Crowdsourcing for Reminiscence Chatbot Design

Svetlana Nikitina
University of Trento and
Tomsk Polytechnic University
svetlana.nikitina@unitn.it

Florian Daniel
Politecnico di Milano, DEIB
florian.daniel@polimi.it

Marcos Baez, Fabio Casati
University of Trento and
Tomsk Polytechnic University
baez@disi.unitn.it, fabio.casati@unitn.it

Georgy Kopanitsa
Tomsk Polytechnic University
kopanitsa@tpu.ru

Abstract

In this work-in-progress paper we discuss the challenges in identifying effective and scalable crowd-based strategies for designing content, conversation logic, and meaningful metrics for a reminiscence chatbot targeted at older adults. We formalize the problem and outline the main research questions that drive the research agenda in chatbot design for reminiscence and for relational agents for older adults in general.

Context & Objectives

Reminiscence is the process of collecting and recalling past memories through pictures, stories and other *mementos* (Webster and Gould 2007). The practice of reminiscence has well documented benefits on social, mental and emotional wellbeing (Subramaniam and Woods 2012; Huldtgren et al. 2015), making it a very desirable practice, especially for older adults. Research on technology-mediated reminiscence has advanced our understanding into how to effectively support this process, but has reached a limit in terms of the approaches to support more engaging reminiscence sessions, effectively elicit information about the person, and extend the practice of reminiscence to those with less opportunities for face to face interactions.

In our previous work (Nikitina, Callaioli, and Baez 2018) we made a case for conversational agents in this domain, and proposed the concept of a smart conversational agent that can drive *personal* and *social* reminiscence sessions with older adults in a way that is engaging and fun, while effectively collecting and organising memories and stories. The idea of conversational agents for older adults is not new, and they have been explored to support a wide variety of activities and everyday tasks (Tsiourti et al. 2016a; Vardoulakis et al. 2012; Hanke et al. 2016; Tsiourti et al. 2016b), to act as social companions (Ring et al. 2013; 2015; Demiris et al. 2016) and even to engage older adults in reminiscence sessions (Fuketa, Morita, and Aoe 2013).

While these works give us valuable insights into the opportunities of using conversational agents as an instrument to support reminiscence sessions, they also show

us how limited our knowledge is in terms of effective strategies to maintain dialogs with older adults. Success stories are mostly limited to Wizard of Oz evaluations (Schlögl, Doherty, and Luz 2014), in which system functionality is partially emulated by a human operator, or based on fully human-operated agents. The few attempts at autonomous agents highlight issues with the mismatch between user expectations and the actual social capabilities of the agents (Tsiourti et al. 2016a), general challenges with designing conversations suitable to the target population (Yaghoubzadeh, Pitsch, and Kopp 2015), and challenges with engaging older adults in question-based interactions in particular (Fuketa, Morita, and Aoe 2013).

In this position paper we aim at identifying effective and scalable crowd-based strategies for designing content, conversation rules, and meaningful metrics for a reminiscence chatbot targeted at older adults. We build on the concept introduced in (Nikitina, Callaioli, and Baez 2018) and identify *where* and *how* crowdsourcing can help design and maintain of an agent-mediated reminiscence process, while addressing the specific challenges posed by the target population.

Reminiscence Chatbot

The envisioned chatbot is based on the idea of automatically guiding older adults through multimedia reminiscence sessions (Nikitina, Callaioli, and Baez 2018). It has the dual purpose of i) collecting and organising memories and profile information, and ii) engaging older adults in conversations that are stimulating and fun. In Figure 1 we show an example conversation and related main actions.

The example starts from the subject (the elder) providing a memory in the form of a picture. In response, the chatbot poses a contextual question. In order to do so, it must be able to *understand the theme of the picture* (big city) and to *extract and understand information* from pictures and text. In order to keep the conversation natural, it must further be able to *reference related conversation topics* (the city of Trento) and, in order to show empathy, it must be able to *sense the feelings of the subject* as the conversation evolves (e.g., it looks like the subject likes rock music, so it could be an idea to talk about that for some time). It would also be good if the bot be able to *sense the presence of peers* (e.g., family
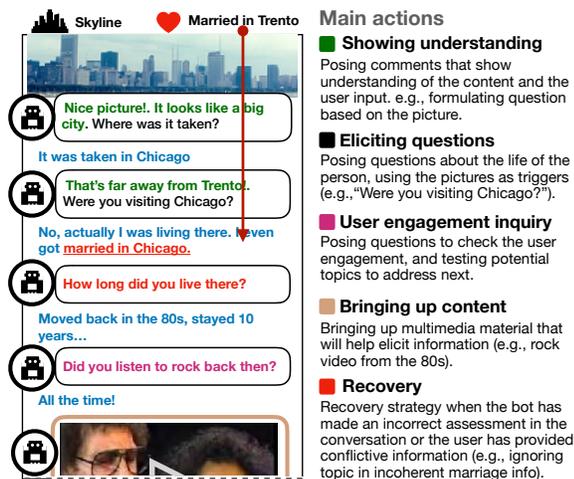
Figure 1: Example reminiscence session with bot actions

members or moderators helping with the chat). All this information helps the bot *decide on appropriate next actions* taking into account possible *conversational goals* (e.g., elicit basic user profile data). Among the most complex decisions to be taken is deciding if and when to *change context* in a conversation (e.g., to make the elder laugh).

All these requirements are particularly challenging since special attention must be paid to the subject's abilities and limitations (Nurgalieva et al. 2017; Hawthorn 2000). For instance, it is hard to *cope with user-initiated context switches* or to *keep knowledge about subjects coherent* due to cognitive decline associated with age (Park, O'Connell, and Thomson 2003). Coping with these challenges is difficult even for humans (Miron et al. 2017).

In the long term, our goal is to develop a crowd-powered chatbot that implements the necessary conversational logic, sensibility and tricks to engage older adults in pleasant and satisfactory reminiscence sessions. The crowd should not be involved in direct interactions with the elderly (like in some real-time crowdsourcing approaches studied in literature (López et al. 2016; Ring et al. 2015)), nor should it be used just to train black-box AI algorithms. The idea is to involve the crowd to elicit and represent reminiscence-specific conversation knowledge explicitly in the form of some dedicated model, in order to be able to actively steer the conversation into specific directions (e.g., to elicit health issues or family memories). In this paper, we focus on an intermediate set of research objectives: identifying (i) how to *model* the conversational knowledge the chatbot may rely on and (ii) how to use the crowd to *learn and evaluate* the model.

Crowd-Supported Chatbot Design
Conversational Model Representation
Conceptually, a simple model we can imagine for a chatbot is a state machine (S, A, δ, π, F), where S denotes the states (a state includes the information on the subject and the conversation history), F denotes the final states, A is the set of (conversational) actions, δ is a state transition function (our

conversational policy), $\pi : S \times A \to \{(s, p)\}$ associating to each state and action a set $\{(s, p)\}$ of possible target states s and the probability p with which that action should be chosen (to model that conversations are not deterministic).

In practice however the state space is infinite and the possible conversations are also infinite so this FSM is not the right model. An alternative model is based on Event-Condition-Action (ECA) rules, where the event for example is the sentence by the subject (the elder) and the condition is some expression over what we know about the subject as well as past events. This has however the same limitations just discussed.

We observe that what we really want to have is a definition of the domain and range of the policy function π so that we can learn a useful policy that can be applied to real life conversations. On the action side (the range), we approach the problem by clustering similar actions along several dimensions, such as i) the type of actions (ask information, make a comment, show interesting content) and ii) the topic of conversation (talk about the picture you are showing, or about childhood, or about hobbies). Given the action type and topic, there are many actual conversations and utterances, but at this level we are focused on learning types and topics rather than conducting an interaction within a topic or paraphrasing sentences.

In terms of the domain a policy is defined on, what we wish to have is a description of the characteristics of the state (or event and condition) to which the policy applies. For example, the crowd may tell us that after they learn the date of birth, they show newspaper covers of that year, or famous people born the same day, or songs that where popular when the subject was very young. In this case the trigger of the action is the last conversation element where the subject is notifying the state of birth (or, in terms of events, it is the event of the system, somehow, coming to know the date of birth of the person).

The challenge here is therefore to understand what is the reasoning of crowd workers when they decide to take actions, and based on this reasoning identify the classes of state and event information we need to attach policies to.

Crowdsourcing tasks

The counterpart of the model is the learning process, which has to do with how to design and process the results of crowdsourcing tasks. The objective we have in seeking the proper task designs are the following: (i) identifying *action types and topics* (unless we want to fix them a-priori), (ii) identifying *when* (based on which state or trigger) a person changes topic or shows specific content, and (iii) identifying *why* (based on which state or trigger) the agent initiates a conversation on a topic.

To do this, we envision crowdsourcing tasks that aim at (i) exploring possible conversations (these can be Wizard of Oz simulations), (ii) reflecting over previous conversations by the same worker or other workers to derive the "rules" that made the worker take a certain course of action, and (iii) aggregating these "rules" into a smaller coherent set that reveals the characteristics that the policy model should have.

For example, the crowd may reveal that they change topic whenever they sense that the person is sad talking about the current topic. This would tell us that an important component of the policy domain is the perceived emotional state, something that therefore the agent should try to detect, and that change in this emotional state should be a trigger to either continue or change topic.

We thus focus on the following research question (**RQ**): *Which crowd-based strategies can help elicit effective conversation logic for conversations (reminiscence sessions) targeting older adults, and how?*

Conversational logic includes understanding of: composition of Dialog State, when and how the State has to be changed, and what are the most important variables that affect the state. That is, given:

- the set of States $S = \{S_1, S_2, ... S_m\}$, where S is the state of the conversation that consists of multiple features (such as user profile info, dialog history, sentiments);

- the set of possible Goals in the conversation $G = \{G_1, G_2, ... G_n\}$, where G is the current goal aimed at (e.g., elicit information, tell a joke, show engagement content); and

- the set of Actions $A = \{A_1, A_2, ... A_n\}$, A being the chatbot action performed, which changes the state and satisfies the current goal (e.g ask question to elicit info);

the aim is to:

- identify the composition of current State; and

- identify the *policy*, i.e., which Action to take given current state S and the Goals G

- such that
$$\pi(G, S) \to S'$$
where Policy π is a rule that defines the transition from state S to state S' and depends on the *Current State* S and current *Goals* G of the conversation.

The research question is actually of more general nature, and the resulting approach can be applied to any social chatbot. To us, reminiscence is an application domain we have experience with and we want to contribute to.

Success Metrics

Different metrics have been proposed for evaluating the quality of conversations with dialog agents, such as: i) user engagement (Cervone et al. 2017; Fitzpatrick, Darcy, and Vierhile 2017), ii) task completion (Huang, Lasecki, and Bigham 2015), iii) conversation quality: including dialog consistency and memory of past events (Lasecki et al. 2013), iv) human-like communication (Kopp et al. 2005). The approach to evaluation – and therefore the choice of metrics – is based on the aim of the agent: having an engaging chat or performing a specific task (e.g., booking a flight). In our case, the reminiscence chatbot is a combination of conversational and task-based agent, as it aims at both having an engaging conversation with the user and collecting information while doing so. Therefore, we consider metrics for both types of agents, including: i) engagement (as subjective

measure); ii) number of turns of conversation made before it drops; iii) times conversation drops overall; iv) domain-specific metrics like the amount of content which the user has provided during one conversation session (amount of pictures uploaded, amount of data attributes filled about a relevant person), and other task-completion metrics.

Related work

Crowdsourcing has been used to support all aspects of chatbot design, from holding direct conversations with final users, to supporting conversation design – the latter being the family of approaches under which we position our work. Prior work on crowdsourcing has addressed the *bootstrapping challenge*, investigating strategies to create dialog datasets to train algorithms (Takahashi and Yokono 2017; Lin, D'Haro, and Banchs 2016), infer conversation templates (Mitchell, Bohus, and Kamar 2014) or declarative conversation models (Negi et al. 2009). It has also been explored to *enrich conversation dialogs* to provide meaning and context, by annotating dialogs with semantics and labels with, for example, polarity and appropriateness (Lin, D'Haro, and Banchs 2016), extracting entities (Huang 2016), as well as providing additional utterances for more natural conversations (paraphrasing) (Jiang, Kummerfeld, and Laseck 2017). Other approaches incorporate the crowd in the *evaluation of chatbot quality*, making sure crowd contributions are valid and safe (Chkroun and Azaria 2018; Huang et al. 2016) and even allowing users to train chatbots directly (Chkroun and Azaria 2018). Acknowledging that chatbot conversations are not perfect, some approaches explore strategies to *escalate conversation decisions* to the crowd in cases where the chatbot is not able to interpret or serve the user request (Behera 2016).

The above highlight the potential of crowdsourcing for designing chatbots. We take these approaches as the starting point for exploring the specific challenges of designing and maintaining a reminiscence bot. Previous work in this domain – though valuable in insights – has been limited to human-operated chatbots and Wizard of Oz evaluations, highlighting the complexity of chatbot design in general and in particular for our target population (Tsiourti et al. 2016a; Fuketa, Morita, and Aoe 2013; Yaghoubzadeh, Pitsch, and Kopp 2015).

Ongoing and Future Work

Next, we are going to define concrete crowdsoursing strategies to elicit the nature of the states, goals and actions that will give structure to the model. Then, we will focus on tasks to fill the model with data and on algorithms to effectively aggregate and apply the elicited knowledge.

Acknowledgments

References

Behera, B. 2016. Chappie-a semi-automatic intelligent chatbot.

Cervone, A.; Tortoreto, G.; Mezza, S.; Gambi, E.; Riccardi, G.; et al. 2017. Roving mind: a balancing act between open–domain and engaging dialogue systems. In *Alexa Prize*, volume 1. https://developer. amazon. com/alexaprize/proceedings.

Chkroun, M., and Azaria, A. 2018. did i say something wrong?: Towards a safe collaborative chatbot.

Demiris, G.; Thompson, H. J.; Lazar, A.; and Lin, S.-Y. 2016. Evaluation of a digital companion for older adults with mild cognitive impairment. In *AMIA Annual Symposium Proceedings*, volume 2016, 496. American Medical Informatics Association.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health* 4(2):e19.

Fuketa, M.; Morita, K.; and Aoe, J.-i. 2013. Agent–based communication systems for elders using a reminiscence therapy. *International Journal of Intelligent Systems Technologies and Applications* 12(3-4):254–267.

Hanke, S.; Sandner, E.; Kadyrov, S.; and Stainer-Hochgatterer, A. 2016. Daily life support at home through a virtual support partner.

Hawthorn, D. 2000. Possible implications of aging for interface designers. *Interacting with computers* 12(5):507–528.

Huang, T.-H. K.; Lasecki, W. S.; Azaria, A.; and Bigham, J. P. 2016. " is there anything else i can help you with?" challenges in deploying an on-demand crowd-powered conversational agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Huang, T.-H. K.; Lasecki, W. S.; and Bigham, J. P. 2015. Guardian: A crowd-powered spoken dialog system for web apis. In *Third AAAI conference on human computation and crowdsourcing*.

Huang, T.-H. K. 2016. Crowd-powered conversational agents.

Huldtgren, A.; Mertl, F.; Vormann, A.; and Geiger, C. 2015. Probing the potential of multimedia artefacts to support communication of people with dementia. In *Human-Computer Interaction*, 71–79. Springer.

Jiang, Y.; Kummerfeld, J. K.; and Laseck, W. S. 2017. Understanding task design trade-offs in crowdsourced paraphrase collection. *arXiv preprint arXiv:1704.05753*.

Kopp, S.; Gesellensetter, L.; Krämer, N. C.; and Wachsmuth, I. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents*, 329–343. Springer.

Lasecki, W. S.; Wesley, R.; Nichols, J.; Kulkarni, A.; Allen, J. F.; and Bigham, J. P. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 151–162. ACM.

Lin, L.; D'Haro, L. F.; and Banchs, R. 2016. A web-based platform for collection of human-chatbot interactions. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, 363–366. ACM.

López, A.; Ratni, A.; Trong, T. N.; Olaso, J. M.; Montenegro, S.; Lee, M.; Haider, F.; Schlögl, S.; Chollet, G.; Jokinen, K.; et al. 2016. Lifeline dialogues with roberta. In *International Workshop on Future and Emerging Trends in Language Technology*, 73–85. Springer.

Miron, A. M.; Thompson, A. E.; McFadden, S. H.; and Ebert, A. R. 2017. Young adults concerns and coping strategies related to their interactions with their grandparents and great-grandparents with dementia. *Dementia* 1471301217700965.

Mitchell, M.; Bohus, D.; and Kamar, E. 2014. Crowdsourcing language generation templates for dialogue systems. *Proceedings of the INLG and SIGDIAL 2014 Joint Session* 172–180.

Negi, S.; Joshi, S.; Chalamalla, A. K.; and Subramaniam, L. V. 2009. Automatically extracting dialog models from conversation transcripts. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, 890–895. IEEE.

Nikitina, S.; Callaioli, S.; and Baez, M. 2018. Smart conversational agents for reminiscence.

Nurgalieva, L.; Laconich, J. J. J.; Baez, M.; Casati, F.; and Marchese, M. 2017. Designing for older adults: review of touchscreen design guidelines. *arXiv preprint arXiv:1703.06317*.

Park, H. L.; O'Connell, J. E.; and Thomson, R. G. 2003. A systematic review of cognitive decline in the general elderly population. *International journal of geriatric psychiatry* 18(12):1121–1134.

Ring, L.; Barry, B.; Totzke, K.; and Bickmore, T. 2013. Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 61–66. IEEE.

Ring, L.; Shi, L.; Totzke, K.; and Bickmore, T. 2015. Social support agents for older adults: longitudinal affective computing in the home. *Journal on Multimodal User Interfaces* 9(1):79–88.

Schlögl, S.; Doherty, G.; and Luz, S. 2014. Wizard of oz experimentation for language technology applications: Challenges and tools. *Interacting with Computers* 27(6):592–615.

Subramaniam, P., and Woods, B. 2012. The impact of individual reminiscence therapy for people with dementia: systematic review. *Expert Review of Neurotherapeutics* 12(5):545–555.

Takahashi, T., and Yokono, H. 2017. Two persons dialogue corpus made by multiple crowd-workers. In *Proceedings of the 8th International Workshop on Spoken Dialogue Systems (IWSDS)*.

Tsiourti, C.; Moussa, M. B.; Quintas, J.; Loke, B.; Jochem, I.; Lopes, J. A.; and Konstantas, D. 2016a. A virtual assistive

companion for older adults: design implications for a real-world application. In *Proceedings of SAI Intelligent Systems Conference*, 1014–1033. Springer.

Tsiourti, C.; Quintas, J.; Ben-Moussa, M.; Hanke, S.; Nijdam, N. A.; and Konstantas, D. 2016b. The cameli frameworka multimodal virtual companion for older adults. In *Proceedings of SAI Intelligent Systems Conference*, 196–217. Springer.

Vardoulakis, L. P.; Ring, L.; Barry, B.; Sidner, C. L.; and Bickmore, T. 2012. Designing relational agents as long term social companions for older adults. In *International Conference on Intelligent Virtual Agents*, 289–302. Springer.

Webster, J. D., and Gould, O. 2007. Reminiscence and vivid personal memories across adulthood. *The International Journal of Aging and Human Development* 64(2):149–170.

Yaghoubzadeh, R.; Pitsch, K.; and Kopp, S. 2015. Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users. In *International Conference on Intelligent Virtual Agents*, 28–38. Springer.