Ryuichi Uehara

The University of Electro-Communications
Chofu, Tokyo
Japan

r-uehara@uec.ac.jp

1 Research interests

My research interest involves **persona dialogue systems**, which use the profile information of a character or real person, called a persona, and responds accordingly. Persona dialogue systems can improve the consistency of the system's responses (Li et al., 2016), users' trust (Higashinaka et al., 2018), and user enjoyment (Miyazaki et al., 2021).

My current research focuses on persona dialogue systems, especially dialogue agents that role-play as fictional characters. The first task involves obtaining the dialogue and personas of novel characters and building a dialogue corpus. The second task involves evaluating whether the dialogue agent's responses are character-like relative to the context. The goal of these studies is to allow dialogue agents to generate responses that are more character-like.

1.1 Constructing a Dialogue Corpus for Role-playing

The main focus when assessing a dialogue system that simulates a character is the accuracy with which the system reflects the character's traits in its responses. To compare the system's responses with those of the character, we need a corpus containing the character's dialogue data. Dialogue corpora related to character role-play include ChatHaruhi (Li et al., 2023), CharacterEval (Tu et al., 2024), and TimeChara (Ahn et al., 2024), which were constructed by extracting character dialogue from novels and movies. Another dialogue corpus involving the role-play of historical figures is Character-LLM (Shao et al., 2023), which generates scenes and dialogues based on Wikipedia profile information.

However, these corpora rely on external or preexisting knowledge about the characters' personas and are often limited to well-known works, extracting dialogues directly from them. For characters from popular works, personas can be inferred from external sources such as Wikipedia or assumed based on the model's parameter size. Some datasets, such as the Harry Potter Dialogue Dataset (Chen et al., 2023), include information on relationships with other characters but are restricted to a few major works.

There is a need for dialogue corpora containing characters from minor works to better assess the role-playing capabilities of large language models (LLMs) such as GPT-4, which has vast parameters and extensive training data. Current corpora mainly construct personas using data available on the Web, evaluating role-playing by comparing the model's output to the expected persona. However, such an approach may not accurately assess LLMs trained on comparable sources.

To address this gap, I focus on collecting character dialogues from novels and deriving personas directly from narrative texts and character utterances. My corpus includes not only major works but also minor ones lacking Wikipedia coverage. Persona extraction from novels allows for more authentic character representation as described by the original authors. While I manually acquired utterances and personas at first, ongoing research explores methods for automating this process using LLMs, facilitating corpus expansion. In the future, I aim to use the corpus constructed by my method to evaluate the role-playing performance of a spoken dialogue system. Since there is no definitive "correct" voice for a character in a novel, I am interested in determining the type of voice the system should select to ensure that users perceive it as matching the character's persona.

1.2 Evaluating Responses of Persona Dialogue Systems

In the assessment of the response performance of a persona dialogue system, criteria such as naturalness and fluency are important, similar to those used in open-domain dialogue systems. However, one vital evaluation pertains to whether the responses align with the designated persona.

Several evaluation methods exist for how well personas are reflected in responses. These methods use persona descriptions (Jiang et al., 2020; Zheng et al., 2020), sample monologues (Su et al., 2019; Wu et al., 2020), and evaluations without references (Miyazaki et al., 2021) and involve LLM assessments (Shao et al., 2023; Wang et al., 2024).

The first three types of methods primarily assess individual responses, which may overlook nuances where responses are contextually incongruent with the persona. For example, if a user with a persona stating "I live with

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 73–75
September 16–17, 2024. ©2024 Association for Computational Linguistics

my family" asks the system, "Do you live alone?" and the system replies, "Yes," although "Yes" alone does not contradict the persona, in context it implies that the system lives alone, which contradicts the persona.

LLM-based methods involve feeding the model persona information and calculating scores to determine if responses align with the persona. For instance, Character-LLM (Shao et al., 2023), generates prompts based on dialogue history and persona traits to evaluate memorization, values, personality, hallucination, and stability criteria.

However, a significant issue with this method is that the correlation between LLM evaluation and manual evaluation has not been consistently explored. InCharacter (Wang et al., 2024) evaluates the performance of persona dialogue systems using psychological scales focusing on personality and has confirmed a correlation with human evaluations. Nonetheless, the assessment of role-playing performance should consider factors beyond the personality reflected in responses. Aspects such as speaking style and the fidelity of character memories may also need to be correlated with human evaluations.

Human evaluation also has its drawbacks. The first is that the evaluation results vary depending on the evaluator's subjectivity and preferences. The other is that, depending on the popularity of the work, it may be difficult to recruit evaluators who know all the information about the characters (Chen et al., 2024). To address these issues, it is possible to have evaluators learn the evaluation rules and character information, but this would be a very complicated process.

Furthermore, research has indicated that GPT-4 tends to give higher ratings to text generated by the same model (Jiang et al., 2020). Typically, researchers use the best-performing model for dialogue systems and response evaluation. Consequently, when GPT-4 evaluates responses generated by itself, there arises a risk of inaccurate evaluation. Hence, it may be necessary to assess persona dialogue systems using a model other than GPT-4.

I am developing a model that takes both dialogue context and responses as input to determine whether the response aligns with the persona. To train the model, I have built a dataset consisting of pairs of responses that align the persona and those that do not. The responses that align with the persona are extracted directly from a novel, while the non-aligning responses are generated using a LLM based on the former. The dialogue context leading up to each response is also generated using an LLM. The goal is to fine-tune smaller language models so that they can provide evaluations highly correlated with human judgments.

2 Spoken dialogue system (SDS) research

To realize a persona dialogue system in a voice dialogue system, it is important to reflect the persona not only in the speech content but also in tone of voice and emotional expression. Depending on the persona, reflecting dialects and accents in the voice may also be necessary. Studies are already being conducted on changing tone and emotion during speech synthesis. With recent advances in multimodal language models, I believe it will be possible to synthesize speech that suits any persona. However, when setting up speech synthesis from a text-based persona, preventing social bias is important.

Regarding the reflection of dialects and accents, research is being conducted in speech synthesis and text translation for languages and dialects with some level of resources. Studies are also ongoing with respect to low-resource languages to overcome limited resources. In the future, this will allow a spoken dialogue agent to reproduce the dialect of any persona, essentially from any region. However, in cases where the person is from a fictional region where no model exists or is an alien, the methods proposed so far may not address the situation.

Despite some challenges, I believe that realizing a persona dialogue system in a spoken dialogue system (SDS) is a promising endeavor. This will allow for the creation of a more humanlike SDSs and is expected to further deepen the relation between dialogue systems and humans.

3 Suggested topics for discussion

I suggest discussing the following topics:

- When incorporating a persona into an SDS, what content should be considered for speech synthesis?
- Will the evolution of multimodal LLM lead to an SDS that can manage all tasks with a single model?
- What additional features would make a SDS feel more human?

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of ACL*.

Nuo Chen, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484* .

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 8506–8520. https://doi.org/10.18653/v1/2023.findings-emnlp.570.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, pages 264–272. https://doi.org/10.18653/v1/W18-5031.

Bin Jiang, Wanyue Zhou, Jingxu Yang, Chao Yang, Shihan Wang, and Liang Pang. 2020. PEDNet: A persona enhanced dual alternating learning network for conversational response generation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 4089–4099. https://doi.org/10.18653/v1/2020.coling-main.361.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597* .

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 994–1003. https://doi.org/10.18653/v1/P16-1094.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Singapore and Online, pages 178–189. https://doi.org/10.18653/v1/2021.sigdial-1.19.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 13153–13187. https://aclanthology.org/2023.emnlp-main.814.

Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019. Personalized Dialogue Response Generation Learned from Monologues. In *Proc. Interspeech 2019*. pages 4160–4164. https://doi.org/10.21437/Interspeech.2019-1696.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275* .

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976* .

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 53–65. https://doi.org/10.18653/v1/2020.acl-main.7.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pages 9693–9700. https://doi.org/10.1609/AAAI.V34I05.6518.

Biographical sketch

Ryuichi Ueahra is a master's student at the Graduate School of Informatics and Engineering, University of Electro-Communications. He is interested in persona-aware dialogue systems and role-playing agents. He participated in several competitions on building dialogue systems, including Dialogue System Live Competition 6 and AIWolfDial2024jp. He is supervised by Assoc. Prof. Michimasa Inaba.