Разговорный искусственный интеллект для инклюзивного образования

Виктория Игоревна Фирсанова, Санкт-Петербургский государственный университет, аспирант 2 года обучения, st085687@student.spbu.ru, +79043372620

Разговорный искусственный интеллект — это область исследований, связанная с разработкой диалоговых систем, например, чат-ботов и виртуальных помощников. Диалоговые модели обеспечивают взаимодействие человека и машины, моделируя человеческую реакцию на текстовые стимулы. Такие агенты, как ChatGPT или Google Bard, доказали свою эффективность в решении задач образовательного (Kasneci et al., 2023), медицинского (Moons, Van Bulck, 2023), информационного (Sobania et al., 2023) и других секторов. Однако применение разговорного искусственного интеллекта в социальной сфере связано с рисками. Робастные генеративные модели требуют особого обращения (промт-инжиниринга), склонны к «галлюцинациям» (порождении фальшивой, но весьма убедительной информации) и дискриминации, вызванной предвзятостью (bias) (Kasneci et al., 2023). Исследование изучает перспективы внедрения моделей разговорного искусственного интеллекта в сферу инклюзивного образования и предлагает несколько вариантов решений перечисленных проблем на примере разработки диалогового агента для людей с расстройствами аутистического спектра.

В работе используется авторский вопросно-ответный датасет на тему инклюзивного образования и расстройств аутистического спектра, с которым можно ознакомиться по ссылке https://doi.org/10.6084/m9.figshare.13295831. Собранный мной набор данных использовался для разработки разговорной модели машинного обучения и экспериментов из следующих областей искусственного интеллекта, как науки: Data-Centric AI (Zha et al., 2023), Transfer Learning (Ruder et al., 2019), Graph Learning (Xia et al., 2021).

Набор данных, использованный в настоящем исследовании, наследует структуру датасета SQuAD 2.0 (Rajpurkar et al., 2018). Свой материал я собирала с помощью краудсорсинг-платформы Toloka. Краудворкерам предлагалось прочитать отрывки текста, извлеченные из русскоязычного веб-ресурса Autistic City (https://aspergers.ru), а затем сформулировать один или несколько вопросов к отрывкам и перечислить возможные ответы. Для оценки влияния структуры данных на производительность модели я преобразовывала дизайн датасета несколько раз в рамках экспериментирования. Более подробную информацию можно найти в моей статье "Two

Approaches to Building Dialogue Systems for People on the Spectrum". Статья была представлена на воркшопе Data Centric AI NeurIPS 2021 и доступна по адресу https://datacentricai.org/neurips21/papers/123_CameraReady_NeurIPS_2021.pdf.

Датасет использовался для тонкой настройки нескольких моделей с архитектурой Transformer (Vaswani et al., 2017) с открытым исходным кодом. Производились попытки создания извлекающих и генеративных вопросно-ответных систем с помощью добавления слоев для извлечения текста к моделям BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) и XLM-RoBERTa (Conneau et al., 2020), а также использования вопросов из датасета в качестве префикса для генеративной модели GPT-2 (Radford et al., 2019), которая должна была научиться отвечать на вопросы на основе информации из датасета с помощью мета-обучения. Демонстрация работы представлена в репозитории и пример использования модели доступны в репозитории https://github.com/vifirsanova/empi. Графовое обучение в данном исследовании является экспериментальной идеей улучшения моделей вопросно-ответных систем путем комбинирования возможностей больших языковых моделей, таких как ChatGPT или InstructGPT (Ouyang et al., 2022), с графовым машинным обучением.

Список источников:

1. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
3. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.
4. Moons, P., & Van Bulck, L. (2023). ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. European Journal of Cardiovascular Nursing.
5. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
7. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.
8. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
9. Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of ChatGPT. arXiv preprint arXiv:2301.08653.
10. Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., & Liu, H. (2021). Graph learning: A survey. IEEE Transactions on Artificial Intelligence, 2(2), 109-127.
11. Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. arXiv preprint arXiv:2303.10158.