Towards Chatbots in the museum

Stefan Schaffer[1], Oliver Gustke[2], Julia Oldemeier[2], and Norbert Reithinger[1]

[1] German Research Center for Artificial Intelligence (DFKI),
Intelligent User Interfaces, Alt-Moabit 91c, 10559 Berlin, Germany
fistname.lastname@dfki.de
[2] Linon Medien,
Steigerwaldblick 29, 97453 Schonungen, Germany
{og,jo}@linon.de

Abstract. In this short paper we report on work in progress in the research project "Chatbot in the museum" (ChiM). ChiM develops a practicable technical solution for the use of chatbots in the museum environment. The paper outlines conceptual work conducted so far, including the comprehension of three important research topics explored in ChiM, namely information processing, multimodal intent detection and dialog management for museum chatbots.

Keywords: Chatbot museum guide · Personal museum assistant · Interactive user interfaces.

1 Introduction

A chatbot is a computer program that attempts to simulate the conversation of a human being via text or voice interactions [4]. In the museum context chatbots offer great potential, as existing "pain points" can be eliminated: in contrast to personal tours ("takes place in an hour", "is cancelled today"), chatbots are always available. Today's digital visitor guidance systems offer only "one-way communication" and are not able to respond to questions from the visitor. Chatbots have the potential to respond meaningfully to the user's input if the input is processed properly.

Conversations with museum visitors showed that they often have specific questions about certain objects. Classical audioguides can not answer specific questions. In the museum a chatbot could be the expert you can take with you, answer questions and provide further information. Fig. 1 exemplifies how the conversational interaction between a visitor and the chatbot could look like.

The aim of ChiM is to explore the usability of a chatbot as an interactive system for knowledge and learning, as well as for effective access and the comprehensible presentation of museum information. A central question is therefore how the existing information must be structured in order to relate to the visitors' questions. ChiM develops new solutions for providing information and explores how the latest research results in the field of intention detection and dialog management can be utilized for chatbots in the museum context. Specifically, the following research fields of human-technology interaction are to be investigated:

Fig. 1. Example chatbot interaction.

- *Information processing*: adaptation of the existing content creation process for classical audio and media guides, to be more "knowledge-based" for the chatbot.
- *Multimodal intention detection*: linguistic or text-based input processing combined with image and other exhibition sensor (e.g. beacons) processing.
- *Dialog management for guided tours*: intelligent dialog strategies to determine context-relevant information and personalized information.

In this paper conceptual work in the ChiM research fields information processing, multimodal intent detection and dialog management for museum chatbots is outlined. In the next section related work in chatbot relevant fields is sketched. Section 3 describes the so far developed concepts for ChiM. Section 4 describes ongoing work on use case development. Conclusion and future work are described in Section 5.

2 Related Work

Current research in the field of interactive museum guides covers a wide range of approaches, from beacons controlling the presentation [3], over agent-based techniques [6], to robotic museum guides [11]. In the museum field, still no elaborate technologies can be found that essentially utilize conversational digital systems such as chatbots. Providers such as helloguide[3] so far only carry over the paradigm of entering numbers from a classical audio guide to chatbots. It is not possible to engage in dialog or ask questions with such museum chatbots. Chatbot/dialog platforms such as Alexa (Amazon), Dialogflow (Google), Wit.ai (Facebook) or Watson (IBM) enable intention detection for many domains (eg.

[3] http://helloguide.de

for ordering a pizza, weather report, flight booking, shopping, etc.) [2]. However, most of these platforms only offer limited customization to their own domains, and are limited in the choice of input and output modalities. In addition, they are not able to manage a dialog with extensive knowledge bases. Therefore, for special domains such as a tour through an exhibition, own solutions for intent detection and dialog management have to be implemented. In many conversational systems domain-specific knowledge is mapped by dialog grammars and state machines [5]. Dialog grammars and state machines provide a reliable method for intent detection and can guide the user targeted-oriented through the dialog, but are limited in flexibility compared to the diversity of natural language [13]. As a result, the intuitive operation of such systems, and thus user acceptance may suffer [9]. Recent successes of statistical methods for natural language processing (NLP) and dialog management open up new possibilities [1]. Statistical approaches enable to create more flexible models for intent detection and dialog management based on existing training data. However, a common problem is that there is little or no data for training. Hybrid methods that combine domain-specific knowledge with statistical approaches, as an alternative for low-data domains, are subject of current research [14].

3 The ChiM approach

ChiM investigates hybrid methods for intention detection and dialog management for the museum sector. A newly developed chatbot-based museum guide usually represents a new knowledge domain for which training data is not yet available. During the process of creating audio and multimedia guides, many data are generated, which can also be used as training data if appropriate transcription and indexing methods are applied. ChiM's approach is to combine the development of a hybrid approach for intention detection and dialog management with the creation process of museum tours, using the data generated by authors and editors as training material for a statistical model. Domain-specific knowledge components, which can not or can hardly be statistically mapped, will be realized through dialog grammars. In addition, the consideration of further information channels (e.g. image recognition, exhibition sensor technology) enables multimodal intent detection. Subsections 3.1-3.3 summarize the so far developed concepts for the research areas explored in ChiM. Subsection 3.4 explains the iterative realization of ChiM.

3.1 Information Processing

In contrast to classical audio and multimedia guides, information for a chatbot system must be structured differently. The entire existing text creation process must be converted into a more "knowledge-based" appraoch. ChiM needs (semi) structured data to allow for a flexible interaction with the content. For data preparation, approaches such as taxonomies and the use of digital asset man-

agement architectures such as Fedora Commons[4] will be integrated into editors' workflows. Further, it will be investigated whether and how the existing data for museum guides and their contents can be prepared in such a way that they can serve as training material.

3.2 Multimodal Intention Detection

Detecting user intent in the museum environment is a complex, multimodal process. Visitors can interact with text or voice input. Thus, on the one hand, the linguistic or text-based input of the user must be processed. On the basis of historical data and the results of the information processing, statistical and rule-based procedures will be evaluated and used. On the other hand, the consideration of further information channels, such as image processing and existing exhibition sensors (e,g. beacons for localization), is of particular interest in a museum context. For this, the MMIR (Multimodal Mobile Interaction and Rendering) framework will be used and extend [12]. The framework supports the creation of multi platform applications and enables straightforward integration of existing libraries, like e.g. OpenCV for image recognition [10], and existing location technologies [7] to explore the multimodal approach. By merging the information channels, the multimodal intent detection transfers the existing information into structured data that can be further processed to enable information retrieval (the determination of the relevant data) and the provision of information (the preparation and comprehensible presentation of the information). This fusion of the individual information channels results in the multimodal intention that represents the input of the dialog management [8].

3.3 Dialog management for museum tours

The individual steps from intention detection over the retrieval of the information to the provision of information are continuously carried out in a dialog between the user and the chatbot. As the core of the chatbot, an intelligent dialog management will process the multimodal intention and determine the system reaction to offer personalized information. The dialog should always be effective, intuitive and continued with the best possible user experience. The input/output modalities will be adapted to the situation using text, speech or multimedia elements.

More specifically, the information presented to the users is decided from the intention, the dialog history, the knowledge model from the information processing and the general context. The processing of the dialog management will be hybrid: from historical data typical museum guide sequences are learned. At the same time, dialog rules will be created and the two approaches combined [14].

[4] http://fedora-commons.org/

3.4 Iterative Realization

The realization takes place in two iterations using a user-centered and participative design approach. In the first iteration, the editors will prepare the knowledge for an exhibition. At the same time the basic ChiM functionality will be investigated with focus groups in an iterative UX process and tested in the lab with potential users. In the second iteration, the ChiM process will be implemented within further exhibition guides and evaluated with users in the context of field studies.

4 Use Case Development

As part of our ongoing work we elaborate on the proposed input and output modalities of the system and exemplify the goal of our approach with respect to the user by outlining first use case ideas which are related to multimodal interaction. As a precondition for all use cases it can be assumed that the chatbot is installed on a smartphone which has the necessary technical specifications, i.e. microphone, camera, and iBeacon, or respectively bluetooth low energy.

As main input modalities touch input on a virtual keyboard and alternatively speech input enabled by automatic speech recognition (ASR) will be utilized to interact with the chatbot. A smartphone camera will be used for computer vision in order to recognize exhibits or other objects relevant for a specific exhibition. Further the proximity to exhibits will be detected by beacons and processed analog to the context.

The output modalities will include visual, auditory and tactile feedback. The visual information comprises mainly the dialog with the chatbot and media content like images and videos. The chatbot will make use of speech synthesis to create auditory feedback for the system promts within the user-chatbot dialog. System promts will be generated for meta communication, i.e. system configuration, as well as for specific information coming from the knowledge model specific to the exhibition. Further auditory content consists of recorded audio material (also in combination with image or video) as in classical multimedia guides. Tactile feedback can be used for alerts, e.g. if the user gets closer to the next exhibit of a tour.

With regard to the context in which this interaction will take place a number of factors have to be considered. At the place itself, i.e. a museum or an exhibition, the acoustic characteristics can be very different, as the exhibition areas can be located in both large halls and smaller rooms. Further, people in the surrounding area can on the one hand produce noise which is harmful for ASR, on the other hand talking to the chatbot could distract other visitors. Within an exhibition the auditory signal should therefore be received by the user via headphones. It can further be necessary to avoid the usage of speech input if other visitors are located right next to someone. The system should therefore always provide touch input on a virtual keyboard as an alternative input mode. To determine relevant information not only the actual user input but also the last user

inputs, questions, or interactions should be considered to enable a personalized request.

Based on these assumptions we so far generated the following ideas for multimodal use cases:

- *Sequential usage of touch or speech with independent fusion*: either touch screen or speech input can be used to enter text based questions about the exhibits to the chatbot.
- *Sequential usage of computer vision and textual information with combined fusion*: the camera can be used to enter visual information that is recognized by means of computer vision. If an image was recognized specific information about the the corresponding exhibit can be asked, e.g. after taking a picture from a sculpture the user can use touch screen or speech input to ask "Who made this work?".
- *Sequential usage of textual information and exhibition sensors with combined fusion*: after specific textual input via touch screen or speech input, e.g. "Where does it go on?", exhibition sensor processing (e.g. beacons) can be used to guide the user to the closet exhibit on the tour.

5 Conclusion and Future Work

The development of a practical technical solution for the use of chatbots in the museum environment represents a challenging task: the editing process for audio and media guides is a highly specialized process, and the extension to knowledge-based approaches for the realization of a chatbot for museums has to be carried out. The interplay of existing approaches for intention detection and dialog management opens up a number of research questions, including how chatbots can be used in complex environments such as museums. A solution for hybrid knowledge processing and the technical implementation of information processing, multimodal intent detection and dialog management for museum chatbots will be explored in ChiM. Different multimodal use cases will be studied. The success of the project will be assessed by a demonstrator implemented iteratively in two development phases. To ensure the acceptance of ChiM the overall system will be designed, developed, analyzed and optimized in cooperation with museums and visitors. Focus groups and field studies in various museums will be conducted.

6 Acknowledgments

We thank the Städel Musem Frankfurt, the Germanische Nationalmuseum Nürnberg and the LVR- LandesMuseum Bonn for their interest in supervising, testing and evaluating the ChiM developments.

References

1. Tom Bocklisch, Joey Faulker, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

2. Massimo Canonico and Luigi De Russis. A comparison and critique of natural language understanding tools. *CLOUD COMPUTING 2018*, page 120, 2018.
3. Zhiqiang He, Binyue Cui, Wei Zhou, and Shigeki Yokoi. A proposal of interaction system between visitor and collection in museum hall by ibeacon. In *Computer Science & Education (ICCSE), 2015 10th International Conference on*, pages 427–430. IEEE, 2015.
4. TechTarget Homepage. https://searchcrm.techtarget.com/definition/chatbot, 2018.
5. Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London:, 2014.
6. Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents*, pages 329–343. Springer, 2005.
7. Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080, 2007.
8. Christophe Mollaret, Alhayat Ali Mekonnen, Isabelle Ferrane, Julien Pinquier, and Frédéric Lerasle. Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.
9. Sebastian Möller, Klaus-Peter Engelbrecht, Christine Kuhnel, Ina Wechsung, and Benjamin Weiss. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 7–12. IEEE, 2009.
10. Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. Realtime computer vision with opencv. *Queue*, 10(4):40, 2012.
11. M Golam Rashed, Ryota Suzuki, Antony Lam, Yoshinori Kobayashi, and Yoshinori Kuno. Toward museum guide robots proactively initiating interaction with humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 1–2. ACM, 2015.
12. Aaron Ruß. Mmir framework: Multimodal mobile interaction and rendering. In *GI-Jahrestagung*, pages 2702–2713, 2013.
13. Blaise Thomson. *Statistical methods for spoken dialogue management*. Springer Science & Business Media, 2013.
14. Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*, 2017.