичные или похожие символы, имеющие разное значение. В статье приводится анализ существующих методов защиты от таких атак и предлагается новый подход на основе модели T5, обученной под корректировку орфографических ошибок на английском языке.

Ключевые слова

Обработка естественного языка, защита от атак на языковые модели, омоглифы,

1. Введение

С развитием глубоких нейронных сетей в области обработки естественного языка [1] стало ясно, что подобные системы уязвимы для атак, искажающих входные данные.

Раньше атаки на языковые модели были заметными для человеческого глаза, например, орфографические ошибки или перефразирования. Сейчас же набирают популярность атаки с использованием омоглифов — графически одинаковых или похожих друг на друга знаков, имеющих разное значение. Они незаметны для человека, но ощутимо различаются при обработке.

Возможность изменить текст, не изменяя его визуально, может быть использована многими злоумышленниками для обхода механизмов фильтрации контента, некорректного машинного перевода, ухудшения качества запросов и индексации поисковых систем. Согласно статье [2] многие системы обработки естественного языка не обращают внимания на возможные проблемы: от снижения производительности до некорректной работы.

Понимание сущности атак с использованием омоглифов и их потенциальных последствий для безопасности систем станет ключом для разработки эффективных методов

6^{th} International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2024), July 01–05, 2024, Irkutsk, Russia

viirtualkuz@yandex.ru (А.А.); mikhailov@icc.ru (А.А.)

0009-0001-1105-0076 (А.А.); 0000-0003-4057-4511 (А.А.)

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)
DOI: 10.47350/ICCS-DE.2024.31

защиты. Анализ реальных случаев атак с использованием омоглифов поможет выявить наиболее уязвимые области и разработать соответствующие меры противодействия.

2. Предлагаемый подход

В статье предлагается новый подход защиты от атак с использованием омоглифов на основе модели T5. Модель T5 — text to text transfer transformer [3, 4], представляет собой предобученную командой SberDevices [5] модель, которая способна корректировать орфографические ошибки на английском языке. Для генерации обучающего корпуса существует два разных алгоритма: с использованием случайной замены символа на его омоглиф и на основе генетического алгоритма [6]. Составленный словарь омоглифов рассчитан на замену латинских символов. В среднем по 6 омоглифов на каждый символ. Размер обучающего корпуса превышает 400 тыс. элементов (алгоритм случайной замены).

Промежуточный результат после первых попыток обучения составляет свыше 80% (Accuracy). Полученные результаты демонстрируют перспективность выбранного подхода и эффективность использования модели T5 для решения задачи защиты от атак с использованием омоглифов. Однако, несмотря на положительные результаты, необходимо дальнейшее развитие проекта, включающее в себя работу с русским языком (кириллицей), расширение существующего словаря омоглифов, генерацию большого обучающего корпуса с использованием генетического алгоритма, добавление новых метрик в процесс оценивания работы модели и оптимизацию модели.

3. Заключение

Таким образом, предложенный подход на основе модели T5 является перспективным направлением для разработки эффективных методов защиты от атак с использованием омоглифов. Дальнейшее развитие проекта позволит расширить область применения предложенного метода и повысить уровень безопасности систем обработки естественного языка.

Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

Список литературы

[1] J. O'Connor, I. McDermott, NLP, Thorsons, 2001.
[2] N. Boucher, I. Shumailov, R. Anderson, N. Papernot, Bad characters: Imperceptible nlp attacks, in: 2022 IEEE Symposium on Security and Privacy (SP), IEEE, 2022, pp. 1987–2004.

[3] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, Overview of the transformer-based models for nlp tasks., In 2020 IEEE 15th Conference on Computer Science and Information Systems (FedCSIS) (2022) 179–183.
[4] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y. H. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences., 2020. URL: https://doi.org/10.48550/arXiv.2010.11934.
[5] N. Martynov, M. Baushenko, A. Kozlova, K. Kolomeytseva, A. Abramov, A. Fenogenova, A methodology for generative spelling correction via natural spelling errors emulation across multiple domains and languages, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 138–155. URL: https://aclanthology.org/2024.findings-eacl.10.
[6] L. Haldurai, T. Madhubala, R. Rajalakshmi, A study on genetic algorithm and its applications, International Journal of computer sciences and Engineering 4 (2016) 139.