SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue
15-17 August 2017, Saarbrücken, Germany

Towards a Natural User Interface for Small Groups in Real Museum Environments

Marco Valentino
Università degli
Studi di Napoli Federico II
m.valentino91@gmail.com

Maria Di Maro
Università degli
Studi di Napoli Federico II
mdimaro17@gmail.com

Francesco Cutugno
Università degli
Studi di Napoli Federico II
cutugno@unina.it

Abstract

In this paper we present a preliminary design and evaluation of a natural user interface for multimodal conversational agents which can be placed in real museums. The natural user interface aims at creating an experience in that users can behave naturally. Specifically, focusing on the requirements imposed by a real museum context, our goal is to implement an interface for small groups in order to allow users to interact through their own bodies without additional and auxiliary devices.

1 System Architecture

The possibility to use embodied conversational agents in real contexts, like museums, has been already investigated in other works (Swartout and et al., 2010; Kopp et al., 2005), with several issues reported. These works, in dealing with real environmental challenges, contrive strategies which restrict the way users can freely interact. Therefore, we are interested in investigating and testing alternative approaches for modelling small groups interactions in real contexts, which allow users to communicate with both verbal and non-verbal actions. With the exclusive use of natural human means of communication, i.e. voice, language and gestures, the virtual agent, projected on a curved screen, understands multimodal dialogue acts performed by users asking for information about paintings or other artworks contained in a 3D scene. Since users can interact in shared environments, the multimodal system is based on a probabilistic model to correctly focus its attention on a single user in the group. Specifically, as primary work, we have implemented three input modules with the purpose of modelling an interaction based on speech and pointing gestures:

- *Natural Language Understanding* (NLU), responsible to process speech signals in order to obtain a semantic interpretation.
- *Pointing Recognition* (PR), in charge of recognising which objects are pointed by users.
- *Active Speaker Detection* (ASD), which allows to identify the last speaker over the time.

The NLU module has been designed through a semi-automatic SRGS grammar extended with a graph database (Origlia et al., 2017). Moreover, Pointing Recognition and Active Speaker Detection have been implemented by vector calculations thanks to a combined use of Unreal Engine 4[1] and Kinect 2. This integration allows avoiding a data-driven approach which usually requires a huge amount of training data. Furthermore, the game engine provides facilitation to create an immersive 3D environment and to directly project users into the scene. The entire setup of the interaction environment consists of a curved screen 2,5m high and 4,4m long. One Kinect is placed on the floor, at the centre of the screen, for tracking users movements and their speech signals in real time. All the users are tracked in a parallel and independent way but the attention will only be focused on the one who has produced the current dialogue act. His verbal and non-verbal signals are therefore combined into a multimodal fusion engine to understand the current request. Linguistic spatial expressions, such as *left* or *center*, are also taken into account to allow users to freely choose the referring strategy. These expressions are used to further reinforce the meaning of the pointing gestures, when they co-occur, in order to make clear what external entity the active speaker is referring to. The multimodal fusion engine adopts a

[1] www.unrealengine.com

hybrid approach based on probabilistic rules (Lison and Kennington, 2015). The designed network is shown in Figure 1. The input fusion process is activated as soon as a user dialogue act is recognised. Therefore, network synchronisation is performed through the *a_u* variable. By adopting this model, the system is able to combine input data into a single random variable representing the current request. Finally, by considering the one with the highest probability, the system selects the action to perform through an utility-based approach.

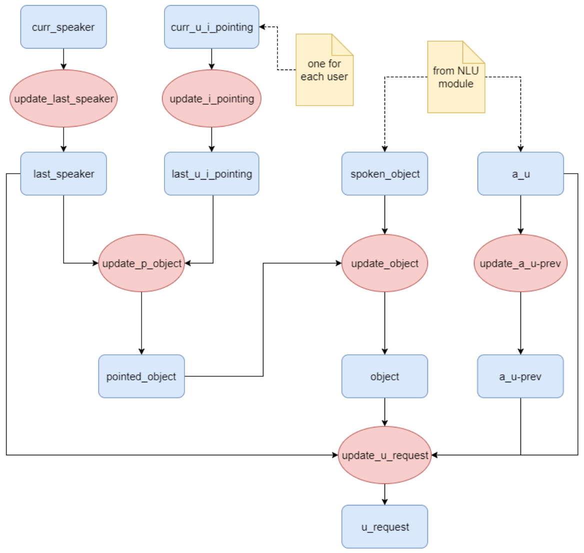

Figure 1: The probabilistic network designed to understand user requests in group interactions. The random variables are represented as blue nodes while the probability rules as red ones

2 Preliminary Evaluation

Preliminary tests were conducted with the aim of getting both limits and potentialities of the implemented architecture. Following what discussed in (Khnel et al., 2010), we analysed system performances by computing the success rate of each input module during simultaneous interactions with groups of two users. System usability was also evaluated by asking participants to compile a 7 point scale USE questionnaire. The obtained results are shown in Table 1 and Table 2.

ASD	PR	NLU
88%	**97%**	**71,4%**

Table 1: Recognition success rates for Active Speaker Detection (ASD), Pointing Recognition (PR) and Natural Language Understanding (NLU)

Usefulness	Ease of Use
6.16	**6.22**
Ease of Learning	Satisfaction
6.77	**6,44**

Table 2: USE questionnaire results

3 Future Works

Promising results prove both the potentiality of this framework and the positive attitude showed by participants. As the NLU error rate is mainly caused by environmental noises, further improvement can be reached by placing more than one Kinect in the interactive environment. Moreover, starting from these results, our purpose is to extend the system functionality by adding new input modalities, such as new gestures, gaze and facial expressions, prosody analysis and modelling of a multi-party dialogue to improve and promote collaborative interactions between users.

Acknowledgments

This paper has been developed within the framework of CHROME (*Cultural Heritage Resources Orienting Multimodal Experience* - PRIN 2015 MIUR), an ongoing Italian project on technologies for cultural heritage.

References

S. Kopp, L. Gesellensetter, N.C. Kraemer, and I. Wachsmuth. 2005. A conversational agent as museum guide design and evaluation of a real-world application. *Intelligent Virtual Agents. IVA 2005. Lecture Notes in Computer Science, vol 3661*, pages 329–343.

C. Khnel, T. Westermann, B. Weiss, and S. Mller. 2010. Evaluating multimodal systems: A comparison of established questionnaires and interaction parameters. In *Proceedings of NordiCHI*, pages 286–294.

P. Lison and C. Kennington. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech and Language, Volume 34, Issue 1*, pages 232–255.

A. Origlia, G. Paci, and F. Cutugno. 2017. Mwn-e: a graph database to merge morpho-syntactic and phonological data for italian. In *Proc. of Subsidia*, page to appear.

W. Swartout and et al. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. *Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science, vol 6356*, pages 286–300.