Защита от незаметных атак на модели NLP с использованием омоглифов

Андрей Александрович Кузнецов, [1] Андрей Анатольевич Михайлов [1,2]

[1] *Институт динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, Россия, 664033, г. Иркутск, ул. Лермонтова, 134*
[2] *Институт системного программирования им. В. П. Иванникова Российской академии наук, Россия, 109004, г. Москва, ул. Александра Солженицына, д. 25*

Аннотация

В работе рассматривается проблема атак на системы обработки естественного языка (NLP), реализуемых с применением омоглифов — символов, которые визуально идентичны или схожи, но обладают различными значениями. Предлагается новый подход, основанный на применении модели трансформера T5, обученной для корректировки орфографических ошибок на синтетически созданном корпусе данных, в котором символы были заменены на их омоглифы. Результаты эксперимента демонстрируют высокую точность модели в исправлении атакованных примеров

Ключевые слова

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

1. Введение

С развитием глубоких нейронных сетей в области обработки естественного языка стало очевидным, что подобные системы уязвимы для атак, искажающих входные данные. В отличие от атак в области компьютерного зрения, атаки на языковые модели в NLP могут быть более заметными, например, нарушение орфографии или перефразирование. Возможность изменить текст, не изменяя его визуально, может быть использована злоумышленниками для обхода механизмов фильтрации контента, некорректного машинного перевода, ухудшения качества запросов и препятствия индексации поисковых систем [1].

2. Обзор работ

Современные работы в области атак на модели обработки естественного языка [2,3,4] предлагают различные способы защиты, которые можно разделить на следующие подходы:

- Система, основанная на правилах замены (rule-based recovery). Суть заключается в символьном проходе по входной строке и замене символа, отличного от языка оригинала. Проблема данного решения заключается в том, что подобная система строится только для одного языка.
- Состязательное обучение (Adversarial Training). Базовый вариант защиты NLP моделей от состязательных атак – добавление «враждебных» примеров в процесс обучения модели. Недостатками данного решения являются увеличение затрат на обучение исходной модели и неполноценная защита от состязательных примеров.

7th International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2025), July 7–11, 2025, Irkutsk, Russia
EMAIL: viirtualkuz@yandex.ru (A. 1); mikhailov@icc.ru (A. 2)
ORCID: 0000-0003-4057-4511 (A. 2)

ICCS-DE 2025 Workshop Proceedings
DOI: 10.47350/ICCS-DE.2025.29

- Создание векторных представлений для каждого символа. Следующая группа работ предлагает создавать векторные представления для каждого символа и, сравнивая с соседями, находить ближайшие аналоги среди других представлений.
- Коррекция визуально схожих слов. В работе по коррекции визуально схожих слов в арабском языке упоминается способ на основе двунаправленных рекуррентных нейронных сетей (BiLSTM), анализирующий контекст слова как в прямом, так и в обратном направлении.
- Коррекция и детекция омоглифов в тексте на основе свёрточных нейронных сетей (CNN). В процессе работы генерируется изображение на основе входного текста, в котором могут присутствовать омоглифы, затем оно поступает на вход модели для последующей детекции и коррекции.

3. Метод коррекции омоглифов

Для решения задачи коррекции омоглифов предлагается использовать языковую модель на базе трансформера, а именно модель семейства T5. Главной её особенностью является то, что она предобучена на большом массиве текстовых данных, покрывающем объёмный спектр NLP задач и используется для дообучения под конкретные цели. Наш метод, T5-GlyF (Glyph Fixing), основан на проекте SAGE команды SberDevices, в котором модель на базе T5 выполняет коррекцию орфографических и пунктуационных ошибок на русском и английском языках. Для создания большого обучающего корпуса авторами был разработан алгоритм аугментации данных, который имитировал ошибки человека в правописании.

4. Эксперименты и результаты

- Разработан алгоритм генерации возмущений с использованием омоглифов.
- Сгенерирован датасет с помощью разработанного алгоритма.
- Проведено обучение модели на сгенерированном датасете.
- Результаты обучения демонстрируют высокую точность модели в исправлении атакованных примеров: 99% для английского языка и 96% при добавлении русского языка.

5. Заключение

Искажение входных данных на основе омоглифов в области NLP является серьёзной угрозой для языковых моделей. Решение на основе трансформера, T5-GlyF, оказалось отличным вариантом системы защиты против атак с использованием омоглифов согласно проведённым тестированиям на реальных NLP задачах. Наша система защиты успешно справляется с атакованными с помощью омоглифов входными данными, а в случае, когда никакой атаки нет, эффективность целевой языковой модели остаётся неизменной.

6. Благодарности

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

7. Список литературы

[1] Boucher N. et al. Bad characters: Imperceptible nlp attacks //2022 IEEE Symposium on Security and Privacy (SP). – IEEE, 2022. – С. 1987-2004.

[2] Eger S. et al. Text processing like humans do: Visually attacking and shielding NLP systems //arXiv preprint arXiv:1903.11508. – 2019.
[3] Almanea M. M. Automatic methods and neural networks in Arabic texts diacritization: a comprehensive survey //IEEE Access. – 2021. – Т. 9. – С. 145012-145032.
[4] Han X. et al. Text adversarial attacks and defenses: Issues, taxonomy, and perspectives //Security and Communication Networks. – 2022. – Т. 2022. – №. 1. – С. 6458488.