

Шамало В.В.

магистрант кафедры информатики и вычислительной техники

Кубанский государственный технологический университет

Россия, г. Краснодар

**АНТРОПОМОРФИЗМ ДИАЛогоВЫХ АГЕНТОВ:
ОБЗОР ПОДХОДОВ К ОЧЕЛОВЕЧИВАНИЮ ОТВЕТОВ**

Аннотация: обычно пользователи дольше общаются с агентом, который отвечает естественно и ненавязчиво, поэтому разработчики стараются сделать ответы агента менее искусственными. В данной статье разобраны четыре группы таких приёмов: закрепление персоны, эмоциональный отклик, стилевая адаптация речи и дообучение по человеческим оценкам. По каждой показано, как она устроена и где перестаёт помогать.

Естественность зависит прежде всего от того, помнит ли агент разговор и держится ли одной линии, а не от богатства словаря.

Ключевые слова: диалоговые агенты, антропоморфизм, обработка естественного языка, персона, эмпатия, языковые модели.

Shamalo V.V.

Master's Student

Kuban State Technological University

Russia, Krasnodar

**ANTHROPOMORPHISM OF DIALOGUE AGENTS:
A REVIEW OF APPROACHES TO HUMANIZING RESPONSES**

Abstract: the naturalness of a dialogue agent's speech affects the duration of interaction with it, so techniques for giving responses human-like properties are widespread in industrial systems. The review systematizes four groups of them: a stable persona, an emotional response, stylistic adaptation of speech, and fine-tuning from human ratings. For each, the implementation mechanism and the

reported limitations are specified. Naturalness is shown to depend primarily on the consistency of the agent's behaviour rather than on lexical variety.

Keywords: dialogue agents, anthropomorphism, natural language processing, persona, empathy, language models.

Введение

Люди принимали программу за человека задолго до появления нейросетевых моделей. Программа ELIZA, построенная на правилах подстановки, не содержала механизмов понимания текста, однако пользователи приписывали ей осмысленное поведение и сообщали сведения личного характера [1]. При переходе к статистическим методам генерации эффект сохранился: образ участника разговора достраивается по ограниченному набору внешних признаков.

Разработчикам это важно по практической причине: чем естественнее отвечает бот, тем дольше с ним разговаривают и реже бросают диалог на середине. В статье собраны приёмы, которыми этого добиваются, и показано, где у каждого предел.

Методы и исследования

Источники отбирали по двум признакам: работы, которые объясняют, как человек воспринимает машину, и публикации, где описано конкретное решение и приведены замеры. Приёмы сгруппированы по тому, на что они влияют: на образ агента, на его реакцию, на речь или на то, как его обучали.

Результаты

Почему человек приписывает машине человеческие черты. Эпли и соавторы называют три причины: человеку не хватает другого объяснения поведению объекта, ему нужно это поведение предсказать, и ему не хватает общения [2]. Ни один из факторов не предполагает наличия у системы внутренних состояний. Данные Насса и Мун подтверждают автоматический характер переноса: участники применяли к компьютеру

нормы вежливости и взаимности, отрицая, что относится к нему как к человеку [3]. Значит, разработчик влияет не на устройство системы, а на то, как её поведение прочитает человек.

Первый приём: постоянная персона. Генеративные модели раннего поколения не сохраняли согласованности сведений о говорящем: ответы на однотипные вопросы различались в пределах одного диалога. В работе [4] предложено кодирование говорящего отдельным вектором, включаемым в модель наряду с контекстом; развитие подход получил в корпусе PersonaChat, где участнику сопоставлен набор утверждений о себе, а модель обучается порождать не противоречащие им реплики [5]. Устойчивость персоны при этом ограничена объёмом контекстного окна.

Второй приём: отклик на эмоцию. Несоответствие тональности ответа состоянию пользователя снижает оценку качества диалога заметнее, чем фактическая неточность. Корпус EmpatheticDialogues построен как набор бесед, где один участник описывает переживание, а второй отвечает с его учётом [6]. В системе XiaoIce выбор реплики оптимизируется по среднему числу ходов в сессии, принятому за меру долгосрочной вовлечённости [7]. Ограничение подхода связано с шаблонностью: набор сочувственных формулировок конечен.

Третий приём: подстройка речи. В открытых диалоговых системах изменчивость речевого поведения достигается совмещением навыков: модель обучается поддерживать тему, проявлять участие и опираться на фактические сведения, а реплика отбирается по совокупности критериев [8].

Четвёртый приём: обучение на оценках людей. Наиболее существенное изменение характера ответов связано не с архитектурой модели, а с критерием оптимизации. Процедура включает обучение на демонстрациях желаемого поведения, сбор сравнительных оценок ответов людьми и дообучение с подкреплением по построенной на этих оценках

модели вознаграждения [9]. Соответствие ответа ожиданиям пользователя оказывается заложено в целевую функцию. Побочное следствие — сглаживание: распределение формулировок сужается, что используется при построении детекторов машинного текста.

Приёмы из разных групп сочетаются: персона задаёт, кем агент себя называет, отклик на эмоцию — как он реагирует, обучение на оценках подтягивает всё остальное. Порядок внедрения при этом имеет значение. Если начать с обучения на оценках, не закрепив персону, модель научится звучать приятно и будет противоречить сама себе, а пользователь заметит именно противоречие.

Заключение

В статье разобраны четыре группы приёмов, которыми реплики агента делают человечнее: постоянная персона, отклик на эмоцию, подстройка речи и обучение на оценках людей. По каждой показано, как она устроена и где перестаёт помогать.

Приёмы различаются по характеру эффекта: воздействие на форму отдельной реплики даёт быстро заметное, но неустойчивое улучшение оценок, тогда как механизмы постоянства агента проявляются медленнее и связаны с продолжительностью взаимодействия. При ограниченных ресурсах разработки приоритет целесообразно отдавать сохранению состояния диалога. Отдельного изучения требует противоречие между следствиями дообучения по человеческим оценкам: приёмы, повышающие субъективную оценку ответов, повышают и их распознаваемость детекторами.

Использованные источники:

1. Weizenbaum J. ELIZA — a Computer Program for the Study of Natural Language Communication between Man and Machine // Communications of the ACM. 1966. Vol. 9. № 1. P. 36–45.

2. Epley N., Waytz A., Cacioppo J.T. On Seeing Human: A Three-Factor Theory of Anthropomorphism // *Psychological Review*. 2007. Vol. 114. № 4. P. 864–886.
3. Nass C., Moon Y. Machines and Mindlessness: Social Responses to Computers // *Journal of Social Issues*. 2000. Vol. 56. № 1. P. 81–103.
4. Li J., Galley M., Brockett C. [et al.] A Persona-Based Neural Conversation Model // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016. P. 994–1003.
5. Zhang S., Dinan E., Urbanek J. [et al.] Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018. P. 2204–2213.
6. Rashkin H., Smith E.M., Li M. [et al.] Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. P. 5370–5381.
7. Zhou L., Gao J., Li D., Shum H.-Y. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot // *Computational Linguistics*. 2020. Vol. 46. № 1. P. 53–93.
8. Roller S., Dinan E., Goyal N. [et al.] Recipes for Building an Open-Domain Chatbot // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 2021. P. 300–325.
9. Ouyang L., Wu J., Jiang X. [et al.] Training Language Models to Follow Instructions with Human Feedback // *Advances in Neural Information Processing Systems*. 2022. Vol. 35. P. 27730–27744.