УДК 004.89

ИНСТРУМЕНТЫ ДЛЯ ИЗМЕРЕНИЯ УСТОЙЧИВОСТИ ЯЗЫКОВЫХ МОДЕЛЕЙ К АТАКУЮЩИМ ШУМАМ

Бессарабов А.А., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Развитие языковых моделей – от классических RNN до современных трансформеров и крупномасштабных моделей наподобие GPT, T5 или LLaMA – привело к качественному скачку в решении задач обработки естественного языка. Однако параллельно с ростом их способности к генерализации, наблюдается рост и чувствительности к незначительным, но намеренным или случайным искажениям входного текста, так называемым атакующим шумам [1].

Даже минимальные отклонения – например, опечатки, перестановки слов, замены синонимов или синтаксические переформулировки – способны вызывать резкие изменения в предсказаниях моделей. Это особенно опасно в таких задачах, как модерация контента, автоматическая классификация юридических или медицинских документов, машинный перевод и извлечение фактов. В этих условиях задача измерения устойчивости языковых моделей к атакующим шумам становится ключевым направлением в построении доверенных NLP-систем. Цель данной работы – рассмотреть современные инструменты, подходы и метрики, которые позволяют формализованно и воспроизводимо оценивать чувствительность языковых моделей к шумам различной природы.

Под атакующим шумом понимаются такие трансформации входного текста, которые при сохранении общей семантики или даже синтаксической правильности провоцируют нежелательные или неверные реакции модели. В отличие от случайных ошибок пользователя, атакующие искажения могут быть намеренно сконструированы с использованием языковых, статистических или дифференцируемых методов. Примеры таких атак включают замену слов синонимами, вставку «незаметных» токенов, перестановку частей фраз, использование омонимов, генерацию контрафактов и подмену важнейших лексем на нейтральные. В задачах генерации это может привести к смещению тона, нарушению фактической достоверности или генерации токсичного контента, а в классификации – к полной смене результата [2].

Для количественной оценки устойчивости языковых моделей к таким воздействиям разрабатывается целый спектр инструментов и фреймворков,

предназначенных для систематического создания атакующих шумов и анализа реакции модели. Один из наиболее известных – TextAttack, который позволяет моделировать более чем 30 различных атак: от простых замен слов по словарю до сложных атак, использующих градиенты модели. Он поддерживает широкий спектр задач (sentiment analysis, NLI, QA, hate speech detection и др.), а также может интегрироваться с HuggingFace Transformers. Помимо генерации атак, TextAttack предоставляет метрики успеха атак (attack success rate), степени модификации, метрики семантической близости между оригиналом и атакой (например, BLEU, BERTScore), а также визуализации различий [3].

Другой важный инструмент – OpenAttack, ориентированный на проведение атак как в white-box, так и в black-box режимах. Он предоставляет унифицированный интерфейс для сравнения моделей по показателю устойчивости в одинаковых условиях и позволяет конструировать сложные сценарии с участием механизмов поиска по пространству семантически эквивалентных выражений. Особенность OpenAttack – модульность, позволяющая быстро добавлять новые методы атаки и адаптировать их под различные языки.

Кроме атакующих фреймворков, активно развиваются инструменты тестирования устойчивости через генерацию парафраз. Например, CheckList предлагает систематизированное покрытие пространства лингвистических трансформаций, включая семантические, синтаксические и прагматические изменения. Вместо «вредоносной» атаки он ориентирован на обнаружение слабых мест модели через перебор разнообразных, но валидных формулировок. Это особенно ценно в задачах генерации вопросов, чат-ботов и классификации пользовательских отзывов, где разнообразие форм выражения может быть практически бесконечным.

Наиболее современное направление – это метрики автоматического измерения устойчивости. Они представляют собой количественные показатели, которые можно рассчитывать без необходимости вручную запускать атаки. К их числу относятся:

- Local Robustness Score – доля семантически близких входов, при которых предсказание сохраняется;
- Semantic Consistency under Noise – изменение уверенности модели при малых трансформациях текста; [4]
- Stability Index – средняя вариация логитов при токенизированных мутациях входа;
- Certainty Gap – чувствительность вероятностного вывода к нейтральным модификациям;
- Token Influence Maps – распределение вклада токенов в итоговое решение, где высокая концентрация может свидетельствовать об уязвимости.

Для реализации этих метрик используют такие инструменты, как Contrastive Test Sets, Adversarial Example Mining, Language Model Scoring, BERT-based similarity embeddings. Они применимы как к классификаторам, так и к генеративным моделям, включая large language models (LLMs) [5].

Однако важным направлением становится не только измерение робастности, но и профилирование уязвимостей модели. Это включает анализ, какие категории лексем (например, местоимения, модальные глаголы, числительные), какие синтаксические конструкции или стилистические параметры чаще всего вызывают ошибку.

Подобный анализ требует интеграции NLP-пайплайнов с модулями морфологического, синтаксического и семантического анализа (например, spaCy, Stanza, AllenNLP), что позволяет строить объяснимые отчёты об устойчивости и локализовать «слабые зоны» модели [6].

Стоит также отметить перспективность мультиязычного тестирования, поскольку модели могут демонстрировать разную устойчивость на разных языках даже при одинаковой архитектуре. Фреймворки вроде TextFlint и Robustness Gym предоставляют возможности для генерации шума в различных языках и культурных контекстах.

Интеграция этих инструментов в пайплайн разработки – ключ к устойчивым NLP-системам. Используя такие фреймворки как MLflow, TensorBoard, DVC и Weights & Biases, можно автоматизировать тестирование на устойчивость, строить регрессии устойчивости по версиям модели, внедрять триггеры на предельно допустимый уровень отклонения, и, при необходимости, инициировать retraining или fine-tuning.

В заключение, инструменты для измерения устойчивости языковых моделей к атакующим шумамстановятся неотъемлемой частью современной практики доверенной разработки NLP-систем. Они позволяют переходить от эмпирического тестирования к формализованной валидации, от ручной отладки к автоматизированной сертификации.

Их дальнейшее развитие критически важно для построения объяснимых, безопасных и социально приемлемых ИИ-систем, которые смогут устойчиво функционировать в условиях лингвистического разнообразия, пользовательских ошибок и даже целенаправленного противодействия.

Список литературы:

1. Ebrahimi, J. HotFlip : White-Box Adversarial Examples for Text Classification / J. Ebrahimi, A. Rao, D. Lowd, D. Dou // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg : ACL, 2018. – P. 31-36. – DOI 10.18653/v1/P18-2006.

2. Alzantot, M. Generating Natural Language Adversarial Examples / M. Alzantot, Y. Sharma, A. Elgohary [et al.] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – Stroudsburg : ACL, 2018. – P. 2890-2896. – DOI 10.18653/v1/D18-1316.

3. Jin, D. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment / D. Jin, Z. Jin, J. T. Zhou, P. Szolovits // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Vol. 34, № 5. – P. 8018-8025. – DOI 10.1609/aaai.v34i05.6311.

4. Li, D. TextBugger : Generating Adversarial Text against Real-world Applications / D. Li, H. Zhang, H. Peng [et al.] // Network and Distributed System Security Symposium. – San Diego : Internet Society, 2019. – DOI 10.14722/ndss.2019.23138.

5. Li, L. BERT-ATTACK : Adversarial Attack against BERT Using BERT / L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. – Stroudsburg : ACL, 2020. – P. 6193-6202. – DOI 10.18653/v1/2020.emnlp-main.500.

6. Wallace, E. Universal Adversarial Triggers for Attacking and Analyzing NLP / E. Wallace, S. Feng, N. Kandpal, M. Gardner, S. Singh // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. – Stroudsburg : ACL, 2019. – P. 2153-2162. – DOI 10.18653/v1/D19-1221.