Towards Healthcare Personal Agents

Giuseppe Riccardi
Signals and Interactive System Lab
University of Trento
Italy
riccardi@disi.unitn.it

ABSTRACT

For a long time, the research on human-machine conversation and interaction has inspired futuristic visions created by film directors and science fiction writers. Nowadays, there has been great progress towards this end by the extended community of artificial intelligence scientists spanning from computer scientists to neuroscientists. In this paper we first review the tension between the latest advances in the technology of virtual agents and the limitations in the modality, complexity and *sociability* of conversational agent interaction. Then we identify a research challenge and target for the research and technology community. We need to create a vision and research path to create personal agents that are perceived as devoted assistants and counselors in helping end-users managing their own healthcare and well-being throughout their life. Such target is a high-payoff research agenda with high-impact on the society. In this position paper, following a review of the state-of-the-art in conversational agent technology, we discuss the challenges in spoken/multimodal/multi-sensorial interaction needed to support the development of Healthcare Personal Agents.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**] – I.2 [**Artificial Intelligence**]

General Terms: Algorithms, Design.

Keywords

Multimodal Interaction, Virtual Agents, Robotics, Affective Computing, Language Understanding, Healthcare Systems

1. INTRODUCTION

Apple's Siri[1] started the trend for mobile Virtual Agents[1] (VA) that provide a speech interface for question-answering in 2011. Awareness was created throughout millions of consumers that indeed smartphones could speak, be funny, answer limited set of questions and execute simple tasks (make a call by-speaking-it). Following Apple, several companies including Google[3], AT&T[5], Nuance[4], etc. have entered the market of Intelligent Virtual Agents.

The common functionality across VAs is their ability to interpret Natural Language via spoken interaction and providing responses either in the form of a software program execution (e.g. opening the contacts folder) or a spoken response (Question Answering). A set of software program actions minimally available to the VAs are: making calls, sending text messages, setting reminders, control of music players, interaction with navigation tools, etc. As for Question Answering capability, the span of supported questions is quite wide and ranges from general factual questions like "*What is the capital of Italy?*" to situational questions like "*Which is the nearest Chinese restaurant?*". The ability to handle situational questions stems from the VAs usually having access to the phone sensors such as GPS, user agendas, etc, while the ability to launch other applications and having access to third-party services such as Yelp or Wolfram Alfa allows VAs to handle factual questions.

However, VAs differ in their approach towards user interactions. Google Now, unlike Siri, does not try to actively engage the user in a spoken conversation. While the user can perform voice searches and give voice commands using Google's always on OK-Google feature, where each command must be preceded by the phrase 'OK Google', Google Now mostly works in the background to keep information ready that the user might need at any point of time. The user has the ability to customize his Google Now feed to edit his work, home locations, and to set his topics of interests. This is referred to as *anticipatory computing* as it tries to predict what a user will need and provides that information even before the user asks for it.

Microsoft introduced its own Personal Assistant , Cortana[2] , in 2014. Cortana, like Siri can use a speech interface to answer questions, find information, or perform tasks. Cortana also combines Google Now's anticipatory computing feature to make proactive recommendations to the user. Cortana, like Google Now learns from user's behavior, and tries to predict and alert the user about weather, meetings, traffic, airlines information.

Several VAs have been developed using AT&T's Speech API[1]. Unlike Google Now, Siri, and Cortana the Speaktoit assistant developed using the AT&T's Speech API has a human face and a body. However, the assistant is a static image and the animation is not interactive. Another Assistant Multimodal Virtual Assistant ([5]) using the AT&T speech API is multimodal in the sense that it can combine speech and location of the user with gesture inputs on a map on the user interface. The user can point to a location on the map, and ask questions which limits the responses to the geographical region indicated on the map.

Another class of VA has spawned out of chat bots. Many of such VAs utilize Pandorabots (e.g. Voice Actions (Jeannie), Skyvi,

[1] Note on terminology: there are various terms to refer to the software like Siri: "***Intelligent Virtual Assistant***" (most common), "Mobile Virtual Assistant", "Virtual Personal Assistant", "Intelligent Software Assistant" (Vlingo). Additionally, they are referred to as "Knowledge Navigators" as the IR function they perform.

RFMIR'14, November 16, 2014, Istanbul, Turkey.
Copyright © 2014 ACM 978-1-4503-0615-7/14/11…$15.00.
http://dx.doi.org/10.1145/2666253.2666266

Iris, CallMom). CallMom[2] is an AIML-based (Artificial Intelligence Markup Language) Virtual Assistant for Android. Besides features common across Virtual Assistants such as having a conversation, dialing a number, etc.; unlike others CallMom has a learning feature for user preferences and can learn to correct Speech Recognition errors. Additionally, the app provides several virtual personalities. The knowledge base of CallMom is based on AIML and fully customizable. Additionally, this knowledge base is open source and allows anyone to create his or her own Virtual Assistant personality. Since historically AIML is used for chat-bots that have very limited usability besides chatting, the app developers state that these bots are immature to perform Virtual Assistant tasks. AIML based VAs inherit the limitations of the AIML specification, which is criticized as being too simple (collection of if-then rules) to be scalable for complex VAs due to its weak pattern matching ability.

2. Limitations

While some VAs are speech-based and can hold limited conversations, others tend to work in the background providing information when and where needed. Most VAs utilize rule-based dialogue strategies such as confirmation and verification to ensure the understanding with the user. However, the extent and adaptability of these strategies is limited with respect to the variability of the user preferences, context and noise in the user input. The current VAs instantiations are similar to Question-Answering or command-and-control systems. Siri and Cortana are not able to maintain a complex dialogue flow. They can remember context across a few (mostly one or two) turns, and tend to treat most utterances as individual queries and commands. In case the VA is unable to understand the context, its fall back strategy is to present the user with the results of a web-search with the query.

Siri became the first mobile virtual assistant to incorporate humor triggers in its conversation. The user can not only ask Siri to tell a joke, but can ask questions such as "*What is the meaning of Life?*" or "*Why did the chicken cross the road?*" and get a humorous response. Cortana, like Siri, also incorporates an hand-crafted sense of humor. While this is certainly one interesting insight, the humorous trait of the VA does not seem to be fully consistent and evaluated to be sustainable along the course of the relationship between a user and *its* smartphone. Even more challenging, is the notion of controlling humorous behavior across situational contexts, user state, languages and cultures.

2.1 Challenges for Personal Healthcare Agents.

Advances in mobile technologies such as voice, video, touch-screens, web 2.0 capabilities and integration of various on-board sensors and wearable computers, have rendered mobile devices as ideal units for delivery of healthcare services [10]. The 2012 survey in [9] reports that in Europe there were more than one hundred health apps in a variety of languages (Turkish, Italian, Swedish, etc..) and domains (mental problems, self-diagnosis, heart-monitoring, etc.). Such growing number of smartphone applications can track user activity, sleeping and eating habits and covert and overt signals such as blood pressure, heart rate, skin temperature, speech, location, movement, etc. by either using the on-board sensors of the smartphone or interacting with various wearable and healthcare monitoring devices.

In the recent years there has been a growing research interest in creating such applications which can interact with people though context-aware multimodal interfaces and have been used for various healthcare services ranging from monitoring and accompanying the elderly [10][11], to providing healthcare interventions for long-term behaviour changes [7].

Such agents can be useful in keeping track of patient activity in-between visits or to ensure the patients are taking their medicines on time, or that they follow their advised health routine.

In the future Healthcare Personal Agent research and development should plan for an agenda where serious limitations are addressed and new avenues are explored. Such agenda can directly impact the quality of life and health of people by disrupting current models of delivering health care services. To this end, such personal agents should have the following characteristics:

a) **Spoken, Multimodal and Multisensorial Understanding**. Agents may take different physical and virtual appearance ranging from avatars to robots (e.g. [11]). Note that in the latter case, although there has been research interest in incorporating the linguistic or paralinguistic communication channel, very little progress has been made, mostly due to the distance (cultural and technical) between the robotics and speech and language research communities. There are many fundamental research questions still open. Amongst others, 1) the understanding of linguistic signs and coordination with other senses (e.g. vision) and 2) the scale of the understanding model in terms domain openness. There are many instances of VAs, although domain-limited, that have been evaluated, in realistic scenarios over many users. In contrast, most conversational robots are designed and evaluated *in-vitro* in research labs. Most of the commercial attempts to take interacting robots into the market have failed. The research and technology community has to aim at taking those robots outside in the wild to be able to do experimental research and ultimately make long-lasting advances in this field.

b) **Understanding Covert Signals**. Covert signal streams from wearable and mobile sensors may be effectively used to model user state in terms of his/her physiological responses to external stimuli or events. A unified model of covert and overt signals being generated can help personalize healthcare virtual agents (IVA) which can continuously monitor and interact with the users, eliciting information and providing help where needed. Such modeling need to take into account user-specific observations as well as age and patient user group requirements

c) **Affectiveness**. Personal Agents need to be able to handle basic and complex emotions such as empathy. Although the research community is making progress in the recognition of emotions from speech and images, the performances and space of emotions is still limited. Even more important is the ability of coordinating and regulating emotions in the personal space of people. We need to understand the affective sustainability of human-interaction. Non-human objects may be recipient of affective signals from humans. For instance people may develop a sense of emotional attachment with their mobile phones [14]. Not only should a Virtual Agent be able to understand the emotion of the user, but should also be able to respond accordingly. In the healthcare domain, the ability to handle emotions is even more critical to

[2] http://callmom.pandorabots.com/static/callmom/

manage and support, for instance, daily healthcare routine. The affective signals and communication need to be adapted for target patient groups such as children, elderly people, or suffering from diseases leading to diminished mental capacity.

d) **Context-Awareness** : A smart VA should be always aware of the user's current context. It should be aware of both physical and emotional state of the user by monitoring and interpreting the personal and world signals of the users. Personal signals can be covert like physiological signals (blood pressure, heart rate and skin temperature), or overt (like language) in nature. World signals could involve the task the user is currently involved in (walking, driving, sleeping) or the situation the user is in (at work, in a party, in a meeting) [15]. The agent might use the on-board sensors on the phone or might interact with other wearable devices to extract such information.

e) Dialoguability

By far one of the most important skill of a conversational agent is the ability to carry out a dialogue with a human. Such ability has been traditionally associated to the so-called dialogue model. The dialogue model drives the dialogue manager which takes the linguistic input (and its interpretation), covert signals from the user and world context to decide what is the best actions (e.g. speak an utterance, grab-a-cup, nod-the-head-up-and-down) to take at any moment of the conversation. The dialogue model requires the optimal selection and coordination of (micro)actions. The dialogue model should have a representation of the *reward* function associated to the goal of conversation, for the optimization of the strategies.

For example, the agent should be aware of the right time to initiate an interaction through opportunistic interruptions. The agent should be able to decide whether some interruptions should be immediate or can be postponed to a later stage. If the agent needs to remind the user to take a medicine, and the user is currently driving, probably it is better to postpone such an interaction. However if the agent realizes that the user is falling asleep while driving, an immediate interruption to wake the user would be desirable. Different models of user interaction might be needed for different users/user-groups and different application domains (e.g. robotic surgery *vs* bank fund transfer *vs* information seeking). An application tracking brushing habits of kids might achieve better results with gamification, while an obesity monitoring agent should use motivational feedback to improve user compliance.

At the current time most dialogue models are limited by an expensive process of rule-based system design. In the case of commercial applications, this process has been engineered in limited-domain cases. More recent research work is heading into the direction of adaptive strategy self-learning (e.g. [18]) or combination of knowledge-based and adaptive systems ([19]). Although these stochastic models have the appropriate requirements for the complex decision making of VA, at the moment the research challenges are the a) scalability, b) the performances and c) the design of reward function that is appropriate for a target domain, user (group) and application.

f) **Sociability**. Since the mid '80s, an important issue related to the anthropomorphism of robots - so-called *uncanny valley effect* - has been posed, hypotheses made, but no ecological, large scale study or evaluation has been carried out. In 2000 [16] the first natural language conversational agent was trained, designed in the research lab and deployed over a very large set of customers. More recently, researchers are facing the challenge of detecting twitter robots that behave "like" any other twitter users [17]. We have not yet understood which are the *sweet-spots*, in the human-machine social relation. We need to strive for large-scale (over a crowd for a relatively short time) or vertical-continuous (over few people or a group for a long time) design, experimentation evaluation of Virtual Agents. The agents may be in the form of chatbots, multimodal avatar or robot with a physical presence in a shared space (e.g. nurse or android at home). The social component is key skill in for humans and we believe that it will be a key research challenge for the human-machine interaction and possibly for machine-machine interaction.

3. CONCLUSION

In this position paper we have highlighted the main limitations and challenges in training, designing and evaluating Personal Virtual and Conversational Agents. We have also identified a class of agents for such research challenges: the Healthcare Personal Agents. Such objective is of high-impact on health and well-being of society, and would have the complexity requirements that are needed to foster non-incremental advances in the science and engineering of agent design.

4. ACKNOWLEDGMENTS

I wish to thank Arindam Ghosh and Evgeny Stepanov who have helped me in the preparation of the paper, and Michael Johnston and Giuseppe di Fabbrizio for useful discussions over the present and future of conversational agents.

5. REFERENCES

[1] http://en.wikipedia.org/wiki/Siri

[2] http://en.wikipedia.org/wiki/Microsoft_Cortana

[3] http://en.wikipedia.org/wiki/Google_Now

[4] www.nuance.com

[5] Johnston, M. et al. 2014 *MVA: The Multimodal Virtual Assistant* In Proc. SIGDIAL Conference, pages 257–259.

[6] Corchado, Juan M., Javier Bajo, and Ajith Abraham 2008 *GerAmi: Improving healthcare delivery in geriatric residences Intelligent Systems, IEEE* 23.2: 19-25.

[7] Ring, Lazlo, et al. 2013 *Addressing Loneliness and Isolation in Older Adults: Proactive Affective Agents Provide Better Support* In Proc. Affective Computing and Intelligent Interaction Humaine Association Conference.

[8] Bickmore, Timothy W., et al. 2010 *Usability of conversational agents by patients with inadequate health literacy: Evidence from two clinical trials Journal of health communication 15.S2*: 197-210.

[9] *European Directory of Health Apps 2012-2013*, Clive Nead Editor, published by Patent View 2012.

[10] Baig, M.M. and Gholamhosseini, H. 2013. *Smart Health Monitoring Systems: An Overview of Design and Modeling,* Journal of Medical Systems. 37:9898.

[11] Minato,M., Nishio, S., Ogawa, K. and Ishiguro, H. 2012. *Development of Cellphone-type Tele-operated Android* In Proceedings of the 10th Asia Pacific Conf. Computer Human Interaction.

[12] http://en.wikipedia.org/wiki/AIBO

[13] Di Fabbrizio, G., Okken, T. and Wilpon, J. G. 2009 *A Speech Mashup Framework for Multimodal Mobile Services* In Proc. ICMI-MLMI , Cambridge, USA.

[14] Vincent, J. *Emotional attachment and mobile phones 2006* Knowledge, Technology & Policy 19.1: 39-44.

[15] Ghosh, A. and Riccardi, G. 2014 *Recognizing Human Activities from Smartphone Sensor Signals* In Proc. 22nd ACM International Conference on Multimedia.

[16] Gorin, A., Riccardi,G. and Wright, J. H., 1997, *How May I Help You ?* Speech Communication, vol. 23, pp.113-127.

[17] Freitas, A. C., Benevenuto, F., Ghosh, S. and Veloso, A., 2014, *Reverse Engineering Socialbot Infiltration Strategies in Twitter* 2014, In Proc. ELREC, Reykiavik.

[18] Williams, J. D. and Young,S. 2007, *Partially Observable Markov Decision Processes for Spoken Dialog Systems*, Computer Speech and Language, 21 (2), 393-422.

[19] Varges, S., Riccardi, G., Quarteroni, S. and Ivanov, I. 2011, *POMDP Concept Policies and Task Structures for Hybrid Dialog Management* 2011, In Proc. ICASSP, Prague.