When Crowd Meets Persona: Creating a Large-Scale Open-Domain Persona Dialogue Corpus

Won Ik Cho[*1], **Yoon Kyung Lee**[*2], **Seoyeon Bae**[2], **Jihwan Kim**[1], **Sangah Park**[3], **Moosung Kim**[4], **Sowon Hahn**[2], **Nam Soo Kim**[1]

[1] Department of Electrical and Computer Engineering and INMC, Seoul National University, Korea
[2] Department of Psychology, Seoul National University, Korea
[3] DeepNatural AI, Seoul, Korea [4] Smilegate AI, Pangyo, Korea

Abstract

Building a natural language dataset requires caution since word semantics is vulnerable to subtle text change or the definition of the annotated concept. Such a tendency can be seen in generative tasks like question-answering and dialogue generation and also in tasks that create a categorization-based corpus, like topic classification or sentiment analysis. Open-domain conversations involve two or more crowdworkers freely conversing about any topic, and collecting such data is particularly difficult for two reasons: 1) the dataset should be "crafted" rather than "obtained" due to privacy concerns, and 2) paid creation of such dialogues may differ from how crowdworkers behave in real-world settings. In this study, we tackle these issues when creating a large-scale open-domain persona dialogue corpus, where persona implies that the conversation is performed by several actors with a fixed persona and user-side workers from an unspecified crowd.

Introduction

Creating a dialogue dataset with two or more participants is a challenging process because a successful conversation between two participants requires several conditions, such as checking for common ground (Stalnaker 2002), forming rapport (Cassell, Gill, and Tepper 2007), and aligning communication style (Tannen et al. 2005). The construction becomes much more complicated when it comes to an open domain dialogue without a specific topic since the participants' common ground and interests usually differ, leading to rapid termination of the conversation.

It is difficult to use strategies adopted in task-oriented dialogues such as manual-based conversation (Wizard-of-Oz) (Wen et al. 2017) or self-play (Shah et al. 2018). Manuals may not cover the variety of topics that can appear in open-domain dialogues (Godfrey, Holliman, and McDaniel 1992; Li et al. 2017; Zhang et al. 2018), and self-play may face the limitations of content and naturalness. This inevitably calls for the participation of multiple speakers in the construction of open-domain dialogues (Dinan et al. 2019; Rashkin et al. 2019; Roller et al. 2021; Xu, Szlam, and Weston 2021).

[*]These authors contributed equally.

Our Study

Our study focuses on creating a large-scale, open-domain persona dialogue dataset, guaranteeing a safe and non-superficial conversation between the participants. Our study is three-fold:

Setting We first assume the setting in which persona participants talk with user participants (Zhang et al. 2018; Roller et al. 2020) and where users first initiate the conversation, given the detail of persona profiles. We prepare conversation guidelines for both sides, with more obligation and the leading role assigned to the persona side. Persona participants should go through an interview to be hired as an actor and participate in the conversation.

Collection After all the settings, we advertise among the worker community of a crowdsourcing platform to recruit user participants who are interested in talking with actors. Conversations are initiated with the users' message, and a web application is constructed to let participants have a conversation–and at the same time to allow the manager of the crowdsourcing platform to check on conversations for persona-user moderation.

Analysis After collecting each conversation, we let the participants fill out a survey form that asks about their satisfaction with the conversation, along with details including fun, friendliness, and connectedness. We also interview with the actors and the moderator to identify their thoughts about the conversation procedure, where they felt fun or experienced difficulties. This process provides supporting details for our strategy and makes our scheme sustainable.

Project Flow

The overall project flow is shown in Figure 1. Three stakeholders appear namely **researchers** who make up persona and user conversation guideline, a **platform (moderator)** that recruits actors and workers and instruct them to conduct persona dialogues, and the **participants** who are engaged in the construction. Users initiate all the conversations, and some conversations between actors and workers are halted after the report & reject process; otherwise, the reward is given after the survey.

Before starting the conversation, workers are informed of the instructions and prepare the dialogue by being famil-

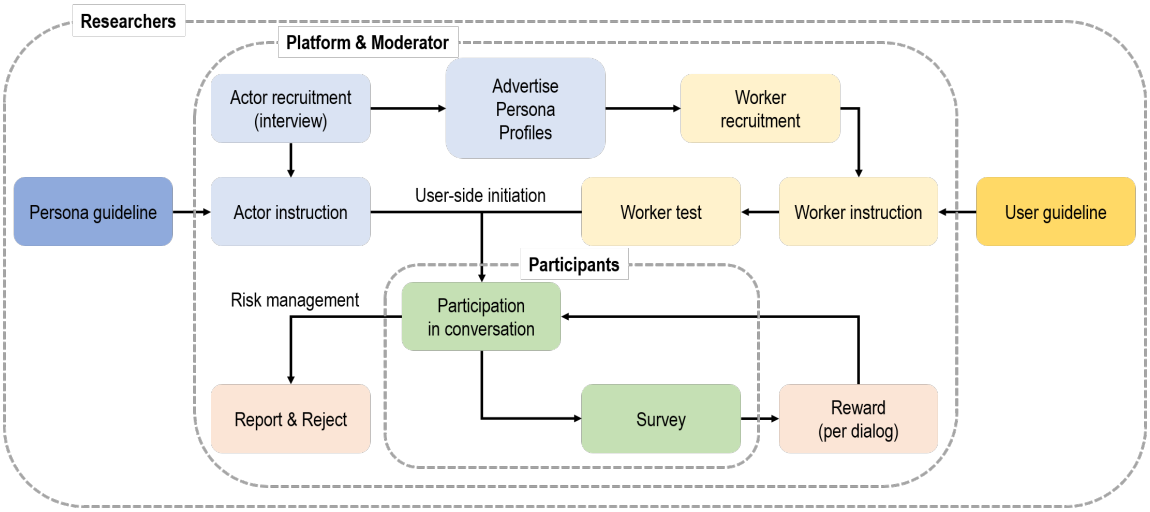

Figure 1: The overall flow of the proposed construction method. Here, *actors* denote participants regarding *persona*, and *workers* indicate who perform *user*. Specific setting of the profile is required only for actors, while conversations are initiated by workers.

iar with the introduction of the persona. The conversation is started by saying hello, and such greetings are recommended to reflect the characteristics of each persona, not just a simple greeting.

The conversation continues over 15 turns. For each turn, which consists of the worker and the actor's words, one's sentence continues until the other side chats. Such chatting is considered a single sentence regardless of the number of lines of text (which equals the number of `enters` that are pressed).

The worker terminates the conversation if they decide to quit the dialogue. Sometimes the actor terminates the dialogue if they feel fatigued or eeriness from talking with the worker. Both sides have to finish the survey after each dialogue. Workers get rewarded if they complete the survey form. The whole project ends if the actors reach the number of dialogues assigned to each, about 300 per persona.

Discussion

With above settings and collection schemes, we aim to show that our strategy helps make up a large-scale, open-domain persona dialogue corpus with a small group of actors and crowd-user participants, providing both groups with a satisfactory experience in a talk-to-earn paradigm.

RQ 1: What should be considered in accommodating the construction of a successful dialogue dataset?

Organizing persona dialogue differs a lot from usual conversations. Especially when the persona is assigned only to actors (and workers initiate the conversation), it is crucial to handle unexpected and unwanted situations that may make actors embarrassed or annoying.

RQ 2: What is the role of the moderator in large-scale dialogue dataset construction?

The usual role of moderators in natural language dataset construction is understanding the researchers' tasks and educating workers, while moderating conflicts and managing finance. In other (linguistic) annotation tasks, adjusting to the above role may lead to a successful project. However, the moderator in large-scale dialogue dataset construction may have to be concerned about the conflict between participants since dialogues inevitably involve the interaction between more than two individuals.

RQ 3: Will the above considerations help reach an intended construction process and output? Will they guarantee the satisfactory experience of workers, at the same time preserving diverse persona characteristics in the corpus?

First, from topic clustering (Van der Maaten and Hinton 2008), we could see that hiring the persona actors with a specific profile and letting them have a dialogue with various users can guarantee diversity of conversation topics. Next, from the few-shot learning example (Brown et al. 2020; Kim et al. 2021), we could check that the persona information can be encoded in the response generation process of the persona side, implying the utility of the constructed dataset in the nowadays prompting paradigm. Turning to worker experience, it was discerned from the survey results that the conversation being fun and the worker being in youth yields the attractiveness of the counterpart, and the experience of open chatting leads the counterpart to feel engagingness.

Conclusion

In this paper, we have proposed a construction scheme for a large-scale persona dialogue dataset, accompanying the building process, crowdworker interviews, and experiments for validation. We suggested three research questions with our findings, confirming that our approach can answer unanswered areas regarding participants' struggles, moderators' roles, and the output product's appreciation and evaluation. We hope our research can make a footstep to facilitating study on successful and worker-centric corpus building, which can bring satisfactory experience for all the stakeholders that participate in the project.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cassell, J.; Gill, A.; and Tepper, P. 2007. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*, 41–50.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Godfrey, J. J.; Holliman, E. C.; and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, 517–520. IEEE Computer Society.

Kim, B.; Kim, H.; Lee, S.-W.; Lee, G.; Kwak, D.; Hyeon, J. D.; Park, S.; Kim, S.; Kim, S.; Seo, D.; et al. 2021. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3405–3424.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Roller, S.; Boureau, Y.-L.; Weston, J.; Bordes, A.; Dinan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

Shah, P.; Hakkani-Tür, D.; Tür, G.; Rastogi, A.; Bapna, A.; Nayak, N.; and Heck, L. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Stalnaker, R. 2002. Common ground. *Linguistics and philosophy*, 25(5/6): 701–721.

Tannen, D.; et al. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gasic, M.; Barahona, L. M. R.; Su, P.-H.; Ultes, S.; and Young, S. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 438–449.

Xu, J.; Szlam, A.; and Weston, J. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213.

Acknowledgments

This work is supported by Smilegate AI[1]. We thank all our crowdworkers and DeepNatural AI[2] for creating high-quality data.

[1]https://smilegate.ai/

[2]https://deepnatural.ai