УДК 004.89

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АРХИТЕКТУР РАЗГОВОРНЫХ АГЕНТОВ С ИСПОЛЬЗОВАНИЕМ ДОПОЛНИТЕЛЬНЫХ ЗНАНИЙ ИЗ СЕТИ ИНТЕРНЕТ

Апанасович К.С.
(Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

В данной работе проводится исследование генеративных моделей, которые получают дополнительные знания из сети Интернет для ведения диалога. Представлены архитектуры таких моделей и результаты экспериментов.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Введение. Разработка разговорных агентов является очень востребованной во многих областях промышленности, а также в повседневной жизни. Рассматриваемые в данной работе системы с открытым доменом не привязаны к какой-либо конкретной задаче или области знаний, а потому способны вести диалог на любые темы. Хотя современные языковые модели способны генерировать осмысленные тексты, их знания ограничены данными, на которых они были обучены.

Основная часть. Были рассмотрены три различные модели, которые тем или иным образом получают знания из сети Интернет. Модель SeeKeR [1], основанная на архитектуре Encoder-Decoder, для генерации ответной реплики последовательно выполняет три задачи: генерация поискового запроса, получение релевантных знаний из полученных документов, генерация ответной реплики. Вторая модель, WebGPT [2], основанная на архитектуре Decoder-only и обучена развернутые ответы на вопросы, путем использования заданного набора команд для нахождения необходимой ей информации в сети Интернет. Последняя модель, LaMDA [3], имеет архитектуру Decoder-only модели, для получения дополнительной информации использует сторонний набор инструментов, в число которых входит информационная ранжирующая система. С ее помощью LaMDA получает информацию из сети Интернет.

Для проведения исследований были выбраны два текстовых набора данных: Wizard of Internet [4] и Toloka Persona Chat Rus. Wizard of Internet состоит из 9633 диалогов на английском языке, в которых второй собеседник использовал поисковые запросы для получения необходимой ему информации. Toloka Persona Chat Rus – корпус из 10013 диалогов на русском языке, каждый из которых сопровождается описанием персон собеседников. Каждое такое описание представлено в виде 5 коротких предложений на каждого собеседника, например “Я рисую” или “Я живу за границей”. В корпусе Wizard of Internet описание персоны представлено только у первого собеседника в виде нескольких предложений, число которых может различаться в каждом из диалогов.

В исследовании использовалась Encoder-Decoder модель ruT5 в двух конфигурациях: на 222 и 737 миллионов параметров. Для получения дополнительных знаний из сети Интернет использовалась трехмодульная архитектура как у модели SeeKeR. Для обучения модели генерировать поисковые запросы использовались данные из корпуса Wizard of Internet. Для извлечения знаний и генерации ответной реплики – данные из корпусов Wizard of Internet и Toloka Persona Chat Rus. При этом на этапе извлечения знаний был использован метод Fusion-in-Decoder, при котором блоком Encoder полученные из сети Интернет документы обрабатываются отдельно друг от друга, а его выходы конкатенируются на входе у блока Decoder.

Выводы. В ходе исследования была разработана нейросетевая модель, которая способна получать новые знания из сети Интернет и вести диалог с собеседником на русском языке.

Список использованных источников:

1. Shuster K. et al. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion //arXiv preprint arXiv:2203.13224. – 2022.
2. Nakano R. et al. Webgpt: Browser-assisted question-answering with human feedback //arXiv preprint arXiv:2112.09332. – 2021.
3. Thoppilan R. et al. Lamda: Language models for dialog applications //arXiv preprint arXiv:2201.08239. – 2022.
4. Komeili M., Shuster K., Weston J. Internet-augmented dialogue generation //arXiv preprint arXiv:2107.07566. – 2021.