Т. А. Гарипов, Д. А. Морозов
Новосибирский государственный университет, г. Новосибирск
УДК 004.912

ОЦЕНКА КАЧЕСТВА ТОЛКОВАНИЙ, СГЕНЕРИРОВАННЫХ С ИСПОЛЬЗОВАНИЕМ LLM

Аннотация. Непрерывная эволюция языка требует поддержания актуальных толковых словарей. Развитие больших языковых моделей позволяет рассмотреть возможность автоматизации этой работы. Проведено сравнение сгенерированных толкований с толкованиями слов из Большого толкового словаря. Для оценки качества толкований проведен опрос среди носителей языка. К обнаруженным недостаткам генерируемых толкований следует отнести неполноту и несоответствие действительности.

Ключевые слова: генерация текста, большие языковые модели, обработка естественного языка, создание толковых словарей.

Введение. Постоянные изменения в окружающем мире оказывают влияние на разные сферы жизни, включая язык. Это проявляется особенно быстро на уровне лексики, где непрерывно появляются новые слова и заимствования из других языков, а существующие слова меняют и приобретают значения или постепенно выходят из употребления. В этом контексте толковые словари играют важную роль, предоставляя актуальные значения слов, часто с дополнительными лингвистическими пометами. Однако составляемые вручную толковые словари часто ограничены в своей способности быстро отражать изменения в языке. Их поддержка требует постоянной и тщательной работы специалистов-лингвистов, а также значительного финансирования. В качестве альтернативы может быть рассмотрено использование больших языковых моделей [1, 2]. Эти модели обучаются на огромных объемах текстовых данных и, в теории, способны улавливать особенности использования слова в различных контекстах и значениях.

Проблема исследования. В ходе работы мы стремились ответить на вопрос: насколько генерируемые толкования близки к толкованиям из широко используемых толковых словарей (в нашем случае, в качестве такого словаря выступал Большой толковый словарь русского языка [3])?

Материалы и методы. Для ответа на поставленный исследовательский вопрос мы провели серию экспериментов, сравнивая толкования из Большого толкового словаря с толкованиями, сгенерированными моделью. Выборка слов для эксперимента была сформирована на основе анализа частотных запросов пользователей в сервисе «Портрета слова» Национального корпуса русского языка [4], а также с использованием списка 100 слов из 5 жанров от специалистов-лингвистов. Для сравнения толкований использовались метрики ROUGE-1, ROUGE-2, ROUGE-L и BERTScore, с усреднением результатов по словам экспериментальной выборки. В данной работе рассматривается подход к генерации толкований слов с использованием модели языковой модели YandexGPT Pro (версия модели от 07.03.2024). Мы использовали два промпта: «Мастер-промпт» и «Диалоговый промпт». Они были разработаны в ходе предварительных экспериментов с участием экспертов-лингвистов. Полученные результаты представлены в табл. 1.

Таблица 1

Метрики нейротолкований с толкованиями из БТС

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
БТС и Мастер-промпт	0.074	0.028	0.072	0.629
БТС и Диалоговый промпт	0.071	0.019	0.068	0.628

Сгенерированные толкования значительно отличаются от словарных на лексическом уровне, что подтверждается низкими значениями метрик ROUGE. В то же время, однозначно интерпретировать полученный BERTScore нам не удалось. Поэтому мы решили провести опрос среди носителей языка, чтобы оценить, насколько генерируемые толкования корректны. Кроме того, с помощью этого опроса мы проверили, какие основные недостатки имеют различные источники толкований. Для опроса был создан телеграм бот. В нем в случайном порядке пользователю выдавалось слово (рис. 1) и спрашивалось, знаком ли респондент с этим словом.

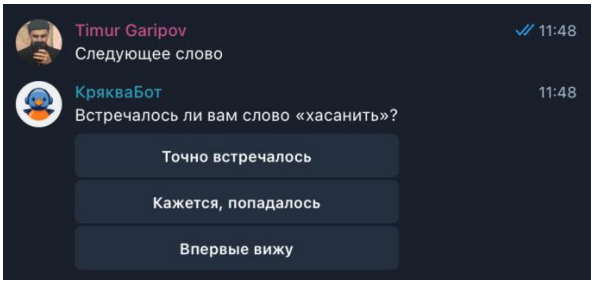

Рис. 1. Первый этап опроса для толкования

После этого респонденту нужно было оценить толкование для слова по следующим критериям (рис. 2):

- не соответствует действительности (если ранее не ответил «впервые вижу»);
- непонятно;
- неграмотно написано;
- содержит лишние слова;
- неполное;
- просто не понравилось.

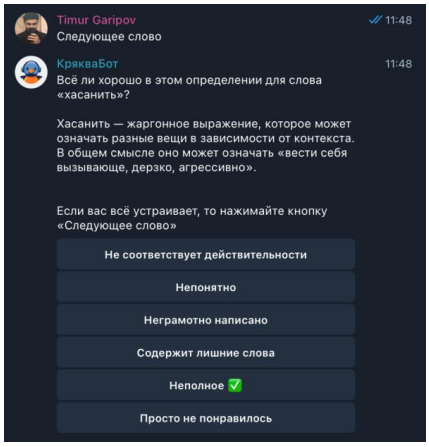

Рис. 2. Второй этап опроса для толкования

Результаты. Всего в опросе приняли участие 51 человек, в сумме толкования были размечены 3246 раз. На рис. 3 представлены полученные отклики для сгенерированных и словарных толкований.

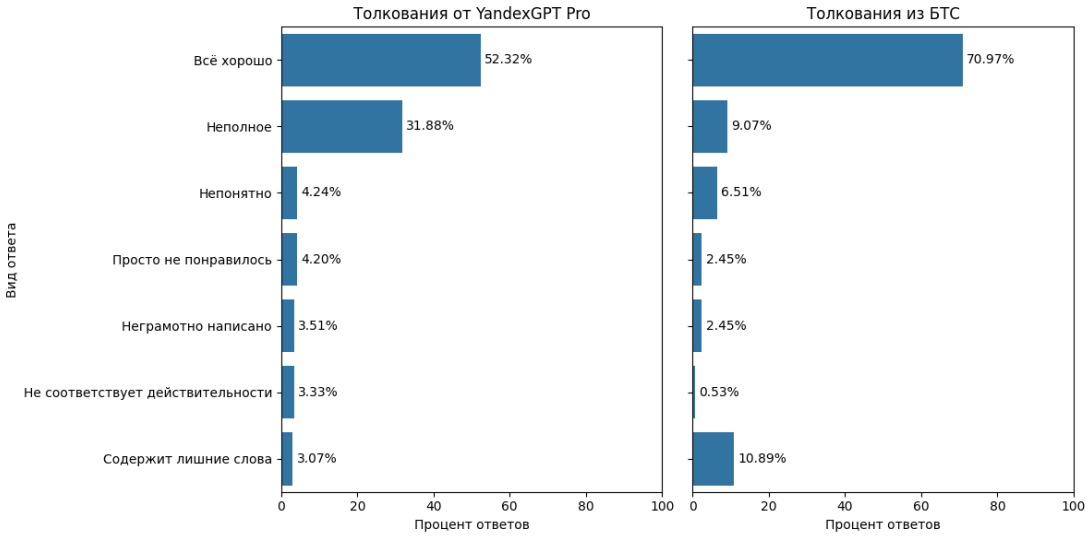

Рис. 3. Гистограммы откликов респондентов на словарные и нейронные толкования

Доля словарных толкований, не вызвавших нареканий у участников опроса, выше аналогичной доли среди сгенерированных. Это связано с тем, что они менее полные и не всегда соответствуют действительности. При этом толкования из БТС чаще непонятны и избыточны. Также проведен анализ результатов отдельно между промптами. Результаты представлены на рис. 4.

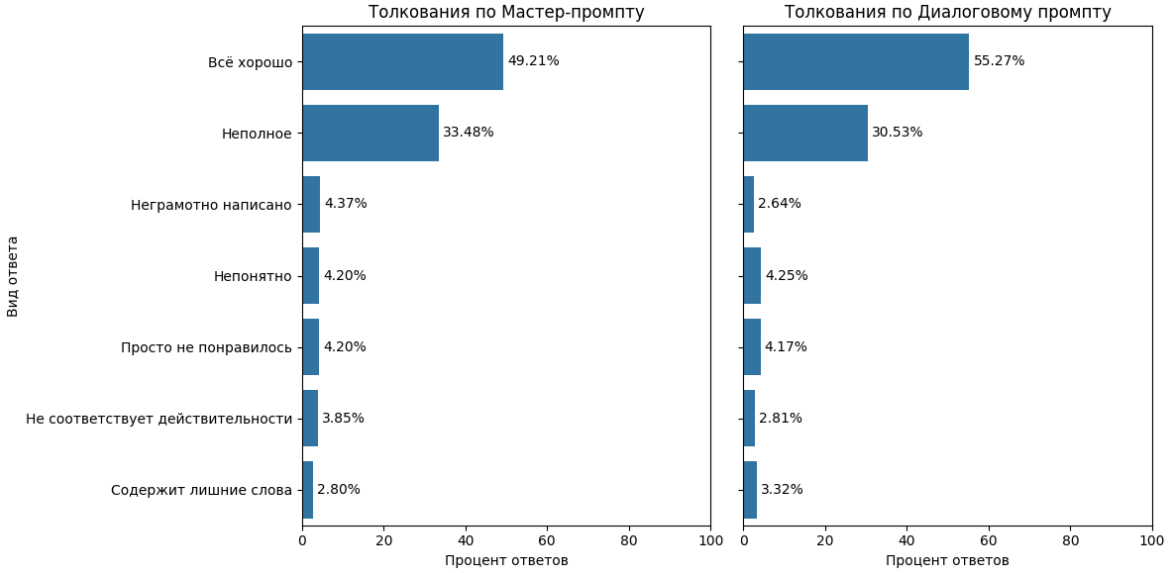

Рис. 4. Гистограммы откликов респондентов на толкования между промптами

Толкования от диалогового промпта чаще получали отметку «Все хорошо». Они оцениваются как более полные, грамотные и соответствующие действительности. В табл. 2 приведен список незнакомых слов и процент респондентов, увидевших это слово впервые.

Таблица 2

Незнакомые слова для респондентов

Слово	*Увидевших впервые, %*
жогать	66.67
парфозный	60.0
фирман	57.89
группаш	50.0
легированный	43.75
заспаунуться	40.0
темнила	40.0
аппарель	34.62
интерпункт	33.33
полудлинный	30.77
перетырка	30.0
сыскарь	28.57
кавайный	28.57
батониться	27.27
поколь	26.67

Заключение. Полученные результаты свидетельствуют о значительных расхождениях между словарными и генерируемыми толкованиями как с точки зрения автоматических метрик, так и с точки зрения носителей языка. Основным недостатком сгенерированных нами толкований следует считать их неполноту. Кроме того, нужно отметить значительную долю толкований, не соответствующих действительности, из-за чего интеграция этого источника толкований в словарные сервисы пока невозможна.

В будущем планируется провести серию экспериментов, включающих:

- использование других моделей для генерации толкований;
- создание более широких и репрезентативных наборов данных;
- оптимизацию промптов;
- дообучение моделей.

Проведение этих экспериментов позволит разработать программный инструмент для автоматической генерации толкований слов, отсутствующих в толковых словарях.

СПИСОК ЛИТЕРАТУРЫ

1. August T. Generating Scientific Definitions with Controllable Complexity / T. August, K. Reinecke, N.A. Smith. — Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers) // Association for Computational Linguistics. — 2022. — P. 8298-8317.
2. Malkin N. GPT Perdetry Test: Generating new meanings for new words / N. Malkin, S. Lanka, P. Goel, S. Rao, N. Jojic. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies // Association for Computational Linguistics. — 2021. — P. 5542-5553.
3. Большой толковый словарь русского языка / под ред. С. А. Кузнецова. — СПб.: Норинт, 1998. — 1534 с. — Текст: непосредственный.
4. Национальный корпус русского языка 2.0: новые возможности и перспективы развития / С. О. Савчук, Т. А. Архангельский, А. А. Бонч-Осмоловская [и др.]. — Текст: непосредственный // Вопросы языкознания. — 2024. — № 2. — С. 7–34.