УДК 004.89

ПРЕВЕНТИВНЫЙ АНАЛИЗАТОР БЕЗОПАСНОСТИ РОБОТОВ

Григорьев Д.С. (ФГБОУ ВО «Пятигорский государственный университет»)
Научный руководитель - кандидат экономических наук, доцент Тимченко О.В.
(ФГБОУ ВО «Пятигорский государственный университет»)

Введение. При разработке автономных систем и робототехники обеспечение надежности и эффективности планирования задач является критически важной проблемой. Возможность валидации высокоуровневых планов задач перед их выполнением может значительно снизить количество ошибок и повысить общую производительность этих систем. Традиционные методы планирования часто генерируют последовательности, которые кажутся корректными, но содержат неявные ошибки, проявляющиеся только при выполнении, что может привести к дорогостоящим сбоям, потере ресурсов или угрозам безопасности. [3, 4]

Основная часть. В работе предложена архитектура VerifyLLM для автоматической валидации высокоуровневых планов задач. Для проведения экспериментальной оценки был выбран метод Zero-Shot Planning с использованием GPT-2 Large (774M параметров) как базовая модель планирования. Тестирование планировщика с верификатором проводилось на наборе данных VirtualHome с задачами из датасета ALFRED, что позволило оценить эффективность системы на реалистичных бытовых сценариях. [1, 3]

Для оценки качества модуля перевода инструкций в формулы линейной темпоральной логики (LTL) использовались три набора данных. Первый набор ATP (Abstract Task Planning) включает 100 инструкций с временными зависимостями и логическими ограничениями. Второй набор GRM (Grounded Robotic Maintenance) содержит 44 сценария для бытовых робототехнических задач. Третий набор ES (Expert Specification) представляет собой 36 эталонных спецификаций, созданных пятью экспертами. Каждый из этих наборов содержит пары “инструкция-формула LTL”, что позволяет оценить точность перевода естественного языка в формальные спецификации. [2, 5]

В исследовании использовались следующие модели: Claude-3 Opus (52B параметров) как основная модель валидации, GPT-2 Large (774M параметров) как базовая модель для сравнения. Дополнительно проводилось тестирование с использованием моделей Qwen/Qwen1.5-0.5B, Qwen/Qwen2.5-3B и TinyLlama/TinyLlama-1.1B-Chat-v1.0.

Для оценки качества работы системы применялись различные метрики. Accuracy (ACC) измеряла точность соответствия сгенерированных и эталонных формул. Semantic Similarity Score (SSS) использовалась для оценки семантической эквивалентности формул. Operator Matching Rate (OMR) определяла точность использования временных операторов. Также применялись метрики Structural Similarity (Struct. Sim.) для оценки сохранения логической структуры планов, Order Preservation (Order Pres.) для измерения сохранения относительного порядка действий и LCS Similarity Score для оценки схожести последовательностей.

Исследование включало несколько направлений экспериментов. При анализе типичных ошибок в генерируемых планах на основе 50 случайно выбранных инструкций из датасета ALFRED были выявлены три основных типа ошибок. К первому типу относятся пропущенные действия (7-9 на план), включая отсутствие проверок безопасности и подготовительных шагов. Ко второму типу относятся лишние действия (6-8 на план), представляющие собой избыточные операции, потенциально мешающие выполнению. Третий тип составляют ошибки упорядочивания (14-15 на план), связанные с нарушением временных зависимостей.

При оценке влияния модуля верификации базовая модель GPT-2 Large показала корректность 0.27, в то время как интеграция с VerifyLLM повысила этот показатель до 0.52. Сравнение подходов к трансляции LTL формул на 15 парах инструкция-формула показало, что

“поднятый” подход обеспечивает лучшее сохранение операторов (0.89), а “заземленный” подход демонстрирует лучшую семантическую схожесть (0.81).

Исследование устойчивости системы к деградации планов показало, что при умеренной деградации сохраняются высокие показатели структурной схожести (0.86) и сохранения порядка (0.98), а при сильной деградации эти показатели составляют 0.79 и 0.97 соответственно. Анализ влияния размера окна контекста выявил, что оптимальная производительность достигается при размере 5 с F1-мерой 0.65, в то время как размеры 3 и 7 показывают более низкие результаты (F1-мера 0.58 и 0.54 соответственно).

Выводы. Эксперименты продемонстрировали эффективность предложенного метода для валидации планов задач. VerifyLLM успешно выявляет и исправляет различные типы ошибок в планах, сохраняя при этом их структурную целостность. Система показала особую эффективность при размере окна анализа в 5 действий и продемонстрировала устойчивость к различным уровням деградации планов. По сравнению с базовым методом Zero-Shot Planning, предложенный подход обеспечивает значительное улучшение корректности планов.

Список использованных источников:

1. Liang J., Huang W., Xia F., Xu P., Hausman K., Ichter B., Florence P., Zeng A. Code as Policies: Language Model Programs for Embodied Control // 2023 IEEE International Conference on Robotics and Automation (ICRA). – 2023. – P. 9493-9500. DOI: 10.1109/ICRA48891.2023.10160591.
2. Kress-Gazit H., Fainekos G.E., Pappas G.J. Temporal-logic-based reactive mission and motion planning // IEEE transactions on robotics. – 2009. – Vol. 25. – №. 6. – P. 1370-1381.
3. Huang W., Xia F., Xiao T., Chan H., Liang J., Florence P., Zeng A., Tompson J., Mordatch I., Chebotar Y. et al. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents // arXiv preprint arXiv:2201.07207. – 2022.
4. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. et al. Language Models are Few-Shot Learners // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 1877-1901.
5. McDermott D., Ghallab M., Howe A., Knoblock C., Ram A., Veloso M., Weld D., Wilkins D. PDDL-The Planning Domain Definition Language // AIPS-98 Planning Competition Committee Report. – 1998.