ISSN 1870-4069

CARLA: Conversational Agent in Virtual Reality with Analytics

Pablo Isaac Macias-Huerta[1], Guillermo Santamaría-Bonfil[2], Maria Blanca Ibañez[3]

[1] CECyT 9, Mexico
[2] CONACYT-INEEL, Mexico
[3] Universidad Carlos III de Madrid, Spain

guillermo.santamaria@ineel.mx, pmaciash1800@alumno.ipn.mx, mbibanez@it.uc3m.es

Abstract. Intrusive interfaces used in online learning environments impose extra cognitive load on learners. Fortunately, current technology trends offer new interactive ways with applications that might ease interactions with learning environments. In this paper, we present a framework to develop a spoken Conversational Agent based in viRtual reaLity with Analytics, namely CARLA. This chatbot is meant to provide interaction and support to any user within an e-Learning platform through a spoken dialogue. For the time being, we only present CARLA components and experimentation related to the Natural Language Understanding, the Dialogue Manager, and the Natural Language Generator using open source and low-cost commercial software. The current development state is presented using the case of Power Systems Education based on virtual reality technology.

Keywords: Conversational agent, virtual reality, natural language processing.

1 Introduction

In accordance to Google Trends, due to the COVID-19 pandemic, e-learning technologies have reached their major peak of interest in the last 14 years[4]. From traditional to corporate training, these technologies are a good mean to control the outspread of the virus by keeping social distance in the classroom. E-Learning ranges from isolated platforms such as Virtual Reality (VR) environments, to a high level of intertwined Learning Ecosystems (LEs) [7].

Virtual Reality (VR) allows the deployment of contextualized learning environments with many benefits: it facilitates learning through exploration and

[4] http://bit.ly/eLearningTrend

repetitive practice, integrates many degrees of freedom, it improves students' motivation and engagement to learn, it ranges from low (and conventional) to highly sophisticated (and costly) equipment, and so on [8]. Unfortunately, most of the existing VR systems use conventional computer inputs and hardware such as keyboard and mouse clicks, whereas highly sophisticated interfaces add an extra cognitive load to the process learning. Fortunately, with the growing maturity of conversational technologies, the possibilities for integrating conversation and discourse in e-learning are receiving greater attention in both research and commercial settings. Conversational agents have been produced to meet a wide range of applications, including tutoring, question-answering, conversation practice for language learners, pedagogical agents and learning companions, and dialogues to promote reflection and metacognitive skills [3].

In this article, we present CARLA, a spoken Conversational Agent in viRtual reaLity with Analytics. This chatbot aims to serve as a tutoring system in a given specific domain situated in a virtual reality environment. It provides knowledge and support to any given user through an interchange of spoken utterances for manipulating objects and navigating the virtual world. As a case of use, the power system education was the topic of choice. Thus, this chatbot is meant to provide knowledge and support for electrical equipment, electrical concepts, energy sources, and other basic principles while allowing the manipulation of the 3D environment.

The rest of this paper is organized as follows: in section 2 materials and tools required by the proposed framework are presented; section 3 details CARLA architecture and explains its functionality; section 4 presents CARLA platform as a whole and an interaction example, also, the DM functionality is analyzed by providing 15 utterances for valid chatbot intentions; finally, section 5 briefly presents conclusions of the project, and discusses future improvements and changes to the platform.

2 Technologies

There are many ways to build a conversational agent, from social media add-ons to real life embodied service robots such as Pepper, an industrially produced humanoid robot created for business and consumer needs [6]. We propose CARLA as a framework for developing a VR spoken chatbot using low-cost or free technologies. We now discuss the learning environment software, the NLP components are emphasized.

2.1 VR Learning Environment

The learning environment is based on non-immersive VR developed using Unity 3D[5]. Unity 3D is a game engine used for developing serious and non-serious games.

[5] https://unity.com/

It is free and a highly compatible VR engine with multiple operating systems. Similarly, it is beginner-friendly reducing the learning curve for future team members.

For the study case which is the power systems education area, 3D scenarios were designed. These are different power systems facilities such as an electrical substation or a microgrid. Each scenario consists of two scenes, a navigable full-fledged and fully sized power system facility and a catalogue where the facilities components can be studied in detail.

2.2 Chatbot Components

A chatbot is a computer system that works as an interface between human users and software via text or spoken utterances. It is comprised of 3 main elements: the Natural Language Understanding (NLU), a Dialog Manager, and Natural Language Generation (NLG) [2].

The purpose of the NLU is to extract the *intention* of an utterance in the form of a semantic representation. Loosely speaking, an *intent* refers to the data, information, or manipulation over the learning environment the user wants to accomplish. Thus, the NLU parses utterances, first by mapping from speech to text and then producing a linguistic structure which can be handled by a dialogue manager.

In the case of CARLA, the NLU is comprised by the Google Speech API[6], and Stanford CoreNLP toolkit [5]. The first, is a web-based Speech-to-Text tool offered by Google. This service receives an audio file and returns a text file with the recognized words. Google speech cost is 0.006 USD for 15 seconds of recording. This platform was chosen due to it having plenty of documentation and an active community. The second, namely CoreNLP for short, is a NLP toolkit for the Java programming language developed by Stanford CoreNLP Group [5]. This library enables users to construct NLP pipelines through building blocks for higher level text understanding applications: tokenization and sentence splitting, the identification of Parts Of Speech (POS), named entities, and numeric and time values, coreferences, sentiment analysis, and so on.

It supports 6 languages, and depending on the language of choice, some may have less functionalities than others, the English being the most complete one. Nevertheless, due to the VR environment we were forced to use Sergey Tihon's Stanford CoreNLP for .NET[7] version of the toolkit, which also requires IKVM[8] as a way to connect both platforms together. IKVM -wordplay on "JVM", which stands for Java Virtual Machine, in which the creator "just took the two letters adjacent to the J"- is an implementation for Java for the Mono and Microsoft's .NET frameworks.

The Dialog Manager is a module that checks the utterance to a database to manage what response or responses the system should produce in return [2].

[6] https://cloud.google.com/speech-to-text
[7] https://sergey-tihon.github.io/Stanford.NLP.NET/StanfordCoreNLP.html
[8] http://www.ikvm.net

Due to the limited amount of utterances CARLA needs to recognize, all of them (for now) are hard-coded into its system.

The NLG is used to synthesize speech to answer back to the user. Depending on the system type, they can be collected via previous user interactions, have a database, or hard-coded. CARLA utilizes Microsoft Azure's Text-to-Speech SDK[9], a set of web-based services, to synthesize speech in an audio file from text.

3 CARLA Architecture

CARLA is a chatbot embedded into a virtual reality environment. It has three main components: (1) Natural Language Understanding (NLU); (2) Dialog Manager (DM); (3) Natural Language Generator (NLG) (see Fig.1). In the following, we describe the dialogue flow of user utterances.

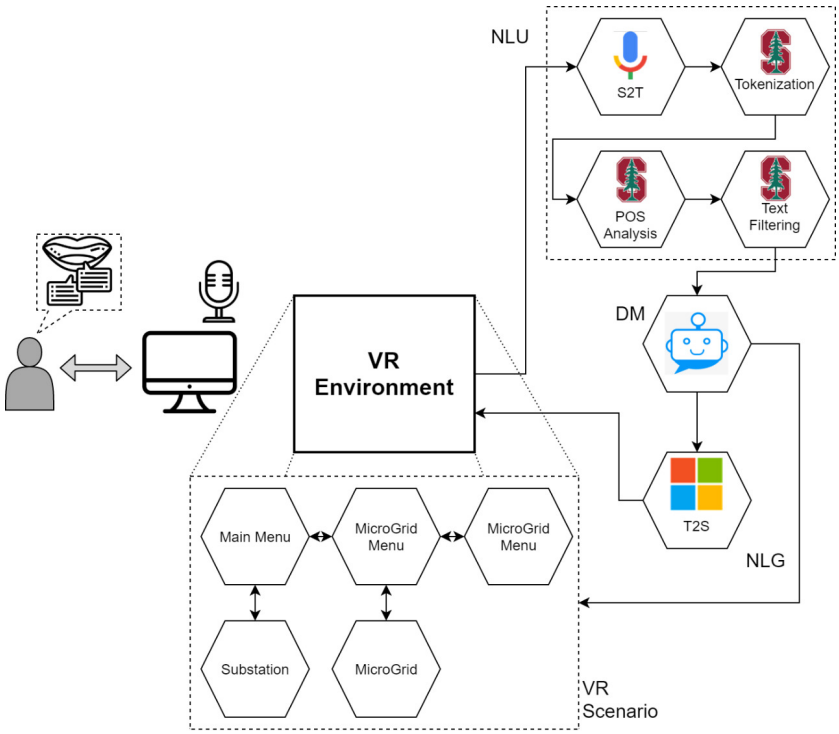

Fig. 1. CARLA framework.

The dialogue step is the NLU step. In CARLA's case, the NLU is a combination of Google Speech-to-Text API and CoreNLP tools. Thus, we turn the user's

[9] https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

voice into a text sentence using Google Speech, implemented in a *push-to-talk* fashion. Once the text of the utterance is obtained, it is necessary to understand the meaning of the sentence.

To this end, three functionalities are provided:

- Tokenization: it is the process of dividing the raw text into individual words.
- Parts Of Speech (POS): it is a categorization of words or lexical items that have similar grammatical properties. Some examples are nouns, verbs, adjectives, and so on. POS *tags* are assigned to tokens to identify the components of the sentence.
- Text filtering: instead of having CARLA analyze the raw text generated by Google Speech, we applied text filtering to analyze only relevant words. Hence, we used the information from the POS analysis and filtered out every single category that we did not need such as stop words like adjectives and punctuation marks.

Once the meaning of the utterance has been extracted by the NLU, an Intent Analysis is performed by comparing the filtered words against the different valid intents. If the intent is valid it modifies one or more objects in the VR environment, as well as provides spoken information. If the intention is not recognized, CARLA will let the user know of this situation. For the time being, this validation is hard-wired but it is expected to be automatized using machine learning algorithms [9].

For the study case, 15 valid intentions were designed namely: *show/what is a solar panel, explain how a solar panel works, show/what is a transformer, explain how a transformer works, show/what is a battery pack, explain how a battery pack works, show/what is a wind turbine, thanks, help, rotate (left and right), stop rotation, zoom in, zoom out, normal zoom, and exit.* The last stage in the human-CARLA interaction is CARLA's spoken answer. For this, we obtain the hard-coded (but can be augmented using generative text content) synthesized into audio feedback answer via the Text-to-Speech SDK in Azure. This tool has more than 140 different voices in more than 45 languages and variants, thus, it provides very expressive voices that sound like human. In particular, this API provides a neural Text to Speech support to display and personalize various styles of speech such as chat, or customer service, and even emotions such as joy and empathy.

A final word regarding the Analytics part of CARLA is deserved. Although for the time being, all analytics is focused at NLP, integration with with a formal knowledge domain model (such as a concept map, classification of dialogue utterances, usage trace data mining, and so on) is expected in the near future.

4 Experimentation

In the following we describe the current development of the VR learning environment shown in Fig. 2, then experimental results regarding CARLA intention recognition performance are presented.

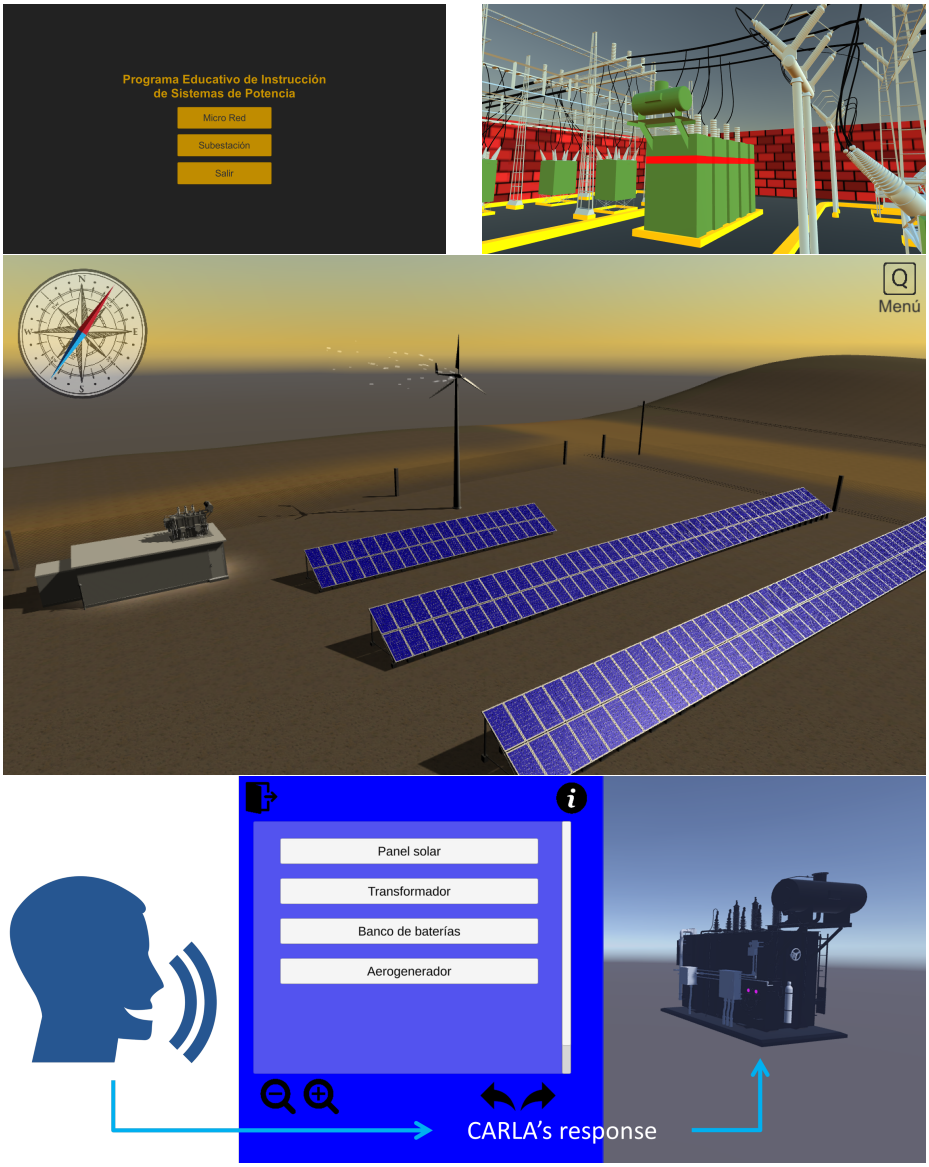

Fig. 2. CARLA VR Learning Environment. On (a) the Main Menu, where the user is greeted by CARLA; on (b) an electrical substation facility; on (c) a real world microgrid with simulation; on (d) the equipment understanding scene.

4.1 CARLA for Power Systems Education

The VR learning environment consists of three main scenes: (1) welcoming scenes, (2) facility navigation scenes, and (3) equipment understanding scenes.

The welcoming scenes are designed to provide learners with information about the available learning facilities and how to navigate through the environment (see Fig. 2).

The navigation scenes are designed to present several power systems facilities, which can be manipulated and explored to understand multiple power systems concepts.

For instance, in Fig. 4, a stylized version of the a electric substation is shown, whereas in Fig. 4 a real world microgrid is presented, where wind and a solar power generation can be manipulated by the learner.

The equipment understanding scenes, present power components in a catalog fashion where users can speak to the system to get information about electrical systems, as well as a general explanations of the physical phenomena under which they operate, depending on the given command.

4.2 CARLA Intentions Recognition Performance

We tested CARLA capabilities in terms of recognition of valid and invalid intentions. A valid intention (a true positive -TP-) is identified and attended given that a specific set of words are contained within a user utterance, whereas an invalid intention (a true negative -TN-) is a request that cannot be attended. In the case of the former, a valid intention would be "CARLA, what is a solar panel?"; in the case of the latter an invalid intention would be "CARLA, why does the sun shine?"; in both cases, CARLA must act accordingly otherwise it will incur in an error (a false negative -FN- if an intention is not recognized using the proper set of words, or a false positive -FP- if an intention is identified when it should not to).

Tests involved a sample of 5 people, 3 males and 2 females with an average age of 37.4 years. Each test subject spoke 40 phrases, of these 10 correspond to invalid intents that had similar pronunciations or letter combinations with valid intents. The results are shown in Table 1.

Table 1. CARLA's confusion matrix.

	True	**False**	**Total**
Positive	140	12	152
Negative	4	44	48
Total	144	56	200

As can be seen the accuracy (i.e., $\frac{TP+TN}{All}$) of the model is very high with 92%, with a True Positive Rate of 97% and a True Negative Rate of 78%. This means that a naive approach such as using just a set of words, has a high rate of detection for valid intentions but is less robust to invalid ones.

5 Conclusion and Discussion

In this work, we propose CARLA as a whole educational platform which grants access to: an interactive spoken conversational agent in virtual reality with NLP capabilities, a stylized recreation of a real electrical substation, a full-fledged and fully sized micro-grid which can be explored, and their corresponding equipment navigation scenes.

It is important to note that CARLA functionalities are only available with an active internet connection, which can create discontinuities in service availability and affect the users perception of the agent [4].

Experimentations regarding the chatbot main capabilities are promising. For the time being, we only analyzed the recognition of valid and invalid intentions using a naive approach. In terms of the overall accuracy, this approach achieved a 92%. Yet, valid intentions can be triggered using invalid utterances with similar pronunciations or letter combinations. Thus, next improvements to CARLA will be focused on the application of machine learning models to boost the recognition of intentions. Similarly, to detect emotional states relevant for learning, future work will be delve to include facial emotional recognition using Deep Learning techniques which excel in this task. These actions will improve CARLA interactions and will allow personalizing them to meet specific users' needs. In this sense, CARLA is meant to become, at some point in the near future, into an Intelligent Virtual Assistant (IVA) [1], a tutor or learning companion.

References

1. Chung, H., Lee, S.: Intelligent virtual assistant knows your life. CoRR abs/1803.00466 (2018)
2. Galitsky, B.: Developing Enterprise Chatbots. Springer International Publishing (2019)
3. Kerry, A., Ellis, R., Bull, S.: Conversational agents in e-learning. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 169–182. Springer (2008)
4. Klopfenstein, L.C., Delpriori, S., Malatini, S., Bogliolo, A.: The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In: Proceedings of the 2017 Conference on Designing Interactive Systems. pp. 555—-565. DIS '17, Association for Computing Machinery, New York, NY, USA (2017)
5. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
6. Pandey, A.K., Gelin, R.: A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. IEEE Robotics & Automation Magazine 25(3), 40–48 (2018)
7. Santamaría-Bonfil, G.: Towards an interactive learning ecosystem:education in power systems. Research in Computing Science 149(1), 6–7 (2019)
8. Santamaría-Bonfil, G., Ibáñez, M.B., Pérez-Ramírez, M., Arroyo-Figueroa, G., Martínez-Álvarez, F.: Learning analytics for student modeling in virtual reality training systems: Lineworkers case. Computers & Education p. 103871 (2020)

9. Schuurmans, J., Frasincar, F.: Intent classification for dialogue utterances. IEEE Intelligent Systems 35(1), 82–88 (2019)