Measuring Subjective Persuasion Effectiveness in Symmetric Human-Agent Dialogues

Carolin Schindler[1,2][0009−0007−5819−3035], Niklas Rach[3][0000−0001−9737−8584], Yuki Matsuda[4,2][0000−0002−3135−4915], Wolfgang Minker[1][0000−0003−4531−0662], and Keiichi Yasumoto[2][0000−0003−1579−3237]

[1] Ulm University, Germany, carolin.schindler@uni-ulm.de
[2] Nara Institute of Science and Technology, Japan
[3] Tensor AI Solutions GmbH, Germany
[4] Okayama University, Japan

Abstract. To effectively change a user's behavior or attitude, it is crucial to adapt the agent's persuasive behavior and strategy to the user at hand. This requires the agent to assess its persuasive effectiveness as perceived subjectively by the individual user. In this paper, we present an empirical approach for acquiring a multimodal dataset for the task of social sensing of persuasion effectiveness in symmetric human-agent dialogues that base their persuasion on argumentation. To the best of our knowledge, there only exist datasets measuring persuasion effectiveness in asymmetric settings where the user (i. e., the persuadee) is merely a passive recipient of the content without any way to influence the interaction. Contrary to this, the persuadee in our study is an active and speaking participant in the discussion, resulting in a symmetric dialogue where both parties are engaged in mutual exchange. Thus, not only the persuadee's visual modality, as in previous work, but also the persuadee's auditory modality can be considered as a meaningful input. While the present work does not yet include the acquisition of the dataset across different cultures and nationalities, the herein discussed limitations, ethical considerations, and research directions will guide future work. Therewith, this work is laying a foundation for enhancing agent adaptability and designing more effective, culturally aware persuasive interactions in symmetric human-agent communication.

Keywords: Computational argumentation · Affective computing · Social sensing · Argument quality · Argumentative dialogue system.

1 Introduction

Audience adaption plays an important role in persuasion [10]. Therefore, a conversational agent trying to evoke a change in the user's opinion, behavior, or general attitude, for example, should adapt its persuasive behavior and strategy towards the respective user. For this adaptation to be performed online during the conversation and for the individual user at hand, the agent needs to be able

to estimate its persuasion effectiveness as subjectively perceived by the user in this moment. To this end, social signals can be utilized [13].

The herein presented work, provides an empirical approach for acquiring a multimodal dataset for social sensing of persuasion effectiveness in symmetric human-agent dialogues that base their persuasion on argumentation. Such a dialogue scenario is depicted in Figure 1, where the agent takes the role of the persuader and tries to convince the user, i.e., the persuadee. Since the dialogue is symmetric, the persuadee is an active interlocutor in the discussion and not merely a passive recipient of the persuasive content.

To the best of our knowledge, this active role of the individual persuadee is not considered in existing datasets for persuasion effectiveness. The datasets are either designed for an objective prediction of persuasiveness for the overall population [2, 4, 5, 11, 16, 18] or a group of people belonging to the same persona, for example by sharing the same personality traits or prior beliefs [1, 6, 9, 14]. Besides a description of the persona, these works base their prediction on the observation of the persuader. Opposed to this, the work in [13] considers solely the non-verbal signals elicited by the persuadee, resulting in a subjective and individual estimation. Nevertheless, their interaction is unidirectional with the persuader being an embodied dialogue system and the persuadee only perceiving its arguments. As such, the dialogue system is speaking to the persuadee, but the persuadee cannot speak to the dialogue system.

We are recording a dataset in a symmetric, bidirectional argumentation setting with both, persuader and persuadee, being active participants in the discussion. Therewith, we directly address limitations and future directions that were also discussed by the authors in [13]. Following the paradigm of symmetric multimodality [19], we embody the persuader with a human-like character. While our focus lies on recording the persuadee to estimate their subjective perception of the persuasiveness of the persuader, we include the recording of the persuader for the sake of completeness and to allow for a broader applicability of our dataset. The ground truth labels of the dataset are obtained during the interaction by having the persuadee rate different aspects related to persuasion effectiveness after each turn. A human baseline for the task is created in a post-hoc annotation setting by a third party watching and labeling the recordings of the conversation.

The contribution of this work is the presentation and discussion of a method for acquiring and labeling a multimodal dataset for the task of social sensing of persuasion effectiveness in symmetric human-agent dialogues. This method is going to be applied in future work to record, label, and analyze such a dataset for different cultures and nationalities in their respective mother-tongue. With this, our work provides the following novelties:

- The persuadee has an active role in the conversation.
- The persuasion effectiveness evolves over a multi-turn conversation and therefore is contextualized within it: Previous turns may influence the effectiveness of current and future turns.

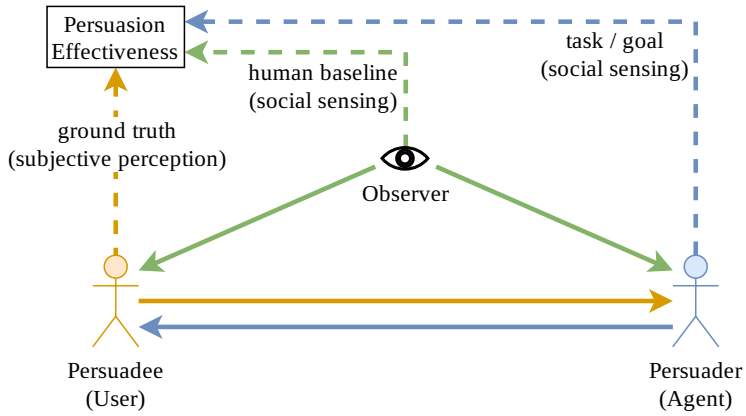

Fig. 1. Different parties' estimation of the persuasion effectiveness on the persuadee.

- For broader applicability, we do not only include the recordings of one interlocutor but of both, the persuadee as well as the persuader.
- Additionally, the dataset will be recorded in a cross-cultural manner.

These novelties will allow to investigate cultural differences and the effect of the persuadee's spoken utterances on the estimation of subjective persuasion effectiveness which are both not possible with existing datasets.

Subsequent to this introduction, Section 2 provides an overview over the scenario that our dataset is situated in. In Section 3, we present our empirical methodology for recording and labeling the data. This is followed by a discussion of limitations, ethics, and expected results, focusing on research directions and tasks that the dataset is applicable to, in Section 4. Finally, we conclude the work in Section 5 with a summary and an outlook on future work.

2 Scenario

As shown in Figure 1, we assume a scenario where a single user is interacting in a turn-wise fashion with a persuasive agent. The agent, the persuader, tries to achieve a change in the user's, the persuadee's, mental state, such as a change in attitude, for example. The persuasion effectiveness of the persuader is subjectively perceived by the persuadee and the goal of the persuader is to estimate this persuasion effectiveness to be able to adapt its behavior and strategy accordingly. Therewith, only the actual persuadee can provide the ground truth labels for our task, while the persuader's estimation of the ground truth is the respective social sensing task. A human baseline for this task, can be obtained by having a third party, i. e., an observer, perform the same task as the persuader. It is important to note that every estimation of the persuasion effectiveness is dependent on the mental model of the party making the estimation, especially

their mental model of the persuadee. This means: Based on the persuadee's mental model, the persuadee subjectively perceives the persuader and its persuasive effectiveness. At the same time, based on the persuader's mental model of the persuadee, the persuader subjectively perceives the persuadee and estimates its persuasive effectiveness. This holds for the observer, accordingly. Since the setting is a conversation with a turn-wise interaction, the influence of previous turns may be considered when performing an estimation for the current turn.

The viewpoint the dataset is taking decides whether the persuasion effectiveness is measured objectively for a group of people or subjectively for an individual person. Datasets taking the viewpoint of the persuadee, rely on the observation of the persuader to predict the persuasion effectiveness. When solely relying on this observation, the estimate is an objective one intended to generalize over the population, thereby ignoring the mental model of the persuadee. In the case of personas, the estimate additionally depends on a selected set of global features from the persuadee, such as personality traits and prior believes. The introduction of the concept of personas leads to the consideration of the persuadee's mental model in a grouped fashion. Nevertheless, going to the individual level and having an estimation of the actual persuasion effectiveness in the specific situation is not feasible with the viewpoint of the persuadee. Therefore, this work is focused on the viewpoint of the persuader, basing the estimation of the persuasion effectiveness on the observation of the individual persuadee. Moreover, reflecting human-human conversation, an agentic persuader may have the viewpoint of the persuader for estimating the persuasion effectiveness.

3 Methodology

This section describes our empirical approach for acquiring a multimodal dataset for social sensing of persuasion effectiveness in symmetric human-agent dialogues. In our dataset, the persuader is the agent, more precisely an embodied argumentative dialogue system, and the persuadee is a human user interacting with it. To investigate persuasion effectiveness in different cultures, all texts are machine translated from English to the respective mother-tongue of the participants. While we focus on the viewpoint of the persuader, we will not only include the recordings of the persuadee but, for the sake of completeness and broader applicability, also screen record the agentic persuader.

3.1 Embodied Conversational Agent

In our study, the persuader is an argumentative dialogue system, which is based on an LLM with retrieval augmentation generation (RAG). The RAG is performed on an unstructured set of argumentative sentences that are relevant to the topic of the discussion. Given a topic, the persuader takes the opposing stance of the persuadee and tries to convince the persuadee of this opposing attitude in a turn-wise interaction. To allow for a comparison to persuasion effectiveness

that was judged objectively on the basis of the text only, we build the knowledge base for the RAG based on these existing datasets. Following the paradigm of symmetric multimodality [19], which states that modalities perceived by the system should also be output by it, we embody the dialogue system by a human-like CG agent. The next utterance of the system is solely based on the textual content of the previous turns. The agent's verbalization and animation depends on its current activity: When the agent is listening or thinking, we utilize an idling animation, including breathing and blinking but no specific reactions to the conversational content at hand. When the agent is speaking, the verbalization and animation, including facial expression and gestures, is based on the utterance to be spoken. The agent will be presented both, either on a regular computer screen or the ARCADE [15] platform. The latter provides an external display that allows for an augmented reality presentation in human-like size.

3.2 Dimensions for Subjective Persuasion Effectiveness

The authors in [16] developed and validated a three-factor scale to measure perceived persuasiveness: *Effectiveness* captures whether the message will cause a change in behavior or attitude in the participant, *quality* checks for the accuracy, trustworthiness, and believability of the message, and *capability* considers the potential of the message to inspire, change, or influence users. The work closest to ours [13], focuses on the categories *convincing* and *interesting* to measure task success of competitive and cooperative argumentation tasks, respectively.

Being only interested in subjective persuasion effectiveness, we do not include the global factor *capability*. Moreover, we do not ask for labeling the factor *quality* since the overall quality of an argument is influenced by multiple aspects [18] that cannot be covered without substantially interrupting the conversation by the labeling procedure. Instead, we refer to the global rating of the quality provided by the datasets that we have extracted the arguments from. Considering the equivalent rational of *effectiveness* and *convincing*, we measure the subjective persuasion effectiveness by rating the categories *convincing* and *interesting*. While persuasion is generally viewed as a competitive task, in our study only the persuader is instructed to be persuasive and therefore we believe that rating the interestingness is still relevant in the light of user engagement and the elaboration likelihood model [12].

3.3 Data Collection and Labeling

The data collection is performed both, in-person in the laboratory and remotely with people's individual hardware. The former will allow for high quality recordings in a controlled environment, while the latter reduces data collection costs, especially in a cross-culture procedure, and is collecting noisier, closer to real-world data. The labeling of the data is performed by the persuadee for the ground truth and by a third party pretending to be the persuader for a human baseline.

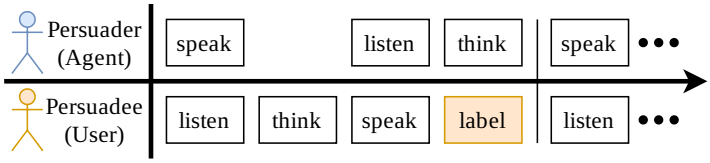

Fig. 2. Activities during the interaction of persuader and persuadee on a timeline.

Procedure The participants are sitting at a table interacting with the embodied argumentative dialogue systems as described above. Figure 2 sketches the activities during the interaction: While the agent is speaking, the user is listening and the user might continue thinking after the speech act of the agent. Afterwards, the user answers to the agent, who in turn is listening and "thinking"/processing before speaking again.

To be as close as possible to the actual perception, the ground truth labels by the persuadee are collected during the interaction and not post-hoc. Possible consistent times for the collection are after the speech act of the persuader and after the speech act of the persuadee. We only collect labels after the speech act of the persuadee to capture the listening, thinking, and reactive speech act of the persuadee without any artificial interruptions. Moreover, this allows for a labeling while the user is waiting for the agent to select and prepare its next utterance, anyways. For the in-person data collection, the labeling will be performed on a separate tablet; for the remote data collection, the participants might use an external device as well or make use of side-by-side viewing.

The labels for the human baseline are collected post-hoc, i. e., any time after the interaction has taken place. Participants in this labeling task are asked to watch the recording of the persuadee from the perspective of the persuader and provide the labels at the same time as the persuadees. There will be no option for repeated playback and the recordings are only interrupted for performing the labeling. Each participant is watching the recordings conversation-wise, allowing knowledge and observations from previous turns to be taken into account. Therewith, the setting is as proximate as can be in a post-hoc annotation to an actual interaction with the human labeler being the persuader.

Sensors We need to record the auditory and visual modality of the persuadee. For in-person experiments in the lab, we utilize a pin microphone and an RGB-D camera capturing not only the face but also the upper body of the participant from a frontal position. Having a real-world applicability in mind, we do not attach any sensors for recording physiological signals to the persuadee and refer to [20] for a survey on extracting physiological measurements from the visual modality. In the remote setting, the (presumably RGB) camera and microphone available to the participant will be utilized for recording. To assure that the recordings captured remotely will be useful for the dataset, we perform preliminary tests before starting the study like extracting the pose of the participant

from the video or extracting a spoken text from the audio recording. Additionally, we screen record the visual and auditory output of the persuader during the interaction. This recording is not directly needed for our task and is included for the sake of completeness and better analysis of the collected data.

Questionnaires Before starting, we ask all participants questions related to their persona. More precisely we gather the following information from the people participating as a persuadee:

- Their demographic data, particularly age, gender, ethnicity, and education.
- The participant's proficiency related to conversational agents by utilizing the affinity for technology interaction (ATI) scale [7].
- We capture their personality traits according to the Big-Five personality dimensions (agreeableness, conscientiousness, extraversion, openness, and neuroticism) through the ten item personality inventory (TIPI) [8].
- Their affective style by applying the Perth emotional reactivity scale (PERS) [3].
- The participant's prior stance and interest in the topic to be conversed about with the agent on a six-point scale.

For people participating as a observer, i. e., in the labeling for the human baseline, we ask the same set of questions without the ATI and the PERS scale.

The labeling of the data with respect to the subjectively perceived persuasion effectiveness, is performed by asking the persuadees to rate the last turn on a six-point scale for its *convincingness* [13, 16] and *interestingness* [13] on them, i. e., "The last turn of the agent was convincing/interesting for me". The observers are rating the same questions for the persuadees based on their recorded reaction. To assure a proper behavior of the LLM-based agent, we additionally ask if the last turn by the agent was coherent [17], i. e., whether the response was comprehensible, relevant, and appropriate.

After the conversation with the agent, we assess the persuadees' perception of the agent and the influence the interaction had on them. For the latter, we repeat the pre-questionnaire questions about their stance and interest in the topic. For the former, following the work in [11], we ask for the agent's Big-Five personality dimensions and additional high-level attributes, namely confident, credible, dominant, entertaining, expert, humorous, passionate, physically attractive, professional-looking, vivid, and voice pleasant. The post-questionnaire for the observers is shown after every conversation asking to rate the difficulty of the task and name the cues that they were using for each of the dimensions of persuasion effectiveness. Additionally, we ask for their perception of the persuadee with the same set of questions as for the perception of the agent.

4 Discussion

Limitations To have better control over the performance of the persuader, we tried to remove the persuader as a factor by investigating persuasion effectiveness in human-agent and not in human-human conversations. Still it cannot

be excluded that the LLM-based agent might hallucinate or that errors occur when translating from English to the respective mother-tongue. With our control question regarding the coherence of the agent's turn, such faulty behavior should become apparent. Additionally, the recordings of the persuader will allow a thorough post-hoc investigation of the agent's behavior. Since the knowledge base of arguments provided to the agent is the same for every study participant, independent of their ethnicity and nationality, topics are discussed on a more global level without considering arguments that are specific to a certain geographical location or geopolitical situation, for example. Moreover, the persuasive framing of the arguments and the agent's strategy are not adapted to the persuadee. While these points could impair the overall relevance and persuasiveness of the discussion, it does not directly harm our analysis of persuadee's displays of subjective persuasion effectiveness since we need a sufficient amount of positive and negative examples in our dataset anyways. The same holds for the non-verbal behavior of the agent and the capabilities of the RAG-based LLM.

Previous datasets investigated the persuasion effectiveness independently for every rating. Since we ask the participants to perform the labeling with respect to the last move of the agent, this should (at least in theory) also be possible for our dataset. However, especially the labels in the human baseline might be subject to a temporal component due to the potentially increasing familiarity with the reactions of the individual persuadee. This could be viewed as a limitation but at the same time it provides a more holistic and realistic view on estimating subjective persuasion effectiveness in the course of a symmetric conversation. Yet it has to be noted that our dataset does not provide any insights on the lasting persuasion effectiveness of the conversation. Our labeling does not go beyond the rating of the momentary perceived persuasion effectiveness by the persuadee. However, this kind of momentary perception is sufficient for a use within the scope of a conversation.

Ethical Considerations In the context of persuasion for good, our dataset can help to tailor the persuasive behavior to the specific user and better achieve a change in the user's behavior or attitude for the better. It should not be overseen that there is the potential for misusing the dataset to create a persuasion strategy that actively manipulates the user in a personalized manner. At the same time, our dataset will allow to detect how persuaded a user is and therefore could be used to train a model that makes users aware of the effect that the system they are using is having on them. With this not only malicious persuasion attempts but also unintended persuasion or manipulation could (at least to a certain extend) be uncovered.

Expected Results: Research Directions and Tasks The machine learning task that our dataset mainly targets is the estimation of the persuader's effectiveness as subjectively perceived by the persuader. To this end, various features can be extracted from the visual and auditory modality of the recordings of the persuadee. For a real-world applicability, the estimation needs to be performed

accurately and in real-time. Once this is achieved, the estimations can be utilized to enable follow-up tasks such as the respective adaptation of the agent's behavior and strategy.

After recording the dataset in a cross-cultural setup, among others the following research directions could be addressed by it:

- The investigation of cultural and individual differences of displaying convincingness and interestingness: Which combination of features are most predictive for estimating the subjective persuasion effectiveness? How does this set of features change when considering the population, a grouping by different personas, or an individual person?
- A comparison of objective and subjective persuasion effectiveness: To what extend are objective predictions a good proxy for the subjective estimation in the situation? What information can bring objective estimations closer to the observed subjective persuasion effectiveness?
- A comparison of the user's behavior in the symmetric setting of our work to the asymmetric setting in [13]. It is hypothesized that the reactions will be more descriptive in the symmetric setting than in the asymmetric one.
- The effect of the contextualization of the persuasion effectiveness in the conversation: How does the accuracy of the estimation of subjective persuasion effectiveness evolve over the course of a conversation? How does the knowledge over previous turns impact the estimation for the current turn?
- We hypothesize that convincingness implies interestingness, i. e., when the point made by an agent is not perceived as being interesting, it cannot be convincing and therefore the persuasion cannot be effective.

5 Conclusion

In the herein presented work, we have discussed an empirical approach for collecting a multimodal dataset that is capturing the subjective persuasion effectiveness in symmetric human-agent dialogues. The symmetric, bidirectional dialogue setting, where the persuader as well as the persuadee are speaking allows to not only obtain meaningful features from the visual but also the auditory modality of the persuadee. Due to the effort and costs of acquiring this dataset in a cross-cultural setting, it is important to ensure a holistic, sound, and rigorously planned methodology. In future work, we are going to record the dataset as described in this work and we will address the research directions discussed above. Going beyond the dataset, we will investigate how predicted persuasiveness and actually observed persuasiveness can interplay in the persuader's mental model of the persuadee.

Acknowledgments. This work was supported by a scholarship of the German Academic Exchange Service (DAAD).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al Khatib, K., Völske, M., Syed, S., Kolyada, N., Stein, B.: Exploiting personal characteristics of debaters for predicting persuasiveness. In: ACL (2020)
2. Bai, C., Chen, H., Kumar, S., Leskovec, J., Subrahmanian, V.S.: M2p2: Multimodal persuasion prediction using adaptive fusion. IEEE TMM (2023)
3. Becerra, R., Preece, D., Campitelli, G., Scott-Pillow, G.: The assessment of emotional reactivity across negative and positive emotions: Development and validation of the perth emotional reactivity scale (pers). Assessment (2017)
4. Chatterjee, M., Park, S., Shim, H.S., Sagae, K., Morency, L.P.: Verbal behaviors and persuasiveness in online multimedia content. In: SocialNLP (2014)
5. Djouvas, C., Charalampous, A., Christodoulou, C.J., Tsapatsoulis, N.: Llms are not for everything: A dataset and comparative study on argument strength classification. In: PCI (2025)
6. Durmus, E., Cardie, C.: Exploring the role of prior beliefs for argument persuasion. In: NAACL (2018)
7. Franke, T., Attig, C., Wessel, D.: A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. International Journal of HCI (2018)
8. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality (2003)
9. Lukin, S., Anand, P., Walker, M., Whittaker, S.: Argument strength is in the eye of the beholder: Audience effects in persuasion. In: EACL (2017)
10. O'Keefe, D.J.: Persuasion and social influence. The International Encyclopedia of Communication Theory and Philosophy (2016)
11. Park, S., Shim, H.S., Chatterjee, M., Sagae, K., Morency, L.P.: Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In: ICMI (2014)
12. Petty, R.E., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: Advances in Experimental Social Psychology. Academic Press (1986)
13. Rach, N., Matsuda, Y., Ultes, S., Minker, W., Yasumoto, K.: Estimating subjective argument quality aspects from social signals in argumentative dialogue systems. IEEE Access (2021)
14. Rescala, P., Ribeiro, M.H., Hu, T., West, R.: Can language models recognize convincing arguments? In: Findings of ACL: EMNLP (2024)
15. Schindler, C., Mayumi, D., Matsuda, Y., Rach, N., Yasumoto, K., Minker, W.: Arcade: An augmented reality display environment for multimodal interaction with conversational agents. In: Companion Proceedings of ICMI (2024)
16. Thomas, R.J., Masthoff, J., Oren, N.: Can i influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. Frontiers in AI (2019)
17. Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., Goel, R., Raju, A.: On evaluating and comparing conversational agents. In: NIPS (2017)
18. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational argumentation quality assessment in natural language. In: EACL (2017)
19. Wahlster, W.: Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In: KI (2003)
20. Wang, J., Shan, C., Liu, L., Hou, Z.: Camera-based physiological measurement: Recent advances and future prospects. Neurocomputing (2024)