Inferring Partner Models for Adaptive Explanation Generation

Amelie Robrecht
Social Cognitive Systems
Bielefeld University
arobrecht@techfak.de

Heike Buhl
Educational Psychology
Paderborn University
heike.buhl@uni-paderborn.de

Stefan Kopp
Social Cognitive Systems
Bielefeld University
skopp@techfak.de

1 Introduction

While most current approaches focus on explanations as single-turn answers to why-questions (Chandra et al., 2024; Lewis, 1986; Anjomshoae et al., 2019), we conceive them as a co-constructive process that may encompass different explanatory questions, including *What?, How?*, and *Why?* (Rohlfing et al., 2021; Axelsson et al., 2022; El-Assady et al., 2019; Lombrozo, 2006; Miller, 2019). Crucially, in human-human explanations, we see interlocutors adapting their utterances to what they think supports their partner best. Although it has long been argued that this requires a model of the partner (Clark and Wilkes-Gibbs, 1986), it is unclear, which features this partner model (PM) consists of. We argue that such a model is a dynamic and fuzzy representation of the interlocutor needed to maintain a shared understanding or grounding of a task. It is influenced by characteristics, experiences, expectations, and stereotypes (Brennan et al., 2010) and must encompass the modeling of a partner's knowledge. That is, a PM is not a box with facts, but rather a mosaic of relevant known fragments about the partner (Dillenbourg et al., 2016). This stresses the importance of inferring the correct fragments for a given task, initially and then constantly updated during the interaction.

Previous work on conversational human-agent interaction has mainly looked at adapting the interaction to the user's evolving understanding (Buschmeier and Kopp, 2018). We extend this view by considering two processes, *inferring* the assumed features in a PM from the user's behavior (Chandra et al., 2024), and then *applying* the PM in a non-stationary decision process to determine the best communicative action. Our previous work focused on applying the simplyfied PM in the decision process (Robrecht and Kopp, 2023). Now we focus on how it is inferred, both in the sense of determining relevant features and then inferring their values during the interaction. That is, we focus on the relations of features and observable information in the PM (analogous to Dillenbourg et al. (2016)). To that end, we go beyond the user's knowledge by examining four additional features hypothesized to be decisive in explanations. We employ a Dynamic Bayesian Network (DBN) to model this inference in order to keep track of the central goal of an explanation: grounding the explanandum.

2 Features of the Partner Model

If the considered features of the user go beyond pure knowledge, the explanation becomes more personal and efficient. The more explicit the state of a dimensional feature is inferred, the more informative they are as an instrument for user adaptation. The belief about a feature is dynamic, independent of the feature's invariance to time. Each feature can be tracked when receiving explicit feature-directed statements or implicitly in the course of interaction. The more meaningful explicit determination is rare, due to its higher costs, while the more fuzzy implicit determination can be executed continuously.

We hypothesize user's **expertise** E to play a significant role in tailoring the explanation to them, as it influences the depth of information required for understanding. Unlike local knowledge, expertise is considered as prior knowledge which is persistent and does not fluctuate during the interaction. Expertise can be observed through explicit user statements S_e or implicit through the frequency of positive user feedback FB_p. A high level of expertise increases the improvement of understanding, as the user already has domain-specific knowledge and can transfer structures and relationships. When adapting to this feature, the agent therefore expects a user with a higher level of expertise to understand more quickly.

Cognitive load L describes the amount of a person's limited working memory resources used in a specific task (Chandler and Sweller, 1991). We assume that adapting an explanation to the personal cognitive load is relevant in order not to leave the listener hanging or bored. Making an adaptive system sensitive to the user's cognitive load is an established approach (Khawaja et al., 2014) and linguistic measures are established in HAI (Khawaja et al., 2014; Arvan et al., 2023). Cognitive load can explicitly be derived from user statements S_L, which are considered the most reliable (Khawaja et al., 2014). Specific linguistic features, such as word count (higher load = longer sentences) (Khawaja et al., 2014), Type-Token Ratio (higher load = lower ratio) (Arvan et al., 2023), or Gunning Fog Index (Gunning, 1968; Khawaja et al., 2014) (higher load = higher index), are proven to correlate with the user's cognitive load and can be used as an implicit measure FB_c. When adapting to the cognitive load of the user, the amount of information per utterance can be varied.

We expect attentiveness to be a relevant feature, as we presume a user with high attentiveness to have a low probability of missing a given information. Just like the cognitive load, the users' **attentiveness** A can change while the explanation evolves. Although there is no explicit measure for attentiveness, there are different ways of implicit measurement: In addition to eye movement and prosody, the frequency of feedback (FB_p and FB_n) is a predictor of attentiveness (Buschmeier et al., 2011; Oertel et al., 2016). Consequently, a low level of attentiveness leads to a higher probability of fully missing an utterance when applying this feature.

According to Allwood et al. (1992), feedback can be illustrated as a ladder with four rungs: contact, perception, understanding, and attitudinal reactions. While attentiveness mainly deals with the lower levels of feedback (contact and perception), **cooperativeness** (C) represents the user's willingness to express understanding and attitude. Consequently, cooperativeness mainly considers so-called *substantive contributions* (Chi et al., 2008), where the user takes the turn. We assume cooperativeness to be an important feature as we expect a highly cooperative user to autonomously interrupt and report non-understanding. The dynamic feature of cooperativeness can be indicated by explicit utterances S_c ,or measured implicitly through the frequency of substantive feedback FB_s. When applying this feature, a higher level of cooperativeness leads to a higher improvement of understanding if no feedback is provided.

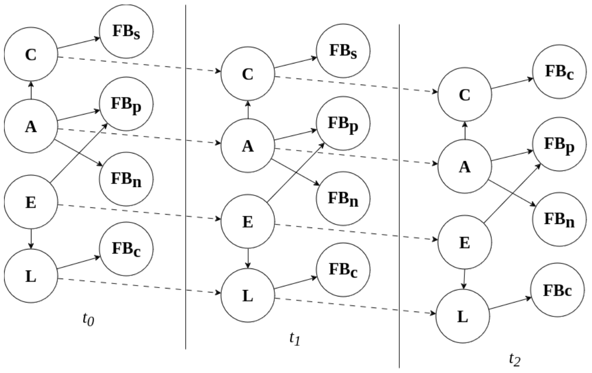

Figure 1: DBN to infer PM. Only implicit Feedback is displayed. See text for variable introduction.

Chandra et al. (2024) argue that a PM is inferred and repaired. Similarly, we look at inferring the PM using a DBN (see Fig. 1). A Bayesian Network is a graphical formalism for representing joint probability distributions, while DBNs are specifically designed to model changes over time, assuming a stationary underlying process with the previous state as a prior (Murphy, 2002). A time step always consists of an explanation move and the associated user feedback. Even if no response is given, the frequency of feedback (FB_n, FB_p, FB_s) changes and the DBN needs to be updated. Certain features are assumed to influence others: As shown in (Khawaja et al., 2014) and discussed earlier, expertise impacts the cognitive load of the user. At the same time, attentiveness is a requirement for cooperativeness (Allwood et al., 1992), which explains their dependency.

3 Discussion and Future Work

This paper explores the relevant features needed for a PM to effectively adapt an explanation. It focuses on features that go beyond pure knowledge. In a second step, it introduces a DBN as a potential tool for implementing such a PM in human-machine interaction. In a next step, the PM will be evaluated by eliminating individual features or combinations of these. The explanations created in this process will be compared with each other and with human-generated explanations, to confirm that the selected features have the hypothesized influence. Additionally, we will merge the improved PM with the current decision model (Robrecht and Kopp, 2023) and assess it in a user study.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Mohammad Arvan, Mina Valizadeh, Parian Haghighat, Toan Nguyen, Heejin Jeong, and Natalie Parde. 2023. Linguistic Cognitive Load Analysis on Dialogues with an Intelligent Virtual Assistant. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling Feedback in Interaction With Conversational Agents—A Review. *Frontiers in Computer Science*, 4:744574.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two Minds, One Dialog. In *Psychology of Learning and Motivation*, volume 53, pages 301–344. Elsevier.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative Listener Feedback in Human–Agent Interaction: Artificial Speakers Need to Be Attentive and Adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden.

Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, Stefan Kopp, and Petra Wagner. 2011. 'are you sure you're paying attention?' - 'uh-huh' communicating understanding as a marker of attentiveness. In *Interspeech 2011*, pages 2057–2060. ISCA.

Paul Chandler and John Sweller. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4):293–332.

Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Joshua Tenenbaum. 2024. Cooperative Explanation as Rational Communication. *arXiv preprint*.

Michelene T. H. Chi, Marguerite Roy, and Robert G. M. Hausmann. 2008. Observing Tutorial Dialogues Collaboratively: Insights About Human Tutoring Effectiveness From Vicarious Learning. *Cognitive Science*, 32(2):301–341.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Pierre Dillenbourg, Séverin Lemaignan, Mirweis Sangin, Nicolas Nova, and Gaëlle Molinari. 2016. The symmetry of partner modelling. *International Journal of Computer-Supported Collaborative Learning*, 11(2):227–253.

Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. 2019. Towards XAI: Structuring the Processes of Explanations. *Proceedings of the ACM Workshop on Human-Centered Machine Learning, Glasgow, UK*, 4:13.

Robert Gunning. 1968. *The Technique of Clear Writing*. New York, McGraw-Hill.

M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *International Journal of Human-Computer Interaction*, 30(5):343–368.

David Lewis. 1986. Causal Explanation. *Philosphical Papers*, 2:214–240.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38. ArXiv: 1706.07269v3.

Kevin P Murphy. 2002. Dynamic Bayesian Networks.

Catharine Oertel, Joakim Gustafson, and Alan W. Black. 2016. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances. In *Interspeech 2016*, pages 2915–2919. ISCA.

Amelie Robrecht and Stefan Kopp. 2023. SNAPE: A Sequential Non-Stationary Decision Process Model for Adaptive Explanation Generation:. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 48–58, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Hab-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728. Publisher: Institute of Electrical and Electronics Engineers Inc.