User Experience of a Conversational User Interface in a Museum

Stefan Schaffer[1](✉), Aaron Ruß[1], and Oliver Gustke[2]

[1] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
stefan.schaffer@dfki.de
[2] Linon Medien, Berlin, Germany
chim@linon.de

Abstract. In this paper, we summarize the initial results of a field test with the ChiM (Chatbot in the Museum) system, a conversational user interface for the museum. The system contains a Natural Language Understanding (NLU) component that translates the user input into intentions and produces a multimodal (mainly spoken and textual) output. Museum visitors can use the system to freely ask questions about the exhibits in the exhibition. We conducted a field test with 140 participants in the Städel Museum, Frankfurt, and recorded over 4600 interactions between the participants and the system. After the test, participants gave their perceived feedback on the user experience (UX) and completed a custom system-specific questionnaire. We exploratively analyzed the feedback. The results show an overall medium UX for the system. We assume that the NLU component must be improved. Participants who rarely or never use audio guides rate the pragmatic quality (PQ) of the system significantly better than people who often or always use audio guides. People who rated the speech quality of the system as good also rated the attractiveness of the system significantly better than people who rated the speech quality as bad. In our future work, we will deepen the UX analysis and further put focus on recorded interaction data.

Keywords: Conversational Interaction · User Experience · Field Study

1 Introduction

Museum visitors often have specific questions about objects and topics. But a personal museum guide is not always available. They often experience that "the guided tour will not start for another hour", or even "the guided tour is canceled today". Today's individual guided tour systems (e.g., media or audio guides), on the other hand, cannot answer specific questions. They can often only be used interactively to a limited extent and tend to offer "one-way communication".

ChiM (Chatbot in the Museum) is intended to fill this gap in the museums' educational offer and to be able to react meaningfully to the user's input: A digital conversational interaction system as an expert and companion that can be taken along, answering

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. L. Brooks (Ed.): ArtsIT 2022, LNICST 479, pp. 215–223, 2023.
https://doi.org/10.1007/978-3-031-28993-4_16

questions, and providing further information. Our vision is an individualized museum experience with a personal chatbot guide.

The long-term goal of the ChiM research project is to develop an application that enables museums and exhibitions to provide all visitors with a personal virtual guide who can answer individual questions just like a modern human guide.

The optimum museum tour is one person, let's say the museum director or a competent guide, who leads each guest through the exhibition individually and personalized according to their interests. Of course, a personal, human guide cannot be available for all individual visitors. A future virtual conversation system, however, can. The visitors "meet their personal museum guide", a freely configurable avatar within an application, at the beginning of their visit to the museum. The guide picks up the visitors (controlled by location-based services) where they are.

The aim of our work is to develop a system for conversational interaction in the museum environment - or "ChiM" for short, the chatbot in the museum. ChiM can meet the future requirements of museum visitors for knowledge transfer and eliminate existing "pain points".

ChiM is therefore an interactive, personalized conversation system combined with an appealing interface that can access various content databases to obtain new information and at the same time is designed to "learn" and, for example, to remember the preferences of its users.

In this paper, we present the initial results of a large field test conducted with the ChiM system. The analysis in this paper focuses on aspects of the user experience (UX). We collected UX, demographic, and system questionnaire data and exploratively examined them for significant differences.

In Sect. 2, we describe the main functionalities of the ChiM system. Section 3 describes the field test and Sect. 4 the results obtained so far. Section 5 concludes and points to future work.

2 About ChiM

2.1 Background

More and more museums are developing chatbots to assist their visitors and to provide an enhanced visiting experience. Most of these chatbots are developed using a chatbot platform, that provides predefined dialogs [1]. In the museum chatbot Ping!, the dialog evolves based on users' decisions between predefined input options [2]. However, such predefined dialogs do not provide a human-like conversation and do not provide answers to visitors' real questions. In recent times, more and more chatbots are appearing that try to employ the latest NLP techniques for enabling museum-related dialogs and question answering [1]. "The Voice of Art", e.g., is an artificial intelligence voice-based interactive guide which allows visitors to ask questions to artworks in the Pinacoteca Museum in Brazil [3]. In the future, such approaches will use advanced machine learning techniques based on deep learning, such as large language models like BERT [4, 5].

The basic Natural Language Processing (NLP) mechanism employed in ChiM follows a multitiered approach using techniques like Rasa, BERT, and cosine-similarity to

generate answers with different degrees of effort [6, 7]. The NLP procedure is described in our paper about "important steps to let AI chatbots answer questions in the museum" [8]. Adopting an approach from [3], we identified distinct content types for questions asked about selected artworks from the Städel Museum and developed Natural Language Processing (NLP) strategies for generating answers by using these content types, complemented by additional annotations.

2.2 The ChiM System

The ChiM chatbot app was developed for the target platform Android based on the open-source cross-platform framework Apache Cordova. The open-source framework MMIR [9] was used for voice input and output, and the open-source frameworks Angular and Ionic were used to develop the graphical user interface.

In the ChiM app, users can select exhibits and ask freely formulated questions by voice or text input; the answers can be rated and commented on. For the selected exhibit, the most frequently asked questions about the exhibit, the location map (where the exhibit can be found in the museum), and other media content such as audio files (with audio guide soundtracks) can also be selected via an "object menu" (see Fig. 1).

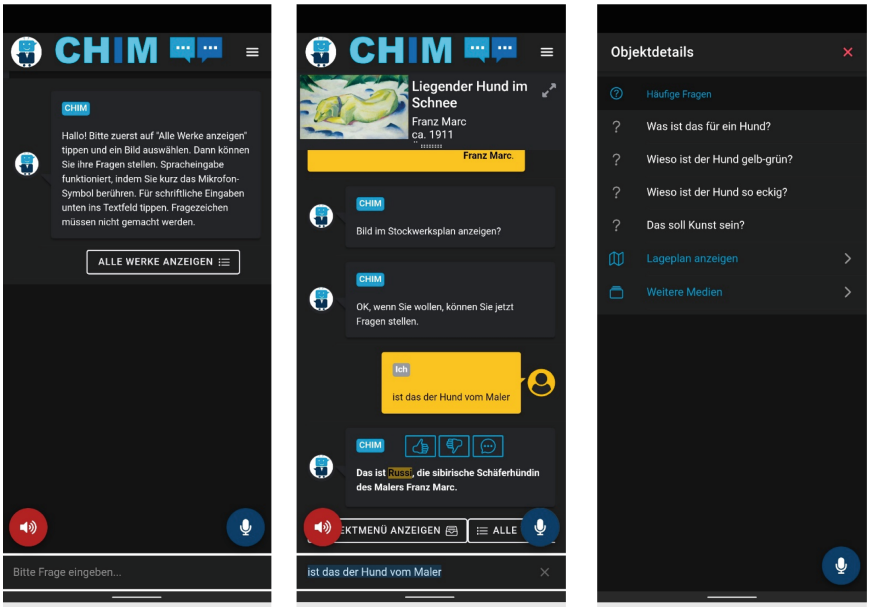

Fig. 1. Screenshots of the Chatbot App. Left: Chatbot with greeting text (active voice output); Middle: Chatbot with an answer to question "ist das der Hund vom Maler" ("is that the dog of the painter" with the answer "That is Russi the Siberian shepherd dog of the painter Franz Marc"); Right: Display of the exhibit details with the frequently asked questions (FAQs).

In addition to voice input and output for questions and answers about the selected exhibits, context-dependent (i.e., depending on the current state of the app, e.g., the displayed content) voice commands can be used to control the app, for example, to display the selection dialogue for exhibits, to select an exhibit from it, to give a rating for an answer, etc. The voice commands that are available in the current context can be listed via the menu or the command "What can I say" and searched with a filter function. Technically, the voice commands are compared with the labels (and alternative labels) of the currently active actions (i.e. defined functional units such as rating-good, show-all-exhibits), and depending on syntactic comparison procedures (e.g. partial string match, tri-gram comparison) and dynamic thresholds (e.g. complete/partial match, XY % match, match of at least XY tri-grams), it is determined whether the user utterance is a valid voice command (and executed if necessary).

3 Field Test

3.1 Preparation

A concept for the implementation of the field test was developed regarding duration, procedure, and methodology used (data collection in line with data protection, standard questionnaires, etc.). A custom questionnaire specially tailored to the project and the standard AttrakDiff questionnaire [10] (on user experience) were implemented in the chatbot app, including a data link to save the collected, anonymized data in a database. In addition, the chatbot app anonymously logs interactions (e.g., selection of a work/image, questions about the work, voice commands) into a database.

AttrakDiff measures the attractiveness of the system on 4 scales. It consists of 28 seven-step items whose poles are opposing adjective pairs (e.g., "confusing – clear", "unusual – ordinary", "good – bad"). Each set of adjective items is ordered into a scale of intensity. Each of the averaged values of an item group creates a scale value for pragmatic quality, hedonic quality, subdivided into users' identity with the system and users' stimulation by the system, and attractiveness.

- Pragmatic Quality (PQ): The ability of a product to satisfy the need for goal attainment by providing useful and usable features. Typical product attributes are: practical, predictable, clear, manageable.
- Hedonic Quality - Stimulation (HQS): the ability of a product to satisfy the need to improve one's knowledge and skills. Typical product attributes are: engaging, creative, original, challenging.
- Hedonic Quality - Identity (HQS): the ability of a product to communicate messages of self-worth to relevant others. Typical product attributes are: brings me closer to people, expert, connecting, stylish.
- Attractiveness (ATT): Global positive-negative evaluation of the product: good, attractive, pleasant.

Shortly before the field test was carried out, the location of 13 exhibits selected for the field test was recorded and site plans were created for each exhibit and integrated into the app and the data backend.

For the field test, 13 smartphones (3× Google Pixel 4a, 5× Google Pixel 6, 5× Asus Zenfone 8) were set up with Android 12 pre-configured with the chatbot app, and a kiosk mode app was installed.

3.2 Execution

The field test was carried out at the Städel Museum from April 26th, 2022, to May 1st, 2022. Two test assistants carried out the field test, who instructed the participants, handed out the devices, and organized the procedure in the museum. The test assistants ensured that the same test schedule was adhered to for all participants. Before starting the visit the following steps were carried out: 1. Welcome; 2. Handing out the task description; 3. Pseudonym selection by the participants; 4. Explanations for "What is CHIM text" and data protection settings; 5. Assistance in trying out the system; 6. Answering any questions that the participants still had. After the visit: 1. if necessary, help with filling in the questionnaires; 2. Goodbye; 3. Reset the system for the next participants.

3.3 Participant Task

The participant's task was to ask at least one question about at least 6 different objects. If the participants wanted more, they were allowed to ask more questions about as many of the selected objects as they wanted. The participants were allowed to move freely through the museum and had to find the ChiM objects on their own. There was no special labeling of the objects in the exhibition.

4 Results and Discussion

A total of 140 participants took part in the test period, of whom 95 provided sufficient data for the following analysis (e.g. completed tasks and AttrakDiff questionnaire). Of these participants, 57 were female, 28 were male, and 10 did not specify female or male; 70 participants indicated their age with an average of 34.3 years (median 30 years).

In total, over 4600 interactions (user inputs into the system) were carried out, with 3722 for participants considered in the following analysis, from which 2331 (or 2909 for all) were questions about the exhibits; on average, participants in the following analysis asked 24.5 questions during their visit.

The overall UX rating via the AttrakDiff questionnaire is illustrated in Fig. 2. The participants gave medium ratings on all AttrakDiff scales. Despite the comparatively high number of participants in the survey, the error bars indicate a high standard distribution for all scales. This indicates that there are inter-individual differences in the participants' assessments. Therefore, a further exploratory investigation of the AttrakDiff ratings was conducted with the system questionnaire data. The main reason why the mean values are only in the middle range is probably the not perfectly functioning NLU component of the system. So far, only qualitative feedback from the participants and the observations of the experimenters are available on the quality of the NLU.

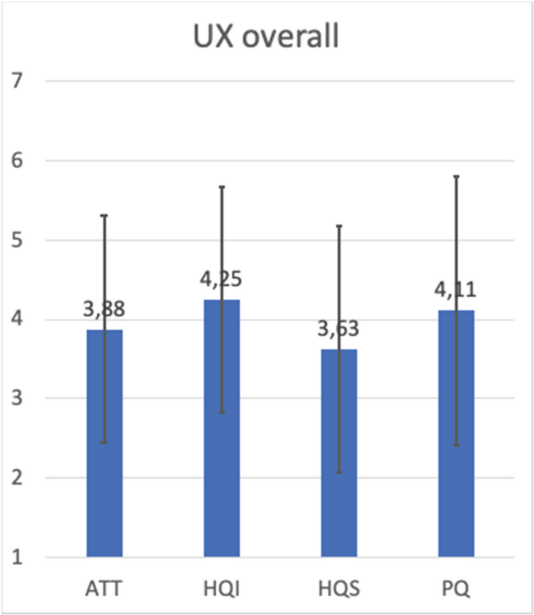

Fig. 2. Overall User Experience (UX) ratings.

A notable correlation could be found with "audio guide usage". Figure 3 illustrates the UX ratings in relation to the usual audio guide usage by the participants. There are no significant differences in the attractiveness (ATT), hedonic quality-identity (HQI), and hedonic quality-stimulation (HQS) scales. For pragmatic quality (PQ) it turned out that participants who rarely or never use audio guides rate the pragmatic quality significantly better than people who often or always use audio guides ($F(2, 85) = 3.40$, $p < .05$). A posthoc analysis revealed significant differences between the groups rare and medium and rare and often. This suggests that the chatbot could be a useful addition for people who would not choose an audio guide system.

Another interpretation could be that people with rare experience with audio guides are just unfamiliar with digital or multimedia information systems in museums and thus rate it as more useful as "experienced" users; as to answering the question "if this could hinder or promote the usage of chatbot-systems as an addition to classical audio guides", further data and analysis would be required.

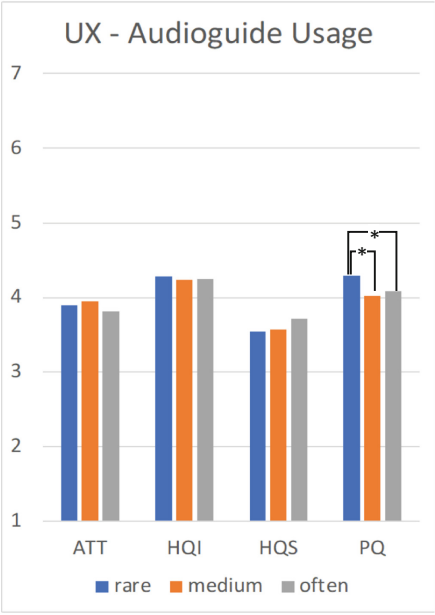

Fig. 3. User Experience (UX) ratings by audio guide use. "*" indicates significant differences.

Another substantial result was found when analyzing the rating for "quality of the voice output" (see Fig. 4). People who rated the speech quality of the system as good also rated the attractiveness of the system significantly better than people who rated the speech quality as bad. This indicates that the quality of the voice output is a decisive factor in evaluating the attractiveness of the system. A poor output voice quality seems to be annoying for many participants.

The custom questionnaire on the system also included the possibility to give qualitative feedback. There was positive and negative feedback. Here are some examples of feedback that were mentioned more frequently.

Positive

1. "You can ask any question that comes to mind. You can also dare to ask simpler questions".
2. "…if the system doesn't give a correct answer to my question, there may still be interesting information about the painting or the artist".

The positive comments indicate that the form of conversational interaction in principle attracts some participants, even in cases where the (first) response is not a direct answer.

Negative

1. "The bot would need to be filled with more data to better answer my questions".

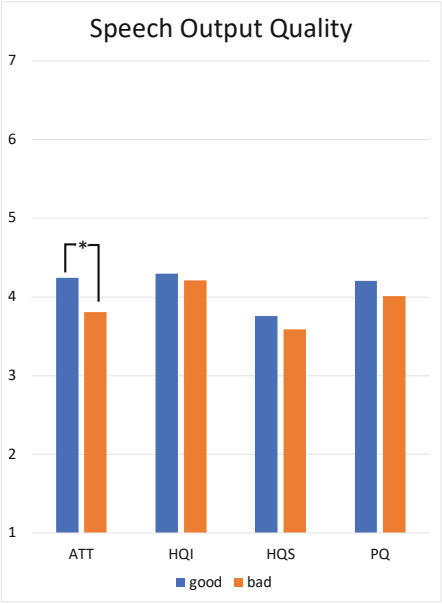

Fig. 4. User Experience (UX) Ratings by speech quality of audio output. "*" indicates a significant difference.

2. "I would have liked to see an introduction so that the questions were not asked at the beginning, but as continuing questions".

Negative comment 1 indicates that more data is needed to give satisfactory answers. Negative comment 2 shows that some users want an adapted interaction strategy of the system.

5 Conclusion and Future Work

In this paper, we have summarized the initial results of a field test with the ChiM system, a conversational user interface for the museum. Participant feedback on user experience and a system questionnaire was analyzed. The results show an overall medium UX for the system. Qualitative feedback from the participants as well as observations from the experimenters indicates that one main reason for the medium rating could be the quality of the NLU component of the system, which could be improved. The exact calculation of NLU quality will be part of our future work.

Participants rated the pragmatic quality (PQ) of the system significantly better if they rarely or never used classical audio guides than participants who often or always used audio guides. This suggests that a chatbot could be a useful addition for museum visitors who usually would not choose an audio guide system. This means that conversational interaction could be a useful additional channel for knowledge communication in museums in the future.

A further investigation regarding the quality of the voice output revealed that people who rated the speech quality of the system as good also rated the attractiveness (ATT) of the system significantly better than people who rated the speech quality as bad. This suggests that good voice output quality might be an important factor in evaluating the overall attractiveness of the system.

In this paper, we could only present initial results from the huge amount of data collected during the field study with ChiM. In our future work, we will focus on analyzing interaction data, which could not be analyzed so far because of resource and time constraints. Extensive work is currently underway to annotate the interaction data, which will form the basis for further publications.

Acknowledgment. This research is funded by the German Federal Ministry of Economics and Climate Protection (BMWK) project ToHyVe.

References

1. Varitimiadis, S., Kotis, K., Pittou, D., Konstantakis, G.: Graph-based conversational AI: towards a distributed and collaborative multi-chatbot approach for museums. Appl. Sci. **11**(19), 9160 (2021)
2. Ping! Die Museumsapp. https://www.landesmuseum.de/digital/apps-und-podcasts/ping. Accessed 25 Oct 2022
3. Barth, F., Candello, H., Cavalin, P., Pinhanez, C.: Intentions, meanings, and whys: designing content for voice-based conversational museum guides. In: Proceedings of the 2nd Conference on Conversational User Interfaces, pp. 1–8 (2020)
4. Gaia, G., Boiano, S., Borda, A.: Engaging museum visitors with AI: the case of chatbots. In: Giannini, T., Bowen, J.P. (eds.) Museums and digital culture. SSCC, pp. 309–329. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-97457-6_15
5. Koroteev, M.V.: BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943 (2021)
6. Zaman, M.M.U., Schaffer, S., Scheffler, T.: Factoid and open-ended question answering with BERT in the museum domain. In: Proceedings of the Conference on Digital Curation Technologies. Conference on Digital Curation Technologies (QURATOR-2021). CEUR Workshop Proceedings (2021)
7. Zaman, M.M.U., Schaffer, S., Scheffler, T.: Comparing BERT with an intent based question answering setup for open-ended questions in the museum domain. In: 32. Konferenz Elektronische Sprachsignalverarbeitung. Elektronische Sprachsignalverarbeitung. Elektronische Sprachsignalverarbeitung (ESSV-2021). TUDpress, Dresden (2021)
8. Schaffer, S., Ruß, A., Sasse, M.L., Schubotz, L., Gustke, O.: Questions and answers: important steps to let AI chatbots answer questions in the museum. In: Wölfel, M., Bernhardt, J., Thiel, S. (eds.) ArtsIT 2021. LNICST, vol. 422, pp. 346–358. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-95531-1_24
9. Ruß, A.: MMIR framework: multimodal mobile interaction and rendering. In: GI-Jahrestagung, pp. 2702–2713 (2013)
10. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In: In: Szwillus, G., Ziegler, J. (eds.) Mensch & Computer 2003. Berichte des German Chapter of the ACM, vol. 57, pp. 187–196. Vieweg+ Teubner Verlag, Berlin (2003). https://doi.org/10.1007/978-3-322-80058-9_19