Multimodal models of repair in social human-agent interactions

Anh NGO
anh.ngo-ha@inria.fr
Inria
Paris, France

Chloé CLAVEL
chloe.clavel@inria.fr
Inria
Paris, France

Catherine PELACHAUD
catherine.pelachaud@sorbonne-universite.fr
ISIR, CNRS, Sorbonne University
Paris, France

Nicolas ROLLET
nicolas.rollet@telecom-paris.fr
Télécom Paris, SES, Institut Polytechnique de Paris,
I3-CNRS
Paris, France

ABSTRACT

People often encounter troubles in everyday conversations, prompting them to initiate repairs, which are various approaches employed to recognize and resolve those problems, fostering mutual understanding across conversational turns. However, maintaining a smooth interaction remains challenging for Conversational Agents (CAs), which are dialogue systems designed to simulate conversation with humans (including chatbots, social robots, and virtual assistants). To foster seamless human-agent interaction, the CA should be able to recognize repairs initiated by humans, utilize multimodal cues, and participate in the repair process. This article, which is an overview of our thesis research project, outlines our ongoing efforts to accomplish this objective. The initial phase involves analyzing repair phenomena in human-human interactions.

KEYWORDS

Conversational repair, multimodal, human-human interaction, human-agent interaction

1 INTRODUCTION

The intricacies of human language exhibit imperfections in everyday communication, marked by frequent problems such as speaking issues, mishearings, misunderstandings, and social norm violations, for which conversational participants continuously identify and fix these troubles to create mutual understandings across conversational turns. All those methods overtly used by human interlocutors are called repair [19]. However, there is a cost to repair, leading to avoidance of repair when the recipient chooses not to initiate repair even when necessary [7, 15], for example, in cases where the encountered problem is unlikely to have obvious consequences.

Repair has been extensively analyzed within conversation analysis and cognitive psychology. Schegloff [19] established a taxonomy of four types of repair by distinguishing between the initiator of the repair and the one that executes the repair solution, including self-initiated self-repaired, self-initiated other-repaired, other-initiated self-repaired, and other-initiated other-repaired. This article, functioning as a summary of our thesis research project, focuses on other-initiated self-repaired, commonly known as "other-initiated repair" (OIR) and more broadly as "interactive repair." This type of repair involves an explicit exchange between two participants to identify and rectify conversational problems, thereby establishing mutual understanding across the turns within a dedicated (repair) sequence [5, 4, 20]. The structure of interactive repair comprises three components, as illustrated in Figure 1: trouble source turn (T-1), repair initiation (T0), and repair solution (T+1). The repair initiation is the pivot tracing back to a trouble source and leading towards a repair solution [4, 21]. Figure 1 depicts an example of repair sequences in our scenario. The human user initiates repair through a polar question, offering a candidate understanding pivot to address the ambiguous trouble in the agent's description sentence, and the agent confirms in response. Since the data is in Dutch, the example is translated into English by DeepL[1].

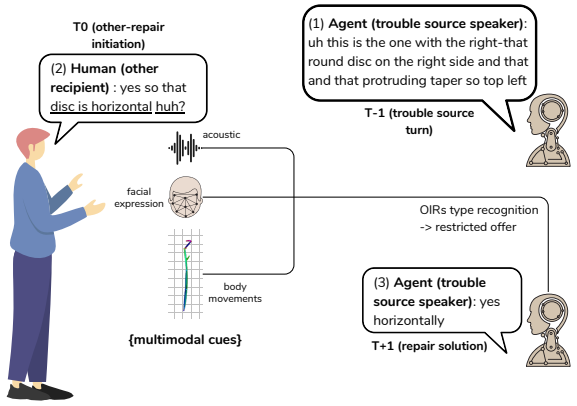

Figure 1: Example of our scenario for multimodal CA with repair capabilities, focusing on a task-oriented corpus [6, 17]

The lack of ability to smoothly repair troubles while interacting with humans limits the freedom and flexibility of CA, negatively affecting the user experience [4]. Equipping the agent with the capability to recognize the human initiation of repair and carry out repair solutions is a prerequisite to establish and maintain fluid interaction. We aim to create a CA capable of identifying when human users initiate repair and effectively implement the necessary repair strategies, thereby facilitating seamless interaction between human and CA. To achieve this primary goal, we have considered four sub-objectives:

- **SO1:** Identify the multimodal cues associated with the OIR process.

[1] https://www.deepl.com/

- **SO2:** Develop a multimodal sequential-based computational model that utilizes verbal (prosodic, acoustic features) and nonverbal (facial expression, gaze, hand gesture, head movement) cues to identify various types of OIR initiated by human interlocutors.
- **SO3:** Implement a computational model to generate the appropriate repair solution.
- **SO4:** Formulate an evaluation protocol to assess the impact of repair strategies on improving the quality of human-agent interactions.

Besides, the importance of repair effort in interactions has been highlighted in [17], where authors explored cost-efficiency in repair using co-speech gestures. We will incorporate similar considerations across our objectives, evaluating repair strategies for both effectiveness and efficiency (minimizing repair cost).

2 BACKGROUND & RELATED WORK

Dingemanse and Enfield [4] emphasized interactive repair as key to human language's flexibility, complexity, and resilience. However, existing studies in repair detection focused mainly on self-repair, in which the trouble source speakers identify and correct their utterances within the same turn, often resulting in disfluency [10, 16, 22]. These approaches exclusively relied on verbal cues (such as syntactic, semantic, and acoustic) to identify repairs, employing Natural Language Processing (NLP) methods within rule-based systems. Höhn [9] developed a rule-based chatbot with repair capabilities, incorporating three components: repair initiation recognizer through rules derived from lexical analysis, trouble source extractor, and repair solution generator based on a predefined template. Similarly, Uchida et al. [23] endowed a conversational agent with the capability to handle dialogue breakdown by employing a rule-based system to detect the repair from participants through keyword matching (including negative keywords and predefined phrases) and suggest repair solutions based on predefined scenarios.

Recent studies in conversation analysis have revealed various nonverbal cues involving bodily expressions (multimodal aspects) related to repair, particularly in interactive repair, encompassing gaze patterns, facial signals, upper body posture, and manual gestures [13, 14, 17]. For instance, Rasenberg et al. [17] observed a synchronized rise and fall in speech and hand gesture efforts across different repair types and sequential positions. Ozkan et al. [13] also found the distinctions in using visual bodily actions during the initiation of repair between people with and without hearing problems. In addition, turns at talk inherently exhibit multimodal characteristics, with construction not limited to verbal elements (words, sentences, phrases) but also integrating with or solely deriving from various modal cues like gestures or movement [8].

Moreover, recent advancements in state-of-the-art multimodal computational models have found applications in diverse areas, including emotion recognition in conversations [3, 24, 26], classification of interruptions in human interaction [25], detection of dialogue breakdowns [12], user engagement breakdown [1], trust detection in human-robot interaction [11], identification of confusion [18], and recognition of ambiguity in human-agent interaction [2]. These findings raise the potential for integrating multimodal computational models that leverage social verbal and non-verbal cues, enabling the agent to discern user-initiated repairs and implement appropriate solutions.

3 RESEARCH PLAN & SELECTED DATASET

To accomplish sub-objective **SO1**, we firstly analyzed a corpus of human-human interactions to examine verbal and nonverbal behaviors (facial expression, body movement, prosody features) associated with OIR. For **SO2**, we plan to implement and train a sequential computational model taking input features derived from **SO1**. To achieve **SO3**, our strategy involves training the computational model to take the detected OIR types as input and generate appropriate repair strategies, incorporating contextual information from trouble source turn. Lastly, to address **SO4**, we plan to conduct studies, such as post-experiment surveys, to assess the effectiveness and efficiency of repair strategies and collect user feedback.

The initial stage of OIR recognition model development involved a comprehensive analysis of both verbal and non-verbal characteristics of OIRs within human-human interactions. We selected a multimodal task-oriented corpus [17] from project CABB [6], which involves twenty dyads engaged in referential communication tasks to describe and locate 16 stimulated geometrical objects called Fribbles. The corpus provides video data from three cameras, audio recordings from head-mounted microphones, and motion-tracking data from Kinect. Setup details align with the CABB dataset and are further explained in [6].

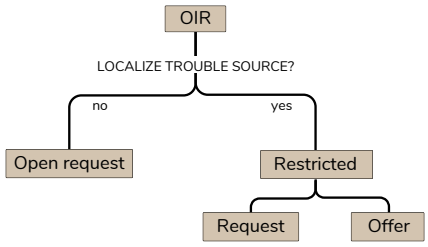

Figure 2: OIRs types annotation schema

The analysis commenced with OIR sequences annotated based on the Dingemanse and Enfield [5]'s coding schema, which classified the OIR based on whether it localized the trouble source. Three categories emerged, depicted in Figure 2: open requests (no trouble source specified in T-1), restricted requests (trouble source specified, requesting repair solution by content interrogative questions), and restricted offers (proposing a solution for confirmation by polar interrogative questions). Restricted offers dominated (83.3%) while open requests were the least frequent (6.4%) in the selected corpus.

4 CONCLUSION

In conclusion, we aim to enhance the conversational agent's ability to recognize the human user's repair initiation and produce appropriate repair solutions to facilitate seamless human-agent interaction. Due to the dataset's specific scenario, we will initially focus on the task-related communication issues (misunderstandings, mishearings) presented in this corpus. Our future work will consider broader interaction troubles.

ACKNOWLEDGMENTS

Data were provided (in part) by the Radboud University, Nijmegen, The Netherlands.

REFERENCES

[1] Atef Ben-Youssef, Chloé Clavel, and Slim Essid. 2021. Early detection of user engagement breakdown in spontaneous human-humanoid interaction. *IEEE Transactions on Affective Computing*, 12, 3, 776–787. DOI: 10.1109/TAFFC.2019.2898399.

[2] Javier Chiyah-Garcia, Alessandro Suglia, José Gabriel Pereira Lopes, Arash Eshghi, and Helen F. Hastie. 2022. Exploring multi-modal representations for ambiguity detection & coreference resolution in the simmc 2.0 challenge. *ArXiv*, abs/2202.12645. https://api.semanticscholar.org/CorpusID:247154975.

[3] Naresh Kumar Devulapally, Sidharth Anand, Sreyasee Das Bhattacharjee, Junsong Yuan, and Yu-Ping Chang. 2024. Amuse: adaptive multimodal analysis for speaker emotion recognition in group conversations, (Jan. 2024). http://arxiv.org/abs/2401.15164.

[4] Mark Dingemanse and N. J. Enfield. 2024. Interactive repair and the foundations of language. (Jan. 2024). DOI: 10.1016/j.tics.2023.09.003.

[5] Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: towards a typology of conversational structures. (Jan. 2015). DOI: 10.2478/opli-2014-0007.

[6] Lotte Eijk et al. 2022. The cabb dataset: a multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264, (Dec. 2022). DOI: 10.1016/j.neuroimage.2022.119734.

[7] Bruno Galantucci, Benjamin Langstein, Eliyahu Spivack, and Nathaniel Paley. 2020. Repair avoidance: when faithful informational exchanges don't matter that much. *Cognitive Science*, 44, (Oct. 2020). DOI: 10.1111/cogs.12882.

[8] Marjorie Harness Goodwin and Asta Cekaite. 2013. Calibration in directive/response sequences in family interaction. *Journal of Pragmatics*, 46, 1, 122–138. Conversation Analytic Studies of Multimodal Interaction. DOI: https://doi.org/10.1016/j.pragma.2012.07.008.

[9] Sviatlana Höhn. 2017. A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, (Eds.) Association for Computational Linguistics, Saarbrücken, Germany, (Aug. 2017), 395–405. DOI: 10.18653/v1/W17-5547.

[10] Julian Hough and Matthew Purver. 2014. Strongly incremental repair detection. In Association for Computational Linguistics (ACL), 78–89. ISBN: 9781937284961. DOI: 10.3115/v1/d14-1009.

[11] Marc Hulcelle, Léo Hemamou, Giovanna Varni, Nicolas Rollet, and Chloé Clavel. 2023. Leveraging interactional sociology for trust analysis in multiparty human-robot interaction. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (HAI '23). Association for Computing Machinery, New York, NY, USA, 484–486. ISBN: 9798400708244. DOI: 10.1145/3623809.3623973.

[12] Wookhee Min, Kyungjin Park, Joseph B. Wiggins, Bradford W. Mott, Eric N. Wiebe, Kristy Elizabeth Boyer, and James C. Lester. 2019. Predicting dialogue breakdown in conversational pedagogical agents with multimodal lstms. In *International Conference on Artificial Intelligence in Education*. https://api.semanticscholar.org/CorpusID:150378290.

[13] Elif Ecem Ozkan, Patrick G.T. Healey, Tom Gurion, Julian Hough, and Lorenzo Jamone. 2023. Speakers raise their hands and head during self-repairs in dyadic conversations. *IEEE Transactions on Cognitive and Developmental Systems*, 15, (Dec. 2023), 1993–2003, 4, (Dec. 2023). DOI: 10.1109/TCDS.2023.3254808.

[14] Kati Pajo and Minna Laakso. 2023. Comparing timing of other-initiation of repair: a multimodal approach. *Frontiers in Communication*, 8. DOI: 10.3389/fcomm.2023.1173179.

[15] Alison Pilnick, Rebecca O'Brien, Suzanne Beeke, Sarah Goldberg, and Rowan Harwood. 2021. Avoiding repair, maintaining face: responding to hard-to-interpret talk from people living with dementia in the acute hospital. *Social Science and Medicine*, 282, (Aug. 2021). DOI: 10.1016/j.socscimed.2021.114156.

[16] Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10, (Apr. 2018), 425–451, 2, (Apr. 2018). DOI: 10.1111/tops.12324.

[17] Marlou Rasenberg, Wim Pouw, Asli Özyürek, and Mark Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, 12, (Dec. 2022), 1, (Dec. 2022). DOI: 10.1038/s41598-022-22883-w.

[18] Mao Saeki, Kotoka Miyagi, Shinya Fujie, Shungo Suzuki, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022. Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent. In vol. 2022-September. International Speech Communication Association, 3988–3992. DOI: 10.21437/Interspeech.2022-10075.

[19] Emanuel A. Schegloff. 2007. *Sequence organization in interaction : a primer in conversation analysis I*. Cambridge University Press, 300. ISBN: 9780521825726.

[20] Emanuel A. Schegloff. 2000. When 'others' initiate repair. *Applied Linguistics*, 21, 205–243, 2. DOI: 10.1093/applin/21.2.205.

[21] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53, (June 1977), 361, 2, (June 1977). DOI: 10.2307/413107.

[22] Elizabeth Shriberg, John Bear, and John Dowding. 1992. Automatic detection and correction of repairs in human-computer dialog. In Association for Computational Linguistics (ACL), 419. DOI: 10.3115/1075527.1075628.

[23] Takahisa Uchida, Takashi Minato, Tora Koyama, and Hiroshi Ishiguro. 2019. Who is responsible for a dialogue breakdown? an error recovery strategy that promotes cooperative intentions from humans by mutual attribution of responsibility in human-robot dialogues. *Frontiers Robotics AI*, 6, APR. DOI: 10.3389/frobt.2019.00029.

[24] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. 2021. Article robust multimodal emotion recognition from conversation with transformer-based crossmodality the title fusion. *Sensors*, 21, (July 2021), 14, (July 2021). DOI: 10.3390/s21144913.

[25] Liu Yang, Catherine Achard, and Catherine Pelachaud. 2022. Multimodal classification of interruptions in humans' interaction. In Association for Computing Machinery, (Nov. 2022), 597–604. ISBN: 9781450393904. DOI: 10.1145/3536221.3556604.

[26] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. Telme: teacher-leading multimodal fusion network for emotion recognition in conversation, (Jan. 2024). http://arxiv.org/abs/2401.12987.