Improving Virtual Suspect Interaction Using Alexa

Gonçalo Baptista[1], Diogo Rato[1,2], and Rui Prada[1,2]

[1] Instituto Superior Técnico, Universidade de Lisboa
{goncalo.baptista,diogo.rato,rui.prada}@tecnico.ulisboa.pt
[2] GAIPS, INESC-ID

Abstract. Interactive Agents need a good interaction in order to showcase their abilities and fulfill their roles. Rato et al. created a Virtual Suspect capable of lying, but its interaction was limited. We took advantage of the tools provided by Amazon Alexa to create a new natural language conversational interaction with the Virtual Suspect. We used an iterative, user-centered approach when designing the new interaction, collecting feedback and data from User Studies in order to improve the interaction with the Virtual Suspect. After we managed to create a good interaction we did another User Study with the original lying algorithm and concluded it still needs improving.

Keywords: Virtual Suspect · Alexa · Natural Language · Conversational Agent · Conversational Interaction

1 Introduction

Interactive Agents can perform a wide variety of roles in our modern world. They can be used for both entertainment and education purposes. Systems like this can be a useful teaching tool because they are less expensive, more accessible and offer increased control of the environment [13]. Whatever their function or context, the quality of an Interactive Agent depends mainly on the quality of its logic and of its interaction. An Interactive Agent needs both good functioning to fulfill its role (whether as a museum guide or a shopping assistant), and needs good interaction to be able to understand and be understood (whether by a visual interface or natural language communication).

These Interactive Agents can sometimes be used to train a person for a specific role or job. One such example of that is a Virtual Suspect[2, 1], an agent that inhabits the role of a suspect in a police interrogation, that can be used to train police officers and detectives in interrogative techniques. Of course, the same technique could also be used for entertainment purposes as part of an investigative video game.

Another technology that has been consistently evolving and becoming more pervasive is Natural Language Understanding and Voice Interaction. Voice Assistants especially are becoming more and more ubiquitous, being present in our homes and our phones, with one such example being the Amazon Alexa, a voice assistant developed by Amazon. Not only is it an extremely versatile voice assistant, it also allows users to create their own functionalities through the form of third-party applications. Truly, it is now easier than ever for people to create interactive and conversational agents.

1.1 Motivation

In 2016, Rato at al.[9, 10] designed and developed a model of a Virtual Suspect with the ability to autonomously create parallel stories to the one initially stored in its memory, and thus allowing it to lie. In order to test their Virtual Suspect, they developed a simple and limited visual interface that allowed users to select questions to ask the Virtual Suspect.

In their conclusion, Rato et al.[9] posited that their interface was too limited and that an approach using Natural Language Processing would highly improve the interaction between the user and the Suspect.

1.2 Problem

The original interaction with the Virtual Suspect[9] was too limited. It only had a pre-defined small number of questions that users could choose from, and did not allow them to ask anything else. It did not cover or showcase the full capabilities of the Virtual Suspect model designed by Rato et al., as it omitted certain types of questions altogether. It did not have a sense of progression or finality, as the users could only ask the same questions over and over again. The order did matter, and sometimes the answers could change, indicating that the agent had previously lied, but there was no flow, and no natural stopping point to the interaction.

1.3 Objective

Our goal is to create a better interaction with the Virtual Suspect designed by Rato et al.[9], by creating a new Natural Language conversational interaction. We will take advantage of the tools provided by Amazon Alexa to create that new interaction. We want to provide a better User Experience (UX) for interacting with the Virtual Suspect, by creating a natural and fluid conversation with the agent, where users can naturally flow from one question to the next at their own pace. Our interaction will give users freedom in interacting with the Virtual Suspect, and will fully showcase all its the capabilities. We want to create an interesting, meaningful and user-driven interaction with the Virtual Suspect.

2 Related Work

As research for our work, we looked at several other works with some similarity to ours.

We reviewed how other Virtual Suspects were defined[1, 2], with focuses on the psychological state of the Suspect, or on different personality models, and how those aspects affected the way the agents responded.

We looked at other Conversation Agents, and studied how to model the internal mental state of the agents[7], how to model context in a conversation[6], and how to model knowledge representation and user's perception of that knowledge[4].

We also looked at the Alexa Prize[8], an Amazon competition to design open-domain socialbots using Alexa technology, but concluded that their models[3] were too complex for us to base ourselves off of.

We could not find anything too similar to our work, so we had to create our model from scratch.

3 Virtual Suspect

In order for us to create a new interaction with the Virtual Suspect, we have to first understand how it was designed and how it works. In "Virtual Suspect - A Lying Virtual Agent" Rato et al.[9] laid out the architecture and functionality of their lying Virtual Suspect.

First, the agent has a memory, its **Knowledge Base**, which contains its story. The agent's story is composed of *events* and *entities.* Entities are the most basic memory fragment, and can represent people, locations, objects, time spans. Events represent distinct episodes in the agent's story and they are composed of an Action and several entities, in different roles. For example, *"John stole a chocolate from the store on September 5th at 4:30 pm"* can be an event where *"Steal"* is the Action, and *"John"*, *"chocolate"*, *"store"* and *"September 5th at 4:30 pm"* are all represented by entities. Entities can have different roles in events. In the previous example, those roles were Agent, Theme, Location, and Time, respectively, but you can also have Manner and Reason. These roles indicate the relation between those entities and the Action in that event, *who* was involved, *what* was the target of the action, *where* it happened, *when* it took place, *how* it happened, and *why* it happened. Events and entities exist separately in the **Knowledge Base**, and events reference the entities that were involved in them, this way the same entity can be referenced by (and thus have participated in) several events. Events can also be true or false, where true events are what really happened in the agent's story and false events are the events the agent uses to lie. These can have an incriminatory value from 0 to 100, depending on how incriminatory each event is, in relation to the crime our Virtual Suspect committed.

The interaction between the agent and the user is done through questions and answers. The user asks the agent a question about its story and the agent responds, with either the truth or a lie (as we will see later). These questions are internally represented as *queries* in the agent's system, and they can either be Validation questions (yes or no), or Information Gathering questions (who, where, when, why, etc.). Each query contains a series of conditions which it seeks to validate in order to find an answer. For example, the question *"When did John steal the chocolate?"* is a Information Gathering question that seeks to retrieve the Time entity from an event that matches the conditions *"Agent Equals John"*, *"Action Equals Steal"* and *"Theme Equals Chocolate"*. The **Query Engine** receives this query, tries to find all the events that match those conditions, and returns a *query response*, which in this case contains the value *"September 5th at 4:30 pm"*. After this step, the Virtual Suspect also contains a **Natural Language Generator** that transforms the query result into a proper English sentence to be returned to the user as its answer.

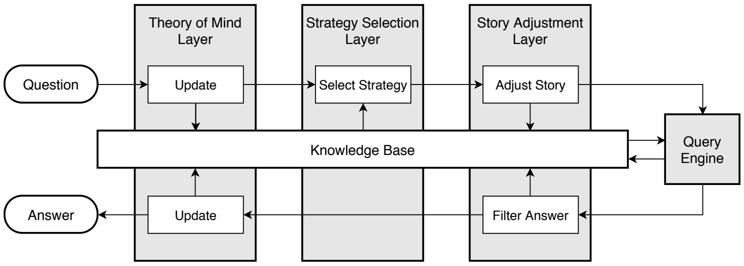

Fig. 1. Virtual Suspect Architecture

What enables the agent to lie is a three layered two-pass control system, as illustrated in Figure 1. When the agent receives a question, it passes through each one of the three layers before being processed by the Query Engine, and the answer passes through the layers again before being returned to the user. The **Theory of Mind Layer** keeps track of what the user already knows about the story, by analysing the information contained inside the query. If the user already knew about John stealing the chocolate, for example, it would not be productive to try to lie about that. The **Strategy Selection Layer** selects an appropriate lying strategy based on the current context, and the **Story Adjustment Layer** creates the fake events that the agent uses in its lies. When the agent encounters a question that would lead it to reveal incriminating information, it instead creates a new fake event with less incriminating information to take the place of the incriminatory event in the version of the story the agent is presenting the user. The agent always keeps track of the true version of events, but is capable of having alternate versions of those events in its memory in order to hide information from the user. After the question has passed through all the layers, it is processed by the Query Engine, and thus the information about the fake events is retrieved instead of the real information, and the result then passes back through the layers again, before being returned to the user.

Figure 2 shows how this was implemented in the original work[9], with the **Response Model** representing the conjunction of the Query Engine and the three layers. The prototype that was originally used to test the Virtual Suspect was a visual interface that contained information about the suspect and the

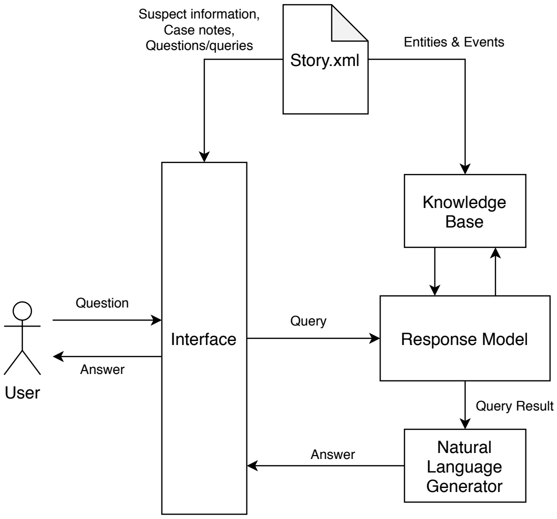

Fig. 2. Virtual Suspect Prototype Implementation

case, and a set number of pre-defined questions that users could select from. All this information, along with the events and entities of the agent's story were all defined in a separate Story file. When the user selected one of the pre-defined questions from the visual interface, it automatically sent the corresponding query to the Response Model, which was then processed as previously described and the answer was displayed back to the user in the interface.

4 Alexa

The Alexa is a virtual assistant developed by Amazon and released with the Echo smart speaker, that is capable of a wide range of features, but the one that is of interest to us is the ability to create third-party applications using the Alexa technology, called Skills. These Skills are made through the Alexa Skills Kit (ASK) and they have two components: the **Interaction Model**, and the **Skill Service**. Figure 3 shows the typical workflow of an Alexa Skill. The user asks a question or gives a command to Alexa, which sends it to the Skill Interaction Model. The Interaction Model disambiguates the meaning of the user's message and sends that information to the Skill Service, which computes the appropriate response and sends it back through the system until it reaches the user.

The **Interaction Model** is the front-end of the Skill, and it is composed of *intents*. An intent contains a selection of sample phrases that could be uttered to invoke that intent. For example, a *HelloWorldIntent* could contain the utterances *"Hello"*, *"Hi World"*, and *"Hey"*, so that when the user says one of these phrases or something similar, Alexa can correctly identify the *HelloWorldIntent* and provide the proper response. The more sample phrases an intent has, the more

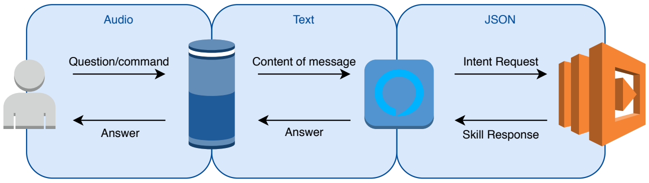

Fig. 3. Workflow of an Alexa Skill

accurate Alexa can be when detecting it, as Alexa trains a model with the information in our Interaction Model to be able to detect meaning from a wide variety of phrases, although the exact method is not publicly known.

Besides having a set of sample utterances, an intent can also have *slots*, which are essentially variables that can be fulfilled by certain values. For example, we can have the sample phrase *"My name is*{name}*"*, where {name} represents a slot that accepts English first names as values. This way, both the sentences *"My name is John"* and *"My name is Mary"*, would equally fulfill that intent. A slot type can be one of many provided by Amazon (like the First Name slot type), or can be a custom list of possible slot values according to the skill's domain. These slot values can also contain synonyms. Slots cannot be iteratively defined, so a slot cannot contain another slot.

This information (intent and slot values), once processed by the Interaction Model, is sent to the **Skill Service**, which is the back-end of the Skill, through a JSON file. The Skill Service takes the information sent by the Interaction Model and computes the appropriate response (for example, *"Hello John"*), and sends it back to the Interaction Model through another JSON file.

5 Solution

In order to create the new Natural Language interaction with the Virtual Suspect, we combined what we studied in the previous sections to create a Virtual Suspect Skill. Our Interaction Model has different intents for the different question types, and we use slots to create the query conditions. Each of our intents needs lots of sample utterances so our model can cover a wide range of questions, and our slot values include the possible entity values for each type. Our Skill Service was created in the same environment as the original Virtual Suspect was developed, so we can use the original Virtual Suspect Response Model as a sort of code black box. The Skill Service takes the intent and slot information from the Interaction Model and uses it to create a query object that can then be sent to the Virtual Suspect Response Model. We also use the Virtual Suspect Natural Language Generator to transform the query result returned by the Virtual Suspect Response Model into a proper answer, before returning it to the Interaction Model.

Figure 4 shows a representation of our Virtual Suspect Skill, showing the Skill Service interacting with the Virtual Suspect modules, and combining what

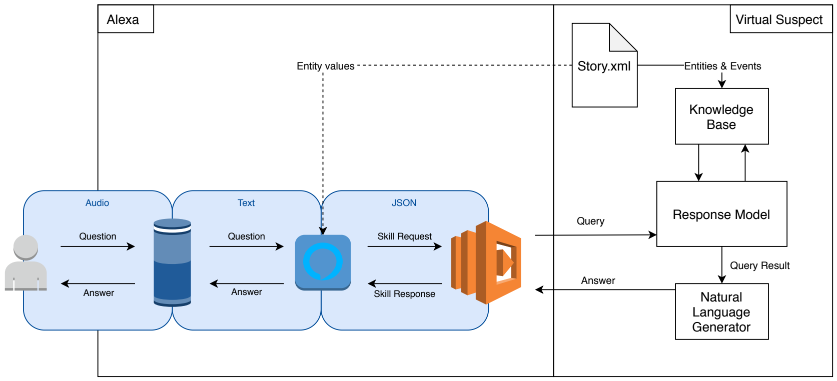

Fig. 4. Solution Architecture

we had already seen in Figures 2 and 3. The connection between the Story file and Interaction Model is merely symbolic, as we cannot directly connect those two entities, but it represents the entity values that populate the slot values.

We used an iterative, user-centered approach when designing the Virtual Suspect Skill. We started by recreating the functionality of the original prototype, and ensuring basic coverage for all the different types of questions, and then we did a User Study to collect data on how users interacted with the Virtual Suspect, what kinds of questions they wanted to ask and how they asked them. We also measured the performance of the agent, collecting data on the problems of the interaction, so we could have a baseline performance to compare to later.

After we collected the data from the First User Study, we used it to improve upon the interaction, making changes and improvements to fix those issues. We then conducted a Second User Study after those changes, to validate if those changes had improved the quality of the interaction, and to measure the User Experience (UX).

During the development and first two User Studies, we kept the lying component of the agent deactivated, so we could better measure its responses to the various questions without the lies obscuring that information. After we did the Second User Study, we turned the lying component of the agent back on and did a final User Study to measure the UX of that interaction, to see how well the original lying algorithm fit into the new interaction.

The next sections describe the development process of the Virtual Suspect, and the three User Studies, respectively.

6 Development

6.1 First Steps

In order to be able to do our fist User Test and collect data on how people interact with the Virtual Suspect, we needed a functional Virtual Suspect Skill prototype. We started off by replicating the functionality of the original Virtual Suspect Prototype[9] whose visual interface only included 13 pre-defined questions. Since those questions were already predefined and the corresponding queries would directly be sent to the Virtual Suspect Response Model (as seen in Figure 2), there was no concern about being able to interpret those questions using Natural Language Processing, and such they did not conform to a consistent style, often having sentences before the question and information that was not relevant for the query. In order to recreate the functionality of being able to ask those original questions (or their corresponding queries at least) and obtaining the same answers, we had to restructure the questions into a more consistent style that we could then expand to the rest of the question types in our Interaction Model.

We ended up with a style where a question like *"Where did you meet John Frey?"* was modeled as *"Where* {question_verb} {subject} {filler_verb} {agent}*"*. In this example, *"Where"* indicates the type of question, {subject} and {agent} are slots that contain information relevant for the query conditions, while {question_verb} and {filler_verb} are slots that allow for a wider range of questions with the same meaning to be identified. This way, questions like *"Did you meet John Frey?"* and *"Have you met John Frey?"* can both be represented by the same utterance, as they both have the same meaning.

After we established a consistent style of question, and managed to recreate the functionality of the original 13 questions, we expanded our Interaction Model to include more questions of each type, by looking at the events of our story and figuring what types of questions could be asked, with which conditions. As we mentioned before, this was done with the lying component turned off, so we could better understand how the agent was processing the information. With a lying agent it would be more difficult to tell if the agent answered *"No"* because he understood the question and decided to lie, or if he did not understand the question at all.

With this functioning prototype, we realized our First User Study to collect data on how people interacted with the Virtual Suspect, so we could expand our Interaction Model with more possible questions, and to measure the performance of the agent, so we could note the problems with the interaction and work to improve it.

6.2 Improving the Interaction

With the data collected from the First User Study, we were able to make a lot of changes and improvements to the Virtual Suspect Skill, to address problems such as:

- **Missing intents:** questions that the users wanted to ask but the agent was not capable of answering.
- **Pronouns:** both direct pronouns (it, him, there) and indirect pronouns (something, anyone).
- **Context:** a knowledge of what was previously asked.

- **Synonyms:** adding more synonyms to the Interaction Model.
- **Missing information:** information about the story that users wanted to know about but it was not represented in the story.
- **Answer generation:** improving the Virtual Suspect Natural Language Generator.
- **Time conditions:** add more cases for different possible time conditions in questions.
- **More utterances:** add more variety of questions to the Interaction Model.
- **Filters in the Skill Service:** to make sure that things are being processed correctly.
- **Feedback:** providing better feedback to the user when the agent cannot answer a question for some reason.

By addressing these and other problems and making all the necessary changes to the Virtual Suspect Skill, we were able to improve the interaction with the Virtual Suspect. We realized the Second User Study in order to verify that improvement and measure the quality of the interaction.

6.3 Last Adjustments

After we validated the improvements we made with the Second User Study, we made a few final minor adjustments before turning on the lying component again and making sure it was still working as intended with all of our changes. After that, we moved on the the Third and final User Study, to test how the lying algorithm was working with the new interaction.

7 User Studies

We conducted three User Studies during the development of our work. In all three studies, users interacted with the Virtual Suspect via a text messaging service, where an account in the name of the Suspect was created to add to user immersion. For all three studies, the conversations between the users and the agent were logged and annotated, in order to identify the problems with the interaction.

For the Second and Third User Studies, a questionnaire was presented to the users after the interaction to measure the User Experience (UX), which used the User Experience Questionnaire (UEQ) developed by Schrepp et al.[11, 12].

7.1 First User Study

For the First User Study, we wanted to collect data on how users interacted with the Virtual Suspect (what kinds of questions they asked and how they asked them) and do an analysis of the problems with the interaction.

Since the interaction was still in an early state and the range of the agent's understanding capabilities was limited, we decided to do two different types of interactions. One where users would interact directly with the Virtual Suspect Skill, and another where users would instead interact with a human pretending to be the Virtual Suspect, answering questions as the Virtual Suspect ideally would without the Skill limitations (following a Wizard-of-Oz technique[5]). This way we could collect data on the problems of the current interaction with the Skill, but also analyse how people would interact without those limitations. When logging the conversations, both types of responses were recorded for either type of interaction, allowing us to do a comparison between what the user actually interacted with and what the other interface would have said instead.

12 people participated in this study, with an average of 23.67 exchanges per conversation with the Virtual Suspect. The average conversation success rate (which is the percentage of exchanges that were correctly identified by the agent) was 37.29%. We classified problems in two categories: Question Problems (which was anything that caused the question not to be recognized by the agent), and Answer Problems (which was anything that cause the agent to give a bad answer). Only 35.92% of exchanges did not have any Question Problems, while only 22.54% did not have any Answer Problems.

Overall, the results were not very good, as we could somewhat expect given the interaction was still in a very early stage. But we were able to collect a lot of data on the problems of the interaction, as well as how users interacted with the Virtual Suspect, and we obtained a baseline performance with which to compare to in the future, so this was a successful study.

7.2 Second User Study

The Second User Study was realized after the changes and improvements made as a result of the data gathered in the First User Study, with the objective of validating those changes, verifying the improvement with the interaction, and measuring the UX.

14 people participated in this study, with an average of 54.07 exchanges per conversation, and an average success rate of 63.39%. This time, 65.13% of exchanges did not have any Question Problems, and 89.83% of exchanges not having any Answer Problems.

These results were a marked improvement over the First Study, validating the changes we made to improve the interaction. Not only that, but the UX results were also very good. Figure 5 shows our results compared to the benchmarks set by the authors of the UEQ[11], and we can see that all of them fall either into the Good or Excellent category.

This Study was a success, as were able to definitively show the improvements we made to our Virtual Suspect Skill, validating our previous choices, and we were able to show that our Skill provides a good User Experience.

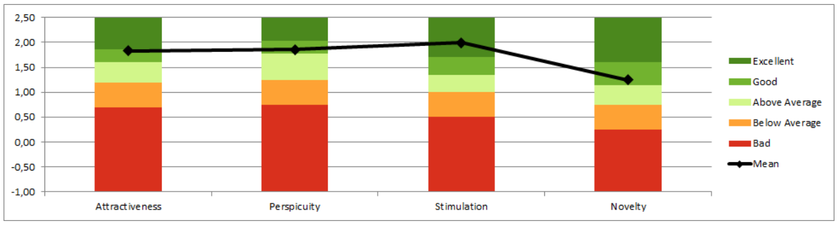

Fig. 5. UEQ Benchmarks

7.3 Third User Study

The Third User Study was conducted after we turned on the lying component of the Virtual Suspect and its objective was to measure the effect it had on the interaction, to see how well it was working.

With 16 participants, there was an average of 46.13 exchanges per conversation, and a success rate of 65.01%. 66.80% of exchanges did not have any Question Problems, and 88.48% did not have any Answer Problems.

The agent performance was largely the same as the Second Study, with the difference in the average number of exchanges being explained by the fact that the interactions of the Second Study were more free and exploratory, while in the Third Study they were more focused on the crimes of the Suspect.

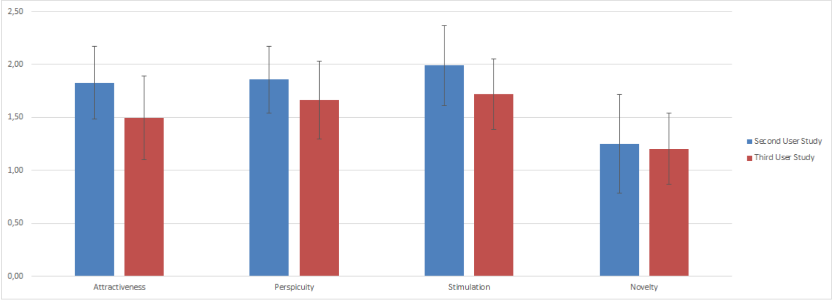

Fig. 6. Comparison of UEQ Results

The UEQ results were more telling, with Figure 6 showing that the results in the Third Study were noticeably worse than the Second Study. Given that agent performance remained at about the same level, and given the feedback we received about the agent's lies not being very believable or realistic, it is safe to conclude that it was the introduction of the lying component that caused this drop in UX. Therefore, as it stands, the current lying algorithm is not very suited for this new conversational interaction.

8 Discussion

Throughout our work, we faced several constraints in different areas that kept us from improving the interaction as much as we could.

Constraints with the Interaction Model, like the inability to have slots inside of slots, led to an inflation of utterances with the same meaning in our Interaction Model, and kept us from achieving a degree of nuance that would have allowed even more questions to be recognized. We believe a different natural language model, something more non-deterministic and grammar-like, could be beneficial in achieving an even better interaction.

There were also constraints with the definitions of the Virtual Suspect Architecture, like the way that Time and Reason entities were defined within events, that kept us from being able to achieve a more realistic story that would have matched the more natural interaction we created. A restructuring of the agent's memory, keeping the same basic concepts but improving upon them, could be beneficial in achieving a more realistic story, which when combined with a more natural interaction could lead to even better User Experience.

And finally, as we saw in the results of our Study, the way the lying algorithm currently works does not fit well with our new conversational interaction. As this was not the focus of our work, we did not make any changes to what was originally implemented, but we believe that a more sophisticated lying algorithm could very much improve the quality of the interaction with the Virtual Suspect.

9 Conclusion

In summary, this was a very experimental work that followed a heavy user-centered approach in pursuit of our objective of improving the interaction with the Virtual Suspect. Whether we were successful in achieving our goals comes down to whether we:

- were able to overcome the limitations of the original Virtual Suspect interaction[9];
- managed to create a Natural Language interaction that showcased the capabilities of the Virtual Suspect;
- created an interaction with good UX.

For each of those points our results were positive. We were able to create a natural and open interaction with the Virtual Suspect, using Natural Language, that showcased all of its capabilities (even adding new ones in the process), and we were able to vastly improve the quality of the interaction and achieve a good UX while doing it. On top of achieving our goal of improving the interaction with the Virtual Suspect, we were able to test whether the original lying algorithm[9] was suited to this type of interaction, and concluded that it needs further improvement.

For future work, there are many things that could be tried in order to further improve the interaction with the Virtual Suspect. Examples include a new Natural Language Model, a restructuring of the agent's memory, a new and improved lying algorithm, and even new stories and characters.

References

1. Bitan, M., Nahari, G., Nisin, Z., Roth, A., Kraus, S.: Psychologically based virtual-suspect for interrogative interview training. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
2. Bruijnes, M., Wapperom, S., op den Akker, R., Heylen, D.: A virtual suspect agent's response model. Affective Agents p. 17 (2014)
3. Chen, C.Y., Yu, D., Wen, W., Yang, Y.M., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., Iyer, S., et al.: Gunrock: Building a human-like social bot by leveraging large scale real user data (2018)
4. Falk, J., Poulakos, S., Kapadia, M., Sumner, R.W.: Pica: Proactive intelligent conversational agent for interactive narratives. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. pp. 141–146. ACM (2018)
5. Hajdinjak, M., Mihelic, F.: Conducting the wizard-of-oz experiment. Informatica (Slovenia) **28**(4), 425–429 (2004)
6. Kenny, I., Huyck, C.: An embodied conversational agent for interactive videogame environments. In: Proceedings of the AISB'05 Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction. pp. 58–63 (2005)
7. Morris, T.W.: Conversational agents for game-like virtual environments. In: AAAI 2002 spring symposium on artificial intelligence and interactive entertainment. pp. 82–86 (2002)
8. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., et al.: Conversational ai: The science behind the alexa prize. arXiv preprint arXiv:1801.03604 (2018)
9. Rato, D., Prada, R., Paiva, A.: Virtual Suspect. Master's thesis, Instituto Superior Técnico (October 2016)
10. Rato, D., Ravenet, B., Prada, R., Paiva, A.: Strategically misleading the user: Building a deceptive virtual suspect. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. pp. 1711–1713. International Foundation for Autonomous Agents and Multiagent Systems (2017)
11. Schrepp, M.: User experience questionnaire handbook. All you need to know to apply the UEQ successfully in your project (2015)
12. Schrepp, M., Hinderks, A., Thomaschewski, J.: User experience questionnaire. Mensch und Computer 2017-Tagungsband: Spielend einfach interagieren **17**, 355 (2018)
13. Sklar, E., Richards, D.: The use of agents in human learning systems. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. pp. 767–774. ACM (2006)