Modelling Proactive Voice Assistants For Collaborative Settings Using Conversational Monitoring

Leon Reicherts[1], Yvonne Rogers[1], Ethan Wood[1]

[1] University College London, London, United Kingdom

Keywords: Voice Assistants, Proactivity, Multi-Modal Interface, Data Visualisation, Conversation Analysis, Social Awareness, Machine Learning.

The goal of context-aware intelligent systems is to help users perform actions while supporting them in accomplishing various tasks. Intelligent Personal Assistants (IPA), such as Alexa or Siri, which are "incarnated" in various smart devices (e.g. smartphones, smart speakers, laptops), have been designed to provide such help, but so far, only by being reactive (i.e. responding to user commands) and for simple tasks (e.g. guiding a user through a cooking recipe). Current research is concerned with developing spatial AI, that tries to understand the setting more, by tracking location, movement and presence of people, in order to become more *adaptive* and even *proactive* (e.g. "There is some breaking news about Brexit. Would you like to hear about it?"). As part of the proactive agenda a key concern is to decide when is an opportune or good time to interrupt users. For this, one approach is to analyse what is being said in the conversations that can be heard in the background [1].

So far, much of the focus on how to make personal assistants proactive in their interactions with people is for single user scenarios, by determining when is a good or bad time to make an announcement or notify the user of an event. To do this without being annoying, the personal assistant needs to be able to "know" *when* to provide *which* information. In contrast, we are interested in the potential benefits of using intelligent assistants for multi-user scenarios, such as taking part in a meeting. While voice apps have begun to emerge to be used in work settings (e.g. Amazon's recent *Alexa for Business* [2]), they have been designed primarily for single user interactions. How might these be used by multiple people present rather than being a personal device for one user?

The situation for multi-user situations, however, is likely to be complex; in particular they need to be more "aware" of the social context, as inadequately timed proactive interventions (e.g. recommendations) could disturb or even "disrupt" groups working. While humans are typically relatively good at knowing when to give information X to person Y, it is much more challenging for machine learning (ML) models, which are based on the relevant social and behavioural indicators related to different social contexts, to know when to act.

How might features of the ongoing conversation within a group be used to enhance the modelling? We propose that the social context can be (partially) modelled by the meta-linguistic context of a conversation, with regards to speakers' turn-takings [3], prosodic features, a dominant participant and pace (c.f. *SocioPhone* [4]). Even when decoupled from a conversation’s content, such patterns can serve as good indicators of states and processes of human-to-human interaction, such as *participation balance* (integrated in the *Reflect* [5] tabletop system). By monitoring patterns of the users’ conversation, machine learning models may be able to improve the timing of their proactive interventions over time. The types of interventions may also be fine-tuned, such as making the right kind of suggestions or hints and in the long run be able to steer the conversation in a particular direction if it appears to be going off topic for too long.

Hence, building ML models of the social context seems promising, as they could be reused for different collaborative tasks/situations and could possibly be integrated in other (existing) systems, which are used for collaboration. Furthermore, it would be interesting to test the extent to which a ML model allows an assistant to naturally intervene in a conversation without having to engage in full-on natural language processing.

Questions related to the rules of conversational systems have already been investigated for AI conversational interfaces, but they typically address single-user scenarios. A classic example is *Eliza* by Weizenbaum [6], arguably the first chatbot, which gave users the illusion of understanding what they were typing in at the keyboard. However, the system’s replies were stock phrases, that transformed the user’s input based on certain rules (e.g. depending on specific keywords). The model that we envision for our ‘collaborative assistant’ is intended to be more advanced; one that can learn over time what to say, and which can give people the feeling that it understands where they need help (while it selects a suggestion from a given set of utterances if the meta-linguistic context matches certain patterns).

Our proposed ML model of social context will mainly build upon aspects of conversation analysis basic, and in particular the analysis of Turn Constructional Units [3]. In addition, it will take into account the significance of different types of silence during a conversation [7]. For this, we plan to use machine learning algorithms to classify turn-taking patterns based on *speaker diarization*, using Zhang et al.’s [8] *unbounded interleaved-state recurrent neural networks (UIS-RNN)*. In addition to the timing of responses, the model will enable the voice assistant to make the conversation more human-like by adding certain conversational elements depending on the current dynamics in the conversation. For instance, as noted by Leviathan and Matias [9] in their discussion on the design of Google’s *Duplex*, their voice assistant can sound more familiar and natural to users if speech disfluencies (e.g. hmm’s and uhuh’s) as well as synthetic waits/latency are added.

Our envisioned collaborative assistant could even go beyond providing suggestions related to the specific task at appropriate times. To this end we will explore whether context-awareness may enable the assistant to take on a moderating or mediating role in the conversation. Based on the detected patterns (e.g. large gaps between speaker activity), the interface prototype will trigger predefined utterances (e.g. “Do you need a hint for analysing X?” or “Shall we move on?”).

The kind of feedback and how they the assistant speaks may also be critical. A study exploring how a tabletop display showing who was talking during a group’s conversation around the table – in the form of growing LED lights – showed how this kind of subtle feedback was effective at letting the group know how much each was participating in a conversation [4]. Empirical studies of its use showed it was successful at raising awareness as well as enabling more balanced interaction. We hypothesise that if a voice assistant provides feedback about who is making contributions to the ongoing conversation, but in the form of verbal feedback, it could be more effective than using visual ambient feedback since the assistant has the potential to intervene in various ways, such as through nudging, guiding or prompting group interactions.

ML models that can trigger suggestions/hints could also be used for building computer-based systems to support collaborative tasks in more structured ways – such as analyses of specific datasets/data types in business meetings or educational scenarios – for which the context of the discussed data/topic is predefined and sufficiently well circumscribed. There is an increasing number of intelligent conversational data analysis tools that could be used for this purpose, based, for example, on chatbots [10, 11]. They could be used to guide users through a series of steps and, in doing so, make the process of data analysis more transparent. *Eviza* by Setlur et al. [12] allows users to interact with a data visualization, using their voice. The use of natural language seems particularly effective, as it requires verbalizations of thoughts, which can encourage users to reflect on their learning and problem-solving through 'slowing down' their thoughts [13].

To explore different proactive behaviours based on the indicators of conversational interactions, as described above, we have created a simple collaborative data exploration task and interface prototype, which involves predefined recommendations (i.e. some of them are task-related, other aim to mediate/moderate conversation) that aim to stimulate the conversation between users. The tool is designed for explorative data analysis and assists groups of users by giving suggestions of what may be interesting to explore in relation to a specific visualisation (e.g. "Did you consider the difference between variable X and Y", where X and Y could be a demographic like age or gender). Users are provided with a set of commands, which they can use to ask the system to display other subsets of the data visualisation (e.g. "Show men and women"). Sometimes, the assistant asks the group a question related to the different comparisons that are possible or by moderating the discussion (e.g. "What do you both think of this?"). The former utterances are predefined for each visualisation/the possible comparison of subsets. The time at which when these utterances are played by our voice assistant with a group is currently through using a Wizard of Oz method (where a human pretends to be the system). Based on the findings of studies conducted with this set-up, we are now investigating how to replace the human Wizard using a ML model. Once a functioning version of this assistant is developed, further user studies will be conducted to understand how well the ML models perform. The findings of these user studies would allow us to provide early design recommendations for proactive moderating/mediating behaviours of IPAs in collaborative contexts. In sum, by providing more social context awareness, based on an analysis of how conversations are progressing, may enable collaborative voice assistants to more effectively interact with groups of people without disturbing or disrupting ongoing human-to-human interactions and discussions.

References

1. McMillan, D., Loriette, A., Brown, B.: Repurposing Conversation: Experiments with the Continuous Speech Stream. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 3953–3962. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702532.
2. Akersh, S.: Delivering Voice-Powered Business Analytics with Alexa for Business : Alexa Blogs, https://developer.amazon.com/de/blogs/alexa/post/8df2b229-c0f4-4333-85dc-131d6bffeb83/delivering-voice-powered-business-analytics-with-alexa-for-business.
3. Selting, M.: The construction of units in conversational talk. Lang. Soc. 29, 477–517 (2000). https://doi.org/10.1017/S0047404500004012.
4. Lee, Y., Min, C., Hwang, C., Lee, J., Hwang, I., Ju, Y., Yoo, C., Moon, M., Lee, U., Song, J.: SocioPhone: Everyday Face-to-face Interaction Monitoring Platform Using Multi-phone Sensor Fusion. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services. pp. 375–388. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2462456.2465426.
5. Bachour, K., Kaplan, F., Dillenbourg, P.: An Interactive Table for Supporting Participation Balance in Face-to-Face Collaborative Learning. IEEE Trans Learn Technol. 3, 203–213 (2010). https://doi.org/10.1109/TLT.2010.18.
6. Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine. Commun ACM. 9, 36–45 (1966). https://doi.org/10.1145/365153.365168.
7. Clift, R.: Conversation Analysis. Cambridge University Press (2016). https://doi.org/10.1017/9781139022767.
8. Zhang, A., Wang, Q., Zhu, Z., Paisley, J., Wang, C.: Fully Supervised Speaker Diarization. ArXiv181004719 Cs Eess Stat. (2018).
9. Leviathan, Y., Matias, Y.: Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone, http://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html, (2018).
10. Fast, E., Chen, B., Mendelsohn, J., Bassen, J., Bernstein, M.S.: Iris: A Conversational Agent for Complex Tasks. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18. pp. 1–12. ACM Press, Montreal QC, Canada (2018). https://doi.org/10.1145/3173574.3174047.
11. John, R.J.L., Potti, N., Patel, J.M.: Ava: From Data to Insights Through Conversations. In: CIDR (2017).
12. Setlur, V., Battersby, S.E., Tory, M., Gossweiler, R., Chang, A.X.: Eviza: A Natural Language Interface for Visual Analysis. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. pp. 365–377. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2984511.2984588.
13. Ahlum-Heath, M.E., Di Vesta, F.J.: The effect of conscious controlled verbalization cognitive strategy on transfer in problem solving. Mem. Cognit. 14, 281–285 (1986). https://doi.org/10.3758/BF03197704.