A Spoken Dialogue System for the EMPATHIC Virtual Coach

M. Inés Torres, Javier Mikel Olaso, Neil Glackin, Raquel Justo and Gérard Chollet

Abstract The EMPATHIC project is devoted to the development of future generations of personalised virtual coaches to help elderly people to live independently. In this paper we describe a proposal to deal with the Dialogue Management of the EMPATHIC Virtual Coach. The paper describes a DM system capable of dealing with both long-term goals and well-being plans, that can implement an effective motivational model. The system to be put into practice aims for high level healthy-ageing, utilising expressive multi-modal dialogue tailored for each specific user, working in tandem with short-term goal-oriented dialogue.

1 Introduction

The EMPATHIC Research & Innovation project is devoted to the development of future generations of Personalised Virtual Coaches to help elderly people to live independently. The EMPATHIC Virtual Coach (VC) will engage the healthy-senior user to take care of potential chronic diseases, maintain a healthy diet, have adequate physical activity as well as encourage social engagement, thus contributing to the older adults' ability to maintain a satisfying and independent lifestyle. Our ambition is to create a personal, friendly and familiar environment for the users, avoiding the threatening effects of unfamiliar new gadgets or an excessive focus on medical supervision. The VC will be capable of perceiving the emotional and social state of a person, in the learned context of the senior users' expectations and requirements, and their personal history, and will respond adaptively to their needs. The VC will put into practice high level healthy-ageing and well-being plans, and implement an effective motivational model, through expressive multi-modal dialogue tailored for

M. Inés Torres and Javier Mikel Olaso and Raquel Justo
Universidad del País Vasco UPV/EHU, Spain; e-mail: manes.torres@ehu.eus

Neil Glackin and Gérard Chollet
Intelligent Voice; e-mail: gerard.chollet@intelligentvoice.com

each specific user. Thus, the research to be carried out is aimed at implementing health-coaching goals and actions through an intelligent computational system, intelligent coach and spoken dialogue system adapted to users intentions, emotions and context.

2 Previous Work

Spoken dialogue systems have been developed in various domains and dealing with different goals but typically providing some information to the user such as flight schedules in pioneering proposal from ATT [8], restaurant recommendation [20] tourism information [7] or customer services [13]. However, spoken dialogue systems supporting tele-medical and personal assistant applications for older people [11], are more aligned with this research. The basic tasks of dialogue management are: to interpret the user utterance in a given context (Natural Language Understanding, NLU component), to decide what action the system should take (Dialogue Manager proper), and produce an appropriate response (Natural Language Generation, NLG component). The input speech is transformed into text using Automatic Speech Recognition (ASR), and the output produced by the Text to Speech Synthesis. The NLU component identifies the intent of the speaker and identifies the semantic content. This classification can be performed using a data-driven statistical approach or knowledge-based approach, such as hand-crafted grammars. The major challenge is the semantic inference due to the ambiguity of natural language and because semantic distinctions tend to vary from application to application.

The dialogue manager (DM) decides upon the actions a system should perform at any given state in the dialogue, given the string of semantic units provided by the NLU and the history of the dialogue. Due to the complexity of this task, traditionally the DM strategy has been designed using hand-crafted approaches based on trees and Finite State Machines which can be easily applied in task-oriented applications. However, for complex dialogues they can be very difficult to manage and impractical to design. A frame or an agenda-based DM [1, 11, 10] provides more flexible dialogue management by decoupling the information state and the actions that can be performed. The most advanced management structure is that of using distributed software agents, with which dialogues can be flexibly designed, making it possible to take the dynamic nature of dialogue into account. Dialogue flow can be modeled as a collaboration between the participating agents, their beliefs, and desires. In this way, intentions can be tracked, and multimodal and contextual information can be taken into account by the system when reasoning about its own state and the next action. A good example is the Ravenclaw DM developed by CMU [1], which was used to implement the LetsGo task [4], and has been used to obtain a large number of dialogues with real users [5].

Statistical dialogue managers were initially based on Markov Decision Processes [8, 21, 17], and Partially Observable Markov Decision Processes [20] where unknown user goals are modelled by an unknown probabilistic distribution over the user states. This approach is up-till-now considered as the state-of-the-art in Spoken Dialogue Systems [3, 22] even if their global optimisation features important computational difficulties that have to be addressed when dealing with real users. Bayesian Networks [9] and Stochastic Finite State models [6, 16] are alternative statistical approaches. Recently other machine learning methods have been proposed to obtain optimal dialogue strategies from data such as deep neural networks and reinforcement learning [15]. End-to-end learning for task-oriented dialogue systems was also proposed in [16] under an interactive pattern recognition approach that allowed dialogue generation based on stochastic bi-automata generative models. Recently hierarchical recurrent Neural networks have also been proposed to generate open domain dialogues [12] and build end-to-end Dialogue Systems. However, the real challenge of end-to-end framework is the decisional problem related to the DM for goal-oriented dialogues [2, 23]. A combination of Supervised learning allowing an LSTM to learn the domain and reinforcement learning that allows learning through interaction, has also been proposed as an alternative to deal with task-oriented dialogue systems [18, 19].

3 Progress beyond state of the art

EMPATHIC will go beyond the state of the art of SDS and DM in four main directions: It will include the perceived user affective status to support contextually relevant DM decisions; It will develop DM strategies that consider not only the user but also the high-level coaching goals to be achieved by a joint optimisation approach for dealing with shared user-coaching goals. It will maximise user and task understanding, will develop active turn-taking to recover the uncertainty, and will keep user engagement when required; and it will develop a DM capable of personalisation and adaptation to the user-specific profile and current status through novel online learning algorithms. Initially, EMPATHIC will define several models of interaction using a Wizard-of-Oz (WoZ) simulation approach. The WoZ has a human in the loop acting as the DM for the purpose of recording dialogues with the users. Then, a statistical DM, able to integrate not only the classical semantic inputs, but also parameters representing emotional status of the user, topics of the conversation and the user's intention, will be developed. This DM will be conducted by policies aimed at improving project-defined well-being metrics, resulting in system-driven interactions. Strategies defined by psychologists will be considered and studied through interaction, for user personalisation purposes. Alternatively, the DM will apply policies aimed at reaching user goals and at dealing with conversation topics previously detected automatically through machine learning techniques. Involved models and policies will get updated during interaction through online learning algorithms. EMPATHIC will use stochastic finite state bi-automata [16] as well as novel end-to-end learning approaches as main methodological frameworks, advancing the state of the art in dialogue management through a combination of supervised and reinforcement learning.

4 Dialogue Management and Intelligent Coach

We will now deal with user estimation and the main decisional systems needed to develop the virtual coach. To this end, we first identify cues and goals for well-being as well as personalized and detailed coaching plans, and actions to be implemented. The plans, developed by health professionals, will be implemented by an Intelligent Coach (IC) in tandem with the DM and modules devoted to user state identification and understanding. Fig-1 shows a diagram of the proposed system.

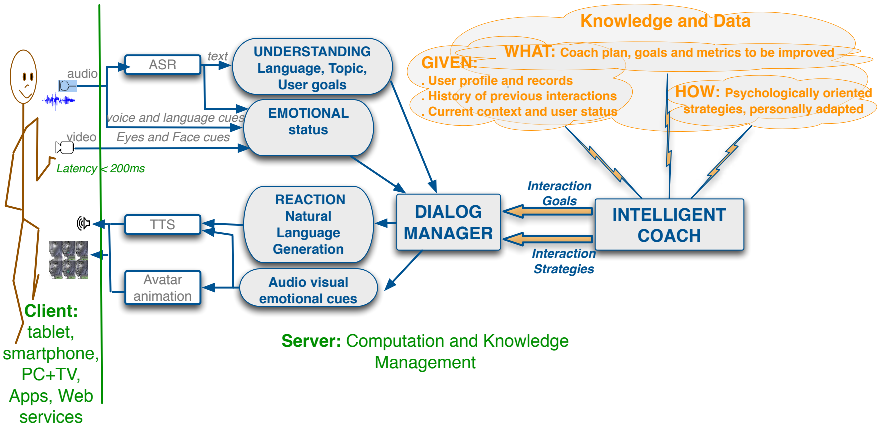
Knowledge and Data
WHAT: Coach plan, goals and metrics to be improved
GIVEN:
. User profile and records
. History of previous interactions
. Current context and user status
HOW: Psychologically oriented strategies, personally adapted
audio
ASR
text
UNDERSTANDING Language, Topic, User goals
voice and language cues
EMOTIONAL status
video
Eyes and Face cues
Latency < 200ms
REACTION Natural Language Generation
TTS
Avatar animation
Audio visual emotional cues
DIALOG MANAGER
Interaction Goals
Interaction Strategies
INTELLIGENT COACH
Client: tablet, smartphone, PC+TV, Apps, Web services
Server: Computation and Knowledge Management

Fig. 1 General view of the Virtual Coach

The Understanding module covers both local and global understanding. By local we refer to the transformation of sequences of words into a sequence of semantic units capturing concepts. Whereas global refers to the aims of the user i.e. the topic and/or the goals the user has when starting a dialogue. The first challenge is to build a component that identifies them using both current input (semantic units) and possibly also some history of the user. The DM layer includes short-term system decision. In a previous step several models of interactions had to be implemented using the WoZ while recording dialogues with the users. Two different kind of dialogues are expected: system-driven and user-driven dialogues. System-driven dialogues will be conducted by policies aiming at improving previously defined well-being metrics learned by the IC module in Figure 1. Strategies defined by psychologists will be considered and learned through interactions for user personalisation purposes. Thus, the DM will implement the dialogue goal and dialogue strategy proposed by the IC. Alternatively, the DM will deal with user-driven dialogues that we foresee to be open-domain, to some extent. In such a case the DM will also be assisted by a User goal tracker and a Topic detection module implemented trough the Understanding module in Figure 1. In both cases the DM will be able to integrate not only the classical semantic inputs, but also additional parameters representing the emotional state of the user. These parameters will be obtained from an emotional module that might consider features extracted from language, speech and/or images. Statistical approaches previously developed by the authors, such as a DM based on Bi-automata where policies are implemented through optimised search of an interaction graph tree will be used. Such approach is based on Stochastic-Finite-State-Transducers (SFST) where three different alphabets are defined: one related to semantic input, one for task related attributes and an additional one for dialog action outputs. The model is easily learned from examples using classical inference algorithms that can be run online allowing task adaptation. Additionally, they allow the development of user models that have been successfully tested [14]. We also foresee end-to-end learning approaches that combine sequence-to-sequence deep learning with reinforcement learning. We want to develop algorithms capable of selecting the specific way to approach each user, e.g. positive tone vs. natural tone, and so on. Since individual users react differently to different types of interaction. Thus, the input to the SDS (given by the ASR) in the form of semantic concepts will be of an n-best type lattice structure for training the correct interface between the SDS and ASR semantic representation. The deepest layer described in Figure 1 is the IC, which deals with the reasoning to take long-term decisions on the basis of evidence-based personalised coaching action plans. From time to time the DM will commence a specific dialogue with the user in order to achieve long-term coaching goals. This is a novel task, as most dialogue systems are user-triggered, while this will be a system-triggered.

5 Concluding remarks

The EMPATHIC Research & Innovation project is devoted to the development of future generations of Personalised Virtual Coaches to help elderly people to live independently. We have described a proposal to deal with the Dialogue Management of a Virtual Coach capable of perceiving the emotional and social state of a person, and which can consider their personal history and respond adaptively to their needs. The paper has described how the EMPATHIC Virtual Coach we will put into practice high level healthy-ageing and well-being plans, and implement an effective motivational model, through expressive multi-modal dialogue tailored for each specific user. The project will be developed under the Social Challenge Pillar and Health Work programme, demographic change and well-being of the EU Horizon 2020 Program (www.empathic-project.eu).

Acknowledgements This work is founded by the European Commission H2020 SC1-PM15 program under RIA grant 769872.

References

1. Bohus, D., Rudnicky, A.I.: The RavenClaw dialog management framework: Architecture and systems. Computer, Speech and Language **23**(3), 332–361 (2009)
2. Bordes, A., Boureau, Y.L., Weston, J.: Learning end-to-end Goal Oriented Dialog. In: International Conference of Learning representations (2017)
3. ci cek, F.J., Thomson, B., Young, S.: Reinforcement learning for parameter estimation in statistical spoken dialogue systems. Computer, Speech and Language **26**(3), 168–192 (2012)
4. Eskénazi, M., Black, A.W., Raux, A., Langner, B.: Let's go lab: a platform for evaluation of spoken dialog systems with real world users. In: INTERSPEECH, p. 219. ISCA (2008)
5. Ghigi, F., Eskenazi, M., Torres, M.I., Lee, S.: Incremental Dialog Processing in a Task-Oriented Dialog. In: InterSpeech, pp. 308–312 (2014)
6. Hurtado, L.F., Planells, J., Segarra, E., Sanchis, E.: Spoken dialog systems based on online generated stochastic finite-state transducers. Speech Communication **83**, 81 – 93 (2016). DOI https://doi.org/10.1016/j.specom.2016.07.011. URL http://www.sciencedirect.com/science/article/pii/S0167639316301984
7. Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J.D., Henderson, M.: The fourth dialog state tracking challenge. In: Dialogues with Social Robots - Enablements, Analyses, and Evaluation, Seventh International Workshop on Spoken Dialogue Systems, IWSDS 2016, Saariselkä, Finland, January 13-16, 2016, pp. 435–449 (2017). DOI 10.1007/978-981-10-2585-3_36. URL https://doi.org/10.1007/978-981-10-2585-3_36
8. Levin, E., Pieraccini, R., Eckert, W.: A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions on Speech and Audio Processing **8**(1), 11–23 (2000)
9. Martínez, F.F., López, J.F., de Córdoba Herralde, R., Martínez, J.M.M., Hernández, R.S.S., Muñoz, J.M.P.: A bayesian networks approach for dialog modeling: The fusion bn. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2009. IEEE, New Jersey, EEUU (2009). URL http://oa.upm.es/5579/
10. Olaso, J., Torres, M.I.: User experience evaluation of a conversational bus information system in spanish. In: 8th IEEE International Conference on Cognitive Infocommunications, Debrecen, Hungary, September 2017 (2017)
11. Olaso, J.M., Milhorat, P., Himmelsbach, J., Boudy, J., Chollet, G., Schlögl, S., Torres, M.I.: A Multi-lingual Evaluation of the vAssist Spoken Dialog System. Comparing Disco and RavenClaw, pp. 221–232. Springer Singapore, Singapore (2017)
12. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pp. 3776–3783. AAAI Press (2016). URL http://dl.acm.org/citation.cfm?id=3016387.3016435
13. Serras, M., Perez, N., Torres, M.I., Del Pozo, A., Justo, R.: Topic Classifier for Customer Service Dialog Systems, pp. 140–148. Springer International Publishing, Cham (2015). URL https://doi.org/10.1007/978-3-319-24033-6_16
14. Serras, M., Torres, M.I., Del Pozo, A.: Online Learning of Attributed Bi-Automata forDialogue Management in Spoken Dialogue Systems, pp. 22–31. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-58838-4_3. URL https://doi.org/10.1007/978-3-319-58838-4_3
15. Su, P.H., Vandyke, D., Gasic, M., Kim, D., Mrksic, N., Wen, T.H., Young, S.: Learning from Real Users: Rating Dialogue Success with Neural Networks for Reinforcement Learning in Spoken Dialogue Systems. In: InterSpeech, pp. 2007–2011 (2015)
16. Torres, M.I.: Stochastic Bi-Languages to model Dialogs. In: International Conference on Finite State Methods and Natural Language Processing, pp. 9–17 (2013)
17. Walker, M.: An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. Journal of Artificial Intelligence Research **12**, 387–416 (2000)
18. Williams, J.D.: End-to-end deep learning of task-oriented dialog systems. In: Keynote in Future and Emerging Trends in Language Technologies FETLT, Seville (2016)

19. Williams, J.D., Asadi, K., Zweig, G.: Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: ACL (1), pp. 665–677. Association for Computational Linguistics (2017)
20. Williams, J.D., Young, S.: Partially observable Markov decision processes for spoken dialog systems. Computer, Speech and Language **21**(2), 393–422 (2007)
21. Young, S.: Probabilistic Methods in Spoken Dialogue Systems. Philosophical Transactions of the Royal Society of London (2000)
22. Young, S., Gašić, M., Thomson, B., Williams, J.D.: POMDP-based Statistical Spoken Dialog Systems: A review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)
23. Zhao, T., Eskénazi, M.: Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA, pp. 1–10 (2016)