УДК 004.58

ПРОЕКТИРОВАНИЕ РАЗГОВОРНОГО АГЕНТА ДЛЯ ВЗАИМОДЕЙСТВИЯ С ПОЛЬЗОВАТЕЛЬСКИМ ИНТЕРФЕЙСОМ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Жуков Д. Е. (ИТМО)

Научный руководитель – доктор экономических наук, профессор Максимова Т. Г. (ИТМО)

Введение. Повсеместное распространение мобильных телефонов делает процесс понимания графического пользовательского интерфейса (далее – GUI) важной задачей. Возможность описать пользовательский интерфейс (далее – UI) на естественном языке способствует пониманию назначения UI; существует множество приложений или групп пользователей, которым такое описание будет крайне полезным. Например, начинающим пользователям бывает сложно сориентироваться в навигации UI, поэтому создаются различные справочники, направленные на обучение пользователей. Аналогично, описание скриншотов UI могут быть вставлены, как «alt-text» на веб-сайтах или маркетплейсах, предоставляя тем самым семантически-осмысленную информацию, которая помогает поисковым системам лучше ранжировать приложения или повышает доступность, например, для пользователей с слабым зрением. Техническая документация, например, спецификация application programming interface, руководство для пользователя или для разработчика, может быть улучшена, путём включения релевантного описания скриншотов. В конце концов, системы с искусственным интеллектом также нуждаются в понимании UI на уровне пользователя. В работе [1] подчёркивается важность создания разговорного агента, который понимает GUI «на уровне компьютера», что нельзя сказать про существующие агенты.

Основная часть. Цель работы – исследовать методы глубокого обучения для упрощения взаимодействия с пользовательским интерфейсом мобильных приложений при ситуативных трудностях пользователя, рассмотреть создание модели глубокого обучения и использование предобученной большой языковой модели на различных этапах выполнения работы разговорного агента. Большие языковые модели не так давно появились и статей, которые изучают их применение в задачах взаимодействия пользователя с экраном мобильных устройств, довольно мало; к тому же, сам по себе разговорный агент имеет три основных этапа: распознание намерения, когнитивная обработка (ведение диалога, понимание контекста) и генерация корректных высказываний: на каком из этих этапов пригодятся большие языковые модели, а на каких лучше справится модель глубокого обучения с иной архитектурой? Исследованы подходы к использованию моделей глубокого обучения, сформированы задачи проектируемого разговорного агента, проведён анализ подобных разговорных агентов [2]: их преимущества и концепции, используемые для ведения диалога, собраны артефакты проектирования аналогов и описаны их недостатки, рассмотрено применение GPT-3 в решении поставленных задач [3], приведена классификация разговорных агентов, их основные отличия и проблема интерпретируемости и «естественного» ведения диалога, описаны руководства для проектирования разговорного агента. Задачи, решаемые разговорным агентом, для каждой отдельной задачи [1]:

- обобщение информации на экране;
- генерация вопроса, исходя из информации на экране;
- генерация ответа, исходя из информации на экране;
- преобразование инструкций в действия пользовательского интерфейса.

Выводы. В результате анализа разговорных агентов и методов глубокого обучения спроектирован разговорный агент с применением концепций «программирование с помощью

демонстраций», «обучение по инструкциям» и больших языковых моделей. В предыдущих работах, посвященных ориентированным на задачу, независимых от доменной области разговорным агентам для задач взаимодействия с мобильным UI, акцентировали внимание на лимит системы при использовании семантических парсеров: агенты не могли различить синонимы, различные ссылки, антонимы, арифметические операции в высказываниях пользователя и многое другое. Текущее исследование подчёркивает эмерджентные свойства больших языковых моделей, которые позволяют ей создавать грамматически верные и относящиеся к элементам UI предложения, поэтому были описаны «промпты» для решаемых задач проектируемого разговорного агента по формуле: преамбула, входные данные, рассуждения в рамках «chain of thoughts», выходные данные; приводятся примеры структуры промптов и для других сценариев взаимодействия пользователя с интерфейсом. Также были «почищены» данные крупных датасетов интерфейсов мобильных приложений из-за ошибок в них, которые делали этот набор данных противоречивым.

Список использованных источников:

1. Enabling Conversational Interaction with Mobile UI using Large Language Models [Электронный ресурс] – Режим доступа: https://arxiv.org/abs/2209.08655 (дата обращения: 26.09.2023)

2. SUGILITE: Creating Multimodal Smartphone Automation By Demonstration [Электронный ресурс] – Режим доступа: https://www.researchgate.net/publication/316451742_SUGILITE_Creating_Multimodal_Smartphone_Automation_by_Demonstration (дата обращения: 30.09.2023)

3. Language Models are Few-Shot Learners [Электронный ресурс] – Режим доступа: https://arxiv.org/pdf/2005.14165.pdf (дата обращения: 27.09.2023)