A Multimodal Chatbot System for Enhancing Social Skills Training for Security Guards

Stein de Bever[1], Daniel Formolo[1(✉)], Shuai Wang[1], and Tibor Bosse[2]

[1] Department of Computer Science, Vrije Universiteit, Amsterdam, The Netherlands
d.formolo@vu.nl
[2] Behavioural Science Institute, Radboud Universiteit, Nijmegen, The Netherlands

Abstract. Chatbots are typically used in dialogue systems for various purposes such as customer service and information acquisition. This paper explores enhancement of social skills training for security guards with the use of chatbots. More specifically, we designed a chatbot using text and voice as input to study the acceptance and the impact of the system to training security guards in deal with stress situations. The result of a pilot experiment and a survey are presented and discussed. Finally, we discuss possible improvements and future work.

Keywords: Chatbots · Serious games · Training · Role-playing · Security employees · Intelligent agents

1 Introduction

Essential skills dealing with inter-personal connections and social interactions can be crucial in the success of many jobs. Good communication and interpersonal skills are therefore essential. However, to improve such skills, one needs to learn from practice rather than books. Typically, companies employ professional actors and deliver role-playing sessions for the training of new recruits, which can be costly, time-consuming and hard to organize. Alternatively, training based on serious games is less costly. Serious games make use of conversational agents, also known as chatbots. Most chatbots uses auditory or textural input/output while some more advanced ones uses both with avatars. These programs are typically used in dialogue systems for various purposes such as customer service and information acquisition. Although it costs a lot to develop such a system, the advantage is that they can be scaled to simultaneously interact with a large group of users, making it viable. In addition, it is local and time independent, which reduces the requirements for interaction, making it more flexible and adaptable than traditional methods. For these reasons, chatbots can potentially be used in a wider variety of instructional situations [5]. In this paper, we study how the use of chatbots for social skills training can make a difference in engagement and knowledge gain of security guards candidates[1].

[1] This paper is based on the bachelor thesis of Stein de Bever.

M. Kurosu (Ed.): HCII 2019, LNCS 11566, pp. 499–513, 2019.
https://doi.org/10.1007/978-3-030-22646-6_37

We designed and implemented the chatbots on campus and conducted our experiment and collected the data at a local mid-scale company named Workrate[2] with the aid of their staff. The company has 634 employees and provides security service for ports, air cargo, company offices and their data centres. Many of the employees are university students who work part-time to cover duties such as reception, control rounds, mobile surveillance, etc. A security guard interact with officers, customers and staff during his or her daily duties. Such duties can be demanding on inter-personal interaction, especially in case of criminal cases or potentially harmful occasions. Improper means of interaction may cause harm to staff, victims or even results in life threatening scenarios and cause damage properties and reputation. New recruits are therefore required to pass a practical exam to demonstrate essential skills in the form of three role-playing games before they get on duty. To improve the passing rate and reduce the cost in training, the companies offer trail role-playing training sessions. These sessions with professional actors can be costly, time-consuming, location-dependent, and therefore hard to organize. The company suffers from low passing rate. The interpersonal skills and the result of such training differ significantly from person to person due to the lack of practice. Since such role-play dialogues are hard to practice alone and practical sessions with actors have several drawbacks, it is natural to consider chatbots as an alternative. Despite chatbots are interactive, less costly and location-independent, they have their own drawbacks. They can suffer from imperfection in the interpretation of voice and the lack of emotion. In this paper, we intend to explore the following research question, which infers two sub-questions:

- Can social skills training for security guards be improved by using chatbots?
  - To what extent do people accept chatbots as tools for social skills training?
  - Which means of interaction with chatbots has better training effects (text input or free speech, or combined)?

In the following section, some background information and related literature are presented. Sections 3 and 4 are the design and the implementation of our chatbots respectively. The results are presented in Sect. 5. Finally, in Sect. 6, we discuss the results and outline some future work.

2 Background and Related Work

Chatbots are interactive computer programs that takes auditory or textual inputs and respond with informative answers. Thanks to the advance in Natural Language Processing in recent years, chatbots embrace applications in customer service, information acquisition, education, professional training and so on.

In the study by Bayan et al. the potential use of chatbots in education is addressed [9]. In later research conducted by Hoffmann et al. [5], a virtual assistant support students with the basics of a study area. That gives more time to

[2] https://www.workrate.eu/en/.

the teacher focuses on more complex topics of that area. They further compared to traditional e-learning systems and studied features in providing knowledge in this interactive way. Both state that chatbots had positive impact on the results of the student [5,9]. Abbassi et al. [1] studied learning with chatbots vs. conventional search engines for the teaching of Object-oriented Programming. They concluded that the learning outcomes by using chatbots are significantly better when compared with learning through conventional search engines. In addition, chatbots have also been used in medical domains to deal with depression [2] and stress [6]. Some distinct chatbots can make use of emotions to provide psychiatric counseling service in mental healthcare [7]. Closer to our research is the case study by Bosse et al. [4] where they studied how conversational agents can be used for public transport employees in dealing with aggressive customers. The case study showed that the employees managed to enhance their social skills after training with conversational agents. As a result, the training employees manage stress situations more effectively.

In contrast with existing work, we focus on engagement and the comparison of the efficiency of text and auditory input for social skills training in contexts where people confront others and have to make decisions under pressure. We also study how adding chatbots to the classical approach in training would improve the passing rate. The next section describes the design of our chatbots for the training of security guards trainees, which we will refer to as users.

3 Chatbot System Design

In this project, two versions of chatbots have been developed using the Watson Assistant[3]. One using voice and text and another using only text. The voice component approximates users of real scenarios. It also includes time pressure to the user replies to the chatbot. By the other hand, translating voice in text can inject wrong inputs to the chatbot by imprecise algorithms, untypical accents or environmental noise. Those disturbs can affect the user experience. Moreover, because the lack of time pressure in speak and finalize the answer in one shot, dummy users would rather start by texting then using voice at beginning. Therefore both voice and text might be considered to study the efficiency of the chatbot applied to this context. The aim is to assist users to learn how to deal with certain scenarios in security service. In the case of the voice and text approach, the idea is that voice is the main input and the text is an alternative channel to the user.

Despite the voice and text inputs, the core of the chatbot shares the same *dialogue tree*. In a *dialogue tree* an answer has to be given to progress from one node to another. According to the user's input and the conditions of the node, the next node is determined. Figure 1 is an example of a *dialogue tree*.

At each round of the conversation, there are 3 types of response: a correct response, a partially correct response and an incorrect response. An incorrect response will redirects to a node that gives the right answer for that round of

[3] Formerly Watson Conversation: https://www.ibm.com/watson/ai-assistant/.

interaction. This prevents the user from restarting the entire use case due to one mistake. In addition, the node redirects back to itself which forces the user to fill in the right answer before continuing. For every partially correct answer, there is a separate branch in the *dialogue tree* which eventually reaches the final node. This branch would lead down to a separate final node. This would be very time consuming as there are partially correct answers at almost every step. For the clarity of the experiment, the complete list of necessary steps at the end of every use case is provided. The users are instructed to check if they take every right step and self-monitor the learning progress.

The system runs on an on-line platform. The interaction with the chatbot is in Dutch. Appendix B contains an sample use case provided by the company. Figure 2 depicts an example of a conversation. To reduce misunderstanding and improve the usability of the system, two instruction web-pages were created with a brief introduction to the experiment. Each version starts with an introduction video per version of the chatbot was embedded to explain the details of the experiment respectively. Further down the web-pages are the links to the chatbot and the corresponding use cases. More details of the web-page and chatbot can be found in Appendix A. The next section provides more details about the implementation and the experimental setup.

4 Research Method

The chatbots have been evaluated in a pilot experiment that consisted of two parts, to which we will refer as separate experiments for practical reasons. In Experiment 1, the objective learning effect of both variants of the chatbot were evaluated. This was done by comparing the skills of a group of users who practiced with the combined (voice and text-based) chatbot with a group of users who used the text-based chatbot (Experiment 1A), as well as with a control group, which are users who did not use any chatbot (Experiment 1B). In Experiment 2, a survey was used to measure the subjective acceptance of people who worked with the chatbot. The design of both experiments is described below. Together with a description is provided of the material that was used for both experiments.

4.1 Material

The use cases selected for our study have been inspired by the material that the users need to study for their final exam. In total, 36 use cases have been provided by the company. These use cases have some similar steps in their structure. At the same time, diverse use cases were chosen that treat different events. For instance, two use cases cover attempted arson, another covers the case with a person trapped in the elevator, while another covers the event of an incorrectly parked car. This was done to show the participants that, across a variety of circumstances, similar steps have to be undertaken to reach the desired result.

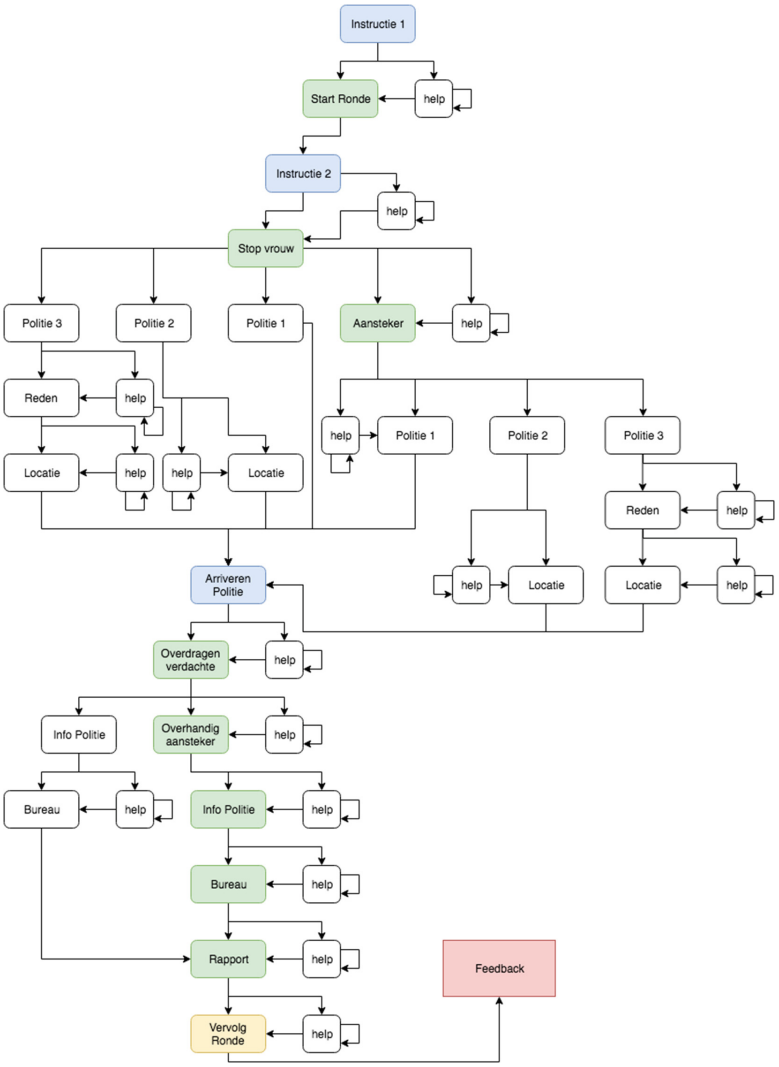

Fig. 1. Dialog tree of use case 1

Nevertheless, the users do not have any guarantee that the same use cases will be used for training as for their final exam.

To pass the practical exam, users have to be familiar with some general protocols that are reflected in many use cases. Some of these protocols can be applied in many different events. For example, some use cases expect the same actions at the beginning with the user saying: 'Security to Central Post, I am starting my round'. Another example is that, when speaking about an unknown

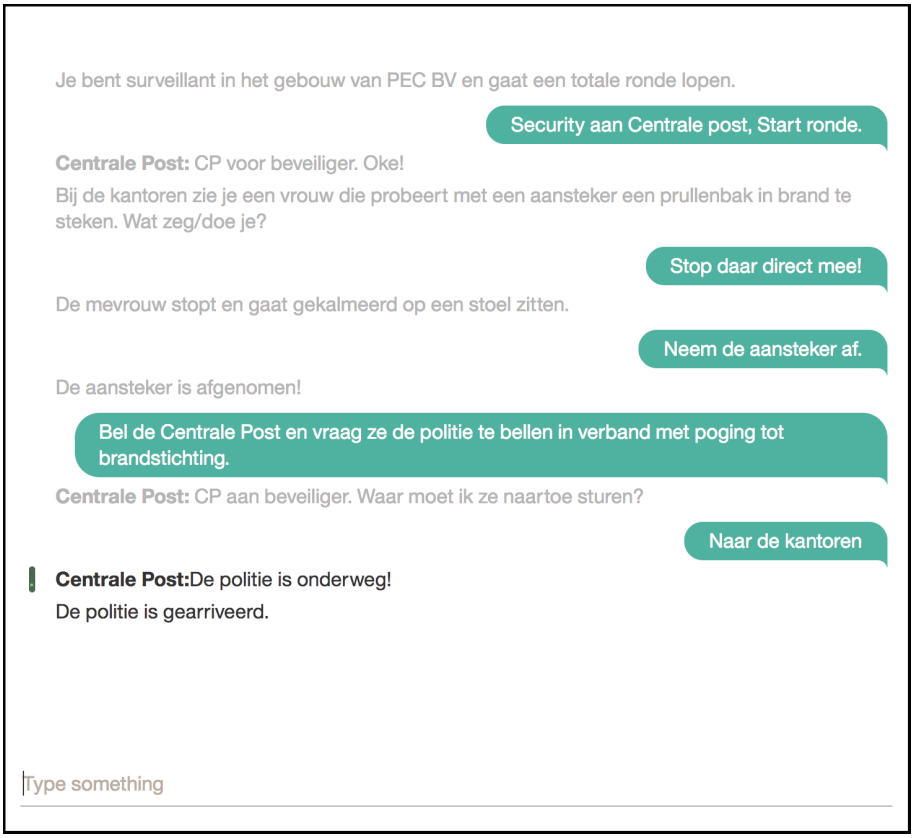

Fig. 2. Example interaction with the chatbot

person or object, one should always describe the primary and secondary features (e.g., 'the red BMW convertible with license plate AB-CD-12').

4.2 Experiment 1

As mentioned above, Experiment 1 (see Fig. 3) consisted of two sub-experiments, called Experiment 1A and 1B. The aim of Experiment 1A was to compare the differences in the learning effect between the two variants of the chatbot. The aim of Experiment 1B was explore the difference between learning with and without chatbots. In total 10 users participated in the experiment, among which 6 were male and 4 were female. The free-speech version was randomly assigned to 5 participants and the text-based version to the other 5 participants, of which one participant dropped out. First, the participants were provided with the learning material, which they could study by themselves before hand for a week. After that, they attended a lesson day (including a practice exam) given by the company. On that day, they were informed about what to expect during the exam

and what are the potential pitfalls. Also some use cases were practiced by role playing with a colleague. After the lesson day, all participants were asked to practice by interacting with (either the combined or the text-based version of) the chatbot every day for 30 min for a period of 5 to 7 days. After that, they took a practical exam with real actors. This practical exam took place at the exam center in Amersfoort. To pass the exam, the users was presented with three use cases in which (s)he was expected to act accordingly to the steps. A third person was assessing and grading the trainee based on a checklist.

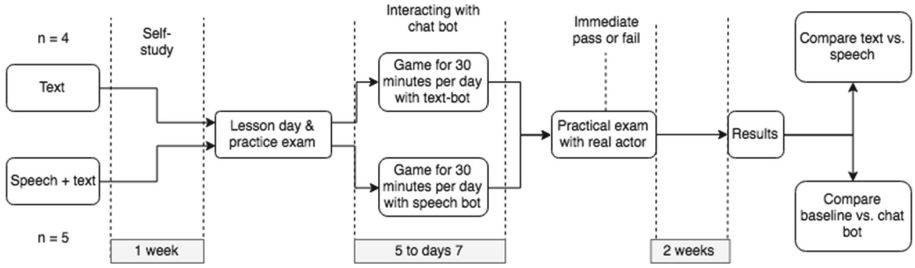

Fig. 3. Design of Experiment 1A

To test the hypothesis that learning with the combined chatbot leads is more effective than learning with the text-based chatbot, the average grades of the two groups were compared. Since the power of the analysis is very low ($n = 9$), no statistical test could be applied.

As a follow-up (Experiment 1B), we were interested in the question whether learning with (any of the two versions of) our chatbot is equally effective (or perhaps more effective) compared to the traditional form of learning. To this end, the grades of the 9 participants of Experiment 1B were compared to the grades of a baseline group of 29 candidates who took the exam after learning the material from a book instead of an interactive chatbot. Among these 29 candidates, 20 were male and 9 were female. The grades of these candidates were collected in the period between January 1st and May 31st, 2018. To test the hypothesis that the performance of both groups is the same, an unpaired t-test was applied. Hence, the independent variable of this analysis was the applied method of interactive learning and the dependent variable the average grade of the examinees.

4.3 Experiment 2

To obtain more qualitative results about people's opinion on the chatbot as a training intervention, 29 users were asked to fill in a questionnaire about their experience with the chatbot. This group consisted of the 9 participants from Experiment 1A and 20 additional participants who have passed the exam. These 20 employees were asked to interact with the combined version of the chatbot

for a week before filling in the questionnaire. Among these 20 employees, 13 were male and 7 were female. The questionnaire contained a number of questions using a 5-point Likert scale as well as some open questions asking for the participants' opinion about their understanding of the use cases and the added value or downside of using chatbots for social skills training. See Appendix C for more details.

Figure 4 summarizes the experiments. In Experiment 1, nine users were divided into two groups: five of them using the combined voice+text chatbot and four of them using the text-based chatbot. After comparing their grades with each other, their grades were also compared with the grades of a baseline group formed by 29 candidates that did not use the chatbot. For Experiment 2, another group of 20 employees that already have took the exam in the past was asked to use the voice+text-based chatbot and fill in a survey. The same survey was filled in by the group of 9 users from the first experiment, and the answers of both groups were analyzed.

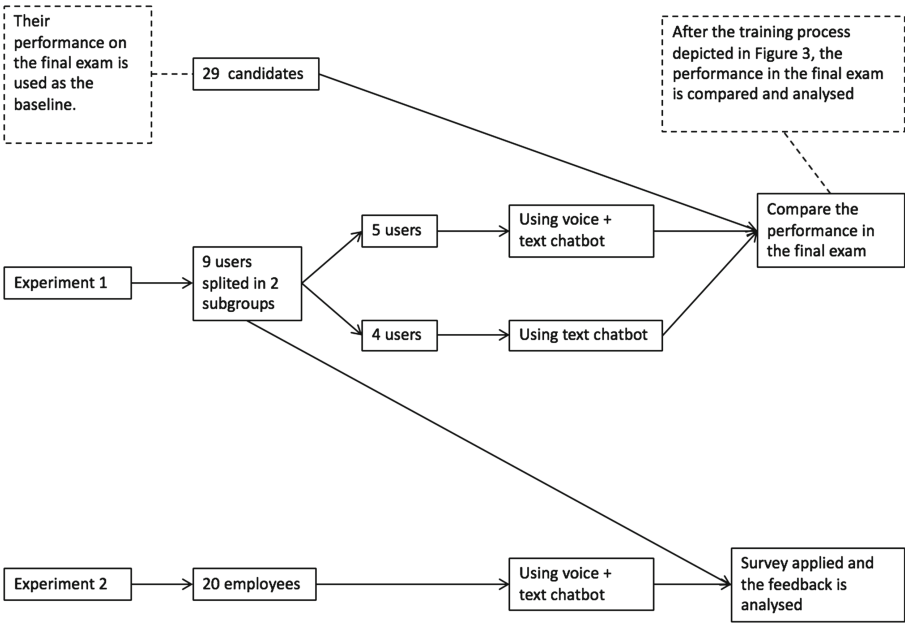

Fig. 4. Summary of the experiments

5 Results

The results of Experiment 1A are shown in Fig. 5. As can be visually observed in the figure, there is not much difference in the grades across the two groups: the users who used the text-based system (mean grade 5.4) performed very similar

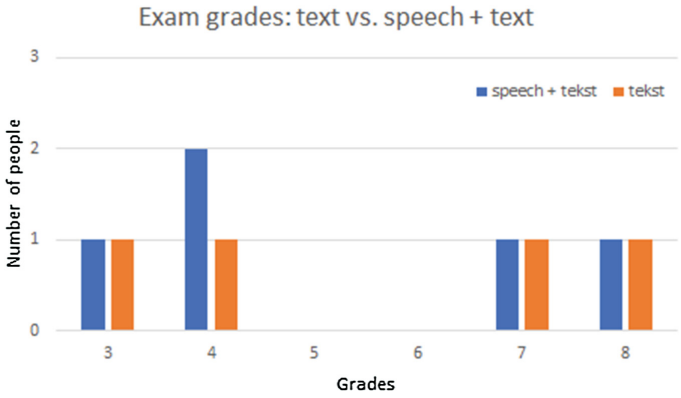

Fig. 5. Experiment 1A: exam grades of users who used the chatbot system.

as the users who used the combined system (mean grade 5.2). Due to the low power of this experiment, no statistical analysis was conducted.

For Experiment 1B, the grades of Experiment 1A were taken together, and were compared to the grades of a group who learned for the exam via traditional means (i.e., from a textbook). The results for this group are shown in Fig. 6. As can be seen, the results are similar to the results of the users who used the chatbot. Both populations have a low frequency in the grades 5 and 6, and a higher frequency in the grades 3, 4, 7 and 8.

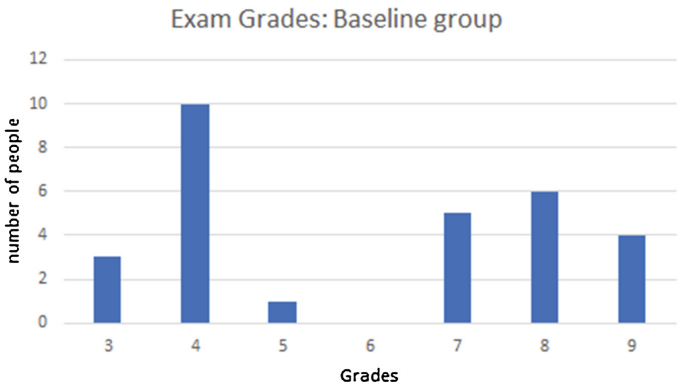

Fig. 6. Experiment 1B: exam grades of users who did not use the chatbot system.

To test if there is indeed no statistical difference in average score between the two populations, a two-sided t-test with unequal variances was applied. The significance level is set to $\alpha = 0.05$ with 14 degrees of freedom. This test pointed out that first group (mean score 5.33) did not obtain a statistically different grade from the second group (mean score 5.96) ($p = 0.45$).

In addition, the aim of Experiment 2 was to obtain some qualitative results about the users' opinion about the chatbot. The main results (for a selection of relevant questions) are reported in Table 1.

Table 1. Experiment 2: results of the survey for questions 6, 7, 8 and 11.

Question	Totally disagree	Disagree	Neutral	Agree	Totally agree
Q.6. There was enough variation in use cases	0%	5%	10%	80%	5%
Q.7. Training with the chat bot is a useful addition for studying for the practical exam	0%	0%	5%	65%	30%
Q.8. By training with the chat bot I perceived studying for the practical exam as less boring	0%	0%	5%	75%	20%
Q.11. I wouldn't mind to keep training with an extended version of the chatbot in the future	0%	5%	40%	35%	20%

Question 6 aimed to find out if the system has enough variation on the use cases as the system covers different use cases for an effective training process. The results in Table 1 are very positive signalling that 85% of the participants agrees with this statement. In other words, it indicates that participants thought that the selection of use cases covered most of the relevant scenarios. Still, there is room for improvement by including more use cases.

Question 7 aimed to find out if the participants had a positive attitude towards using the chatbot. Row 2 of Table 1 shows that 95% considered the chatbot a useful addition for the preparation of the practical exam. Only 5% remained neutral. This gives a positive indication that the chatbot is a useful addition to the regular course material. Question 8 aimed to find out if the participants of the survey had an increased motivation for studying for the practical exam. The results in Row 3 of Table 1 show that again 95% was positive on the statement: studying for the practical exam is less boring when having a chatbot. Similarly, only 5% remained neutral.

Finally, Question 11 was designed to find out whether it is interesting for the company to continue with the chatbot as an addition to the course material. Row 4 of Table 1 shows that 35% agrees with the statement and 20% totally agrees. When asked about which use cases were more beneficial (Question 4), the

answers clearly indicated the most complex use cases, with a long dialogue. They were considered more challenging, realistic and covering several aspects needed in the training. That indicates the advantage of using interactive training that move the users to dynamic interactions. This fact is confirmed by the answers to Question 5 which asked the opposite of Question 4: it inquired about the less useful use cases, i.e. the scenarios that helped people less in the learning process. The answer, again was clear and indicated the shortest and less complex use cases as bringing a small contribution to the training process.

6 Discussion

The primary results of our study indicate that users who prepared for the exam with the chatbot system did not obtain different grades than who used a traditional text book method. This finding can be interpreted both positively and negatively: the bad news is that the chatbot does not lead to better scores, but the good news is that it does not lead to worse scores either. In addition, the results of the survey seemed to indicated that people were positive about the added value of the chatbot system. Therefore, it can still be considered as an interesting alternative to textbooks, which stimulates people to spend more time on learning.

Indeed, other studies have already hinted at the potentially positive impact of using chatbots [1,5,9] or virtual agents [3] on engagement and learning effect. These papers point out similar advantages as were observed in our survey, and during after-session discussions with our participants. Most of the users mentioned as advantages of using a chatbot that it is less repetitive and more engaging than simply reading the use cases. As such, the chatbot system presents an interactive environment which enables users to 'learn by doing'.

Nevertheless, a number of limitations of our study could be mentioned. Firstly, there is no way to assure if the users really practised 30 min per day for a minimum of 5 days. If this is not the case, the results would be biased. Secondly, with a sample size of 9 in the experimental group and 29 in the baseline group the experiment was a bit underpowered for adequate statistical analysis. With a larger sample size and more strict instructions and control the statistical significance would be more meaningful. According to Schreiber et al. each group should have at least 10 participants for the results to draw any conclusions about significance [8]. Finally, a third possible weakness in the experiment was that the document with use cases provided by the company might be outdated. The document was drawn up in 2014 and has not been updated since. Many of the use cases have been replaced because of change in protocols or legislation. This was identified during the training and could affect not only the performance with the chatbot but also the traditional approach, because the same material was given to the participants of the experiment to prepare themselves for the final exam. Moreover, it is important to note that not only the system must behave according to the expectations, but it also has to be attractive and engaging, otherwise the users lose interest. When this happens, a chatbot might lose its advantage

over traditional tools. This phenomenon could clearly be observed in the results of the survey. Some users mentioned that simple use cases were not challenging and some others lost interest after practicing the same use case for some time. The time invested in the practice period with the chatbot also contributes to the learning effect. For this particular aspect, more time and more variability of use cases will help to improve the final results.

To conclude, our study provides also some hints that the system is promising to enhance user experience. Although the chatbot did not have a significant impact on the average pass rate, it captivates the user interest to spend more time studying, as show the survey results. Furthermore, it prepares users to deal with the social and time pressures of the final exam. To improve the system more use cases could be included and minor characteristics should be refined in future work. One example is to extend the library to recognise other constructions of sentences that have similar meaning, to ensure that different conversation styles will be considered by the chatbot as a correct answer. Moreover, further research is needed to better understand the differences between text-based and speech approaches. Although the current design does not allow us to draw any conclusions about this, a more sophisticated experiment (e.g., using different performance indicators) may shed more light on this topic.

Acknowledgement. This research was supported by the Brazilian scholarship program Science without Borders - CNPq scholarship reference: 233883/2014-2.

Appendix

A Online System

The two versions of our system are available online via the following webpages:

- text: http://vesci.labs.vu.nl/steinabacus/
- speech and text: http://vesci.labs.vu.nl/steinvox/

Both webpages contain an introduction with instructional video, followed by links to the ten use cases.

B Example of a Plain Text Use Case

Case: Arson in the fusebox.
Instruction: You are surveillance at PEC BV. and you are walking a full round.

1. Tell the central post that you are starting your full round.
2. The door of a technical room is open and you hear screaming.
3. When you go and look, you encounter a woman that is trying to light a fire in a dust bin.
3. Tell the woman to stop immediately and arrest her on suspicion of Arson.
4. Confiscate the lighter.

5. Ask the Central post if they can call the police.
6. Hand over the woman to the police that has been arrived.
7. Hand over the lighter as evidence.
8. Ask the names of the agents and to which bureau they are taking the suspect.
9. Speak about 'the suspect' and not 'the woman'.
10. Draw up specific report.
11. Tell the central post that everything has been resolved and that you are continuing with your round.

C Survey Questions

1. Which version of the chatbot have you been using?
   - Speech + text input
   - text input
2. How long did you practice with the chatbot each day?
   - under 10 min
   - 10 to 20 min
   - 20 to 30 min
   - over 30 min
3. For how many days did you use the chatbot?
   - under 5 days
   - 5 days
   - 6 days
   - 7 days
   - over 7 days
4. Which use case did you find most useful and provided the best training? (multiple answers possible)
   - 1. Arson in the fuse box
   - 2. Broken elevator
   - ...
   - 10. Mover
5. Which use case did you find least useful and provided the least best training? (multiple answers possible)
   - 1. Arson in the fuse box
   - 2. Broken elevator
   - ...
   - 10. Mover
6. There was enough variation in use cases.
   (a) Totally disagree
   (b) Disagree
   (c) Neutral
   (d) Agree
   (e) Totally agree
7. I believe practicing with the chatbot is a useful addition to the regular course material.
   (a) Totally disagree

(b) Disagree
(c) Neutral
(d) Agree
(e) Totally agree

8. By training with the chatbot I perceived studying for the practical exam as less boring.
   (a) Totally disagree
   (b) Disagree
   (c) Neutral
   (d) Agree
   (e) Totally agree
9. What could have been done better regarding the design of the chatbot?
10. What difficulties did you encounter practicing with the chatbot?
11. In the future I would not mind continuing to train with an extended version of the chatbot.
    (a) Totally disagree
    (b) Disagree
    (c) Neutral
    (d) Agree
    (e) Totally agree
12. What is your name?

References

1. Abbasi, S., Kazi, H.: Measuring effectiveness of learning chatbot systems on student's learning outcome and memory retention. Asian J. Appl. Sci. Eng. **3**(2), 251–260 (2014)
2. Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., Sereika, S.M.: Maintaining reality: relational agents for antipsychotic medication adherence. Interact. Comput. **22**(4), 276–288 (2010)
3. Bosse, T., Gerritsen, C., de Man, J.: Evaluation of a virtual training environment for aggression de-escalation. In: Proceedings of Game-On, pp. 48–58 (2015)
4. Bosse, T., Provoost, S.: Towards aggression de-escalation training with virtual agents: a computational model. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2014. LNCS, vol. 8524, pp. 375–387. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07485-6_37
5. Hoffmann, R., Kowalski, S., Jain, R., Mumtaz, M.: E_universities services in the new social eco-systems: Security risk analysis: Using conversational agents to help teach information security risk analysis (2011)
6. Medeiros, L., Bosse, T.: Testing the acceptability of social support agents in online communities. In: Nguyen, N.T., Papadopoulos, G.A., Jędrzejowicz, P., Trawiński, B., Vossen, G. (eds.) ICCCI 2017. LNCS (LNAI), vol. 10448, pp. 125–136. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67074-4_13
7. Oh, K.J., Lee, D., Ko, B., Choi, H.J.: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In: 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 371–375. IEEE (2017)

8. Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A., King, J.: Reporting structural equation modeling and confirmatory factor analysis results: a review. J. Educ. Res. **99**(6), 323–338 (2006)
9. Shawar, B.A.A., Atwell, E.: A corpus based approach to generalising a chatbot system. Ph.D. thesis, University of Leeds (School of Computing) (2005)