Auto-generating Virtual Human Behavior by Understanding User Contexts

Hanseob Kim*[1,2], Ghazanfar Ali†[1,3], Seungwon Kim‡[1,4], Gerard J. Kim§[2], and Jae-In Hwang ¶[1]

[1]Korea Institute of Science and Technology, [2]Korea University, [3]University of Science and Technology, [4]Chonnam National University

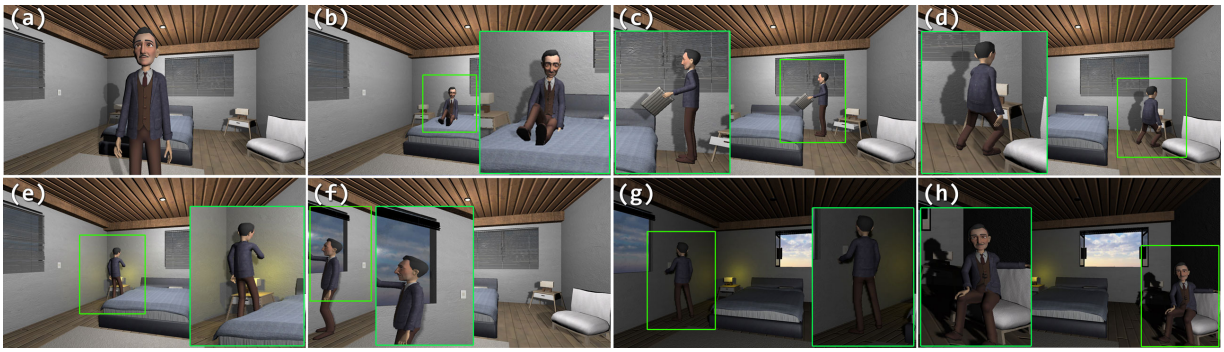


Figure 1: Automatically generated virtual human behavior according to the user contexts: (a) initial state, (b) lay on a bed, (c) bring a pillow, (d) open the drawer, (e) turn on a lamp, (f) open the window, (g) turn off the light, and (h) sit on a chair.

Abstract

Virtual humans are most natural and effective when it can act out and animate verbal/gestural actions. One popular method to realize this is to infer the actions from predefined phrases. This research aims to provide a more flexible method to activate various behaviors straight from natural conversations. Our approach uses BERT as the backbone for natural language understanding and, on top of it, a jointly learned sentence classifier (SC) and entity classifier (EC). The SC classifies the input into conversation or action, and EC extracts the entities for the action. The pilot study has shown promising results with high perceived naturalness and positive experiences.

Index Terms: Human-centered computing—Interaction design process and methods; Human-centered computing—Virtual reality; Computing methodologies—Natural language processing;

1 Introduction

Virtual Humans (VHs) are human-like characters that interact with the user verbally and gesturally according to the situational context. Ali et al. [1] introduced the system in which a VH verbally communicated with users and provided information. They used fixed word terms to generate actions, so users were required to know the terms beforehand. In another study by Chen et al. [3], a virtual crowd was simulated using simple text inputs. This system used a rule-based method and was unable to work with complex sentences. To remove the constraint of fixed phrases or work with complex sentences, the system needs more understanding of the context and semantics of input to predict user intentions. With the recent advancements in natural language processing, pre-trained models like BERT [4] can be used to extract contextual and semantic features of input text, which can then be further used to classify sentences and extract entities. Our research goal is to enable users to interact with a VH, thereby, without memorizing specific phrases.

*e-mail: khseob0715@kist.re.kr
†e-mail: aliust@ust.ac.kr
‡e-mail: Seungwon.Kim@jnu.ac.kr
§e-mail: gjkim@korea.ac.kr
¶e-mail: hji@kist.re.kr; Corresponding Author

We designed a controlled room scenario with one VH and a fixed number of objects. We also prepared a set of plausible interactions with each object. Based on the controlled list of objects and plausible actions, we created a dataset. Each data sample consists of input text and a set containing class of input and entity classes. In the case of input, which is meant to retrieve information, entity classes were set to none. With this method, our model is trained jointly learn the sentence class and entity classes with respect to input sentence. We conducted a pilot study where users were given a possible interaction set and were free to input whatever they want. Users reported high perceived naturalness and positive user experiences. The following sections describe implementations, experiments, and results in detail.

2 Implementations

Our system consists of two parts, a sentence classifier and VH interaction module. The sentence classifier is a learning-based system tasked with the classification of sentences and entities. The interaction module uses information from the classifier to generate and simulate the relevant VH's behaviors.

2.1 Sentence and Entity Classifier

To implement the prototype, we designated a small room as the interactive space with a VH. We then prepared a dataset with 1200 sentence samples of possible interactions in the room. To understand the user's intention to interact with a VH, as shown in Table 1, we defined four categories as follows: *Subject, Action, Position, and Target*. Subject is a binary category that is used to classify the sentence into a conversation or action. Action, Position, and Target consist of 14, 6, 11 classes, respectively. We trained four classifiers simultaneously on top of pre-trained BERT. BERT's weights were frozen during training. The implementation was done in Pytorch.

Table 1: Possible interaction vocabulary.

Subject	Action		Position	Target	
None (Small Talk)	None	Walk	None	None	Floor
Virtual Human	Open	Close	Left	Chair	Drawer
	Sit	Stand up	Right	Bed	Lamp
	Turn on	Turn off	In	Window	Curtain
	Lay	Run	On	Object	Pillow
	Idle	Bring	To	Switch	
	Hold	Put			

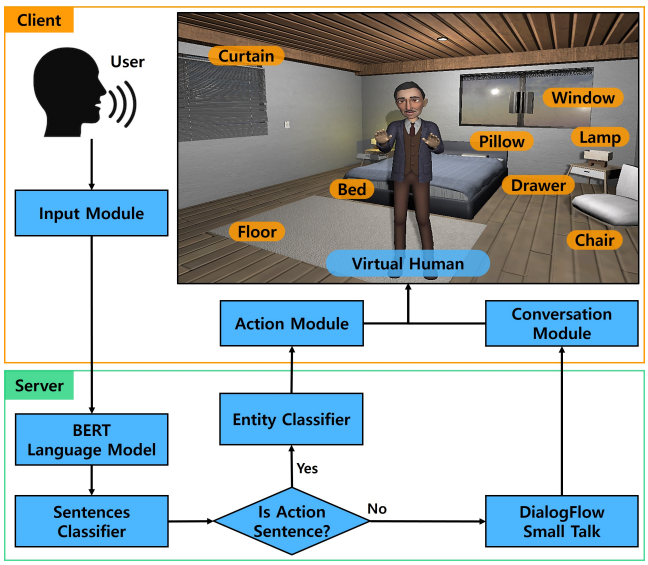


Figure 2: Overview of the system architecture.

2.2 Interaction with a Virtual Human

To provide users with automatically generated interactions from user's sentences, we developed an interactive system with a VH. Figure 2 details our system consisting of a client and server. We first prepared an input module that uses voice/text to communicate with users. If the input was a voice, our system converted it to text using a speech-to-text API from Google. Then we sent the converted text data to the server to understand user's intentions.

In the server, the user sentence was classified as conversation or action by SC. If the user's intention was to converse with the VH, the server used Google DialogFlow to get a simple answer to the user's sentence. Otherwise, if the user's intention was a request for specific actions, the user's sentence was classified by EC, as shown in Table 1. And finally, our server sent each result to the client.

When the client received entity pairs for specific actions, our system combined them to generate actions. For these, we prepared a series of simple character animations in advance. For example, the VH opens, sits, or pushes objects. Our system also had a set of information for each target object (e.g., name, position, direction). Thus, our system can generate new interactions that correspond to the user's intention by combining target information with animation.

3 Experiment

With our initial prototype, we conducted a short experiment to evaluate generation accuracy and gather diverse input sentences by users. We recruited 16 participants (15 male, 1 female, age 24-32, $M = 27.9$, $SD = 2.73$) who were familiar with English from a local community. We provided our system to participants as a web client. The web page included a brief description, and we instructed participants to interact as much as they wanted without a time constraint. During the experiment, we automatically collected user inputs and requested participants to evaluate the demonstrated interactions for each input sentence (1: successful, 2: failed, and 3: unnatural). At the end of the experiment, we asked them to answer the five questions to evaluate user perception with a 5-point Likert scale. From the related study [2], we extracted and modified questions in the following aspects: perceived naturalness, semantic consistency (see Table 2).

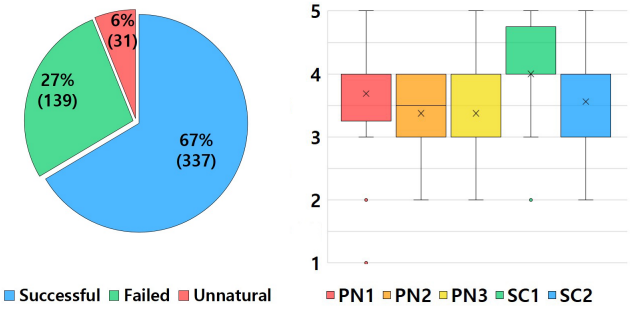
4 Results And Discussion

From 16 participants, we received a total of 507 sentences ($M = 36.29$, $SD = 16.48$). Figure 3 (a) shows the results of user evaluation for each interaction. The response distributions were 73% positive and 27% negative. The positive distribution was divided into 67% successful and 6% successful but unnatural interaction. In case of failure, users often tried more to test using alternative words and implausible (or unexpected) actions (e.g., *"Jump out from the window", "Throw the chair"*), which led to an increase in the number of failures.

Table 2: Questionnaire for perceived naturalness (PN1 to PN3) and semantic consistency (SC1, SC2).

PN1	Action was smooth
PN2	Action was natural
PN3	Action was comfortable
SC1	Action was matched to content
SC2	Action was same as I expected

Results on user perception show higher scores on perceived naturalness and semantic consistency. SC1 showed the highest score, which mean that the generated actions were well-combined from extracted entities. Although this experiment is limited, it provided us valuable feedback that system was indeed feasible and natural to use albeit some performance issues regarding execution of actions.



(a) User evaluation (b) User perception

Figure 3: Subjective measure results.

5 Conclusion and Future Work

We studied and designed an approach to seamlessly invoke conversation or action with a VH, without any predefined phrases and in semantic-aware manner. Our study results indicated a comfortable and natural interaction by users. In future work, we plan to prepare a much larger dataset, replace the classifiers with decoder to cater for complex multiple sentence input.

Acknowledgments

H. Kim and G. Ali contributed equally to this paper. This work has supported by the National Research Council of Science & Technology grant by the Korea Government (MSIT CRC-20-02-KIST).

References

[1] G. Ali, H.-Q. Le, J. Kim, S.-W. Hwang, and J.-I. Hwang. Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable mixed reality environment. In *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*, pp. 47–52, 2019.

[2] G. Ali, M. Lee, and J.-I. Hwang. Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds*, 31(4-5), 2020.

[3] C.-Y. Chen, S.-K. Wong, and W.-Y. Liu. Generation of small groups with rich behaviors from natural language interface. *Computer Animation and Virtual Worlds*, 31(4-5), 2020.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.