Are You Human – Adapting and Evaluating the Bot Interaction Patterns Towards the User Expectation

Andreas Lommatzsch, Danuta Ploch, Benjamin Kille, Sahin Albayrak
TU Berlin, DAI-Labor, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
{firstname.lastname}@dai-labor.de

Abstract. A rapidly growing number of domains employ chatbots to strengthen communication with users. Systems have evolved in such a way that more and more users struggle to detect whether they talk to another human or a machine. The distinction between human and machine, on the other hand, guides users’ expectations. Assuming they are talking to a fellow human, users will inevitably be disappointed as the machine fails to understand their information needs. We study a large-scale chatbot system’s dialogs. The analysis detects dissatisfaction caused by incorrect expectations concerning the chatbot’s abilities. We propose to tackle this problem by better clarifying the system’s capabilities.

Keywords: chatbots, information retrieval, context, human-computer interaction, natural language processing, question answering

1 Introduction

Chatbots—a form of ‘Interactive Dialog Systems’—enter a rapidly growing number of domains. The ubiquity of access to web services along with social media platforms’ growing popularity accelerates the spread of chatbots.

Interactive dialog systems operate in a wide range of domains with varying sophistication ranging from online banking machines, to the Eliza system [8] (simulating a psychotherapy session) to complex smart home assistants (such as Alexa or Siri) to Chit-Chat-bots [1, 2]. Many current chatbots try to imitate humans in a chat using an avatar and interactively showing messages such as ‘reading’ or ‘agent is typing’. This gives users the impression that they can interact with the bot as if it were a human agent [4, 3]. Consequently, users might overestimate the system’s capabilities and use a language too hard to parse. As a result, the system observes unsuccessful dialogs, which leave users disappointed. Thus, chatbot developers should be aware of this fact and apply suitable methods.

This work studies methods for ensuring a successful human-bot interaction and for detecting misunderstandings. We discuss approaches for adapting to the communication partner and evaluate the suggested methods. We analyze our methods on a large scale chatbot system designed for answering questions related to the administrative services of two major German cities. The bot “Bobbi”
handles \approx 30,000 \text{ chat session for the Berlin administration}; the bot “Frag-Den-Michel” handles \approx 3,000 sessions per month. The answers given by the bots must be reliable and correct as the bots act as official representative of the city administration. The bots cannot answer all questions due to limited knowledge bases. Thus, users should understand the chatbot’s capabilities and be aware of the question domain covered by the bot’s knowledge base. The chatbot should use adequate methods to steer user expectations in order to ensure successful dialogs. A selection of the applied methods is discussed in this work.

2 Approach

In this section, we discuss methods for clarifying the capabilities in a dialog and for adapting to the needs of the communication partner.

Our chatbot system has been developed to support customers by reliably answering all questions related to services and locations of the city administration. The customers should understand the chatbot’s competencies. The chatbot should enable customers to implicitly learn how to take advantage of the knowledge. The chatbot ought to convince its users with its knowledge concerning the domain. But the users should also become aware about the chatbot’s limitations. Further, users need to understand that an algorithm processes their inquiries.

2.1 The Bot’s Strategies for a Successful Conversation

Our chatbot makes use of several strategies to manage a successful conversation in order to ensure the desired perception.

– **The Bot Introduction:** Opening the chatbots webpage initiates a new dialog. First, the chatbot introduces itself emphasizing its purpose and capabilities. Keeping the introduction to two short sentences should prevent users from skipping to read it. The introduction notes that users are about to engage with a computer program. Besides, the chatbot is equipped with a non-human avatar.

– **Grounding** Users may ask ambiguously or out-of-scope. Whenever the system responds, it includes a reference to the detected intent. Thereby, users should gain a better understanding of the system’s inner workings. This procedure, referred to as ‘grounding,’ aims to assure that users can comprehend to what information need the system responds. This method has been proven to be effective in critical situations. In our chatbot system, the explicit grounding ensures that users understand the state of the chatbot. Being aware of the chatbot’s assumptions, users can understand where undesired responses originate. They become more aware of the answer’s context. This information helps users to reformulate their questions and avoid ambiguities.

– **Proactive Bot Behavior** The chatbot keeps reminding users about its limitations. Automatic recommendations help to direct the conversation toward topics inside the chatbot’s knowledge base. New users can follow the
suggestions and learn about the system’s capabilities. We have noticed that users express their information needs in a markedly different way than expressed in the documents underlying the system’s knowledge base. Hence, users perceive the suggestions as helpful both with respect to the content and style.

- **Answer and Language Style** The chatbot’s answers follow a fixed scheme (grounding, general service information, details, suggestions) to gain users’ trust. This format might appear a bit monotonous. Still, it guarantees a good readability and emphasizes that the non-human chatbot is responding.

The discussed methods underline the chatbot’s capabilities. Such strategies matter especially for new users. The more users become accustomed with the system, the better they can adapt and adjust their expectations. Personalization could reduce the emphasis that the chatbot is non-human as well as the “verbosity” of the grounding mechanism for recurring users. Further, “short cuts” to frequently asked questions could be defined. Besides, small talk could become less and less relevant to experienced users.

2.2 **Handling Explicit and Implicit User Signals**

Chatbots must monitor users’ reactions to continuously improve. Recorded dialogues serve as data basis for optimization. The dialogues contain both explicit and implicit signals.

- **Emotional Speech** Users contacting a service center—either in the form of a hotline or a chatbot—tend to have issues that require timely solutions. Frequently, users have tried other communication channels before turning to the chatbot and are angry about the insufficient support. Their language can reveal their emotionally charged state of mind. Whenever the system recognizes emotionally charged users, the system ought to attempt calming them down. The bot can try to provide further explanations, emphasize its limitations, refer to alternative information sources, or hand over to a human operator. The situation can be handled without emotions, clarifying the best solution for the problem.

- **Extended Small Talk and Long Monologs** The chatbot can adequately react to some basic questions about the weather, family, food preferences etc. that are unrelated to the target domain. This kind of small talk can serve, e.g., as a nice conversation starter. Our chatbot tries to detect such extended small talk (exceeding politeness), answers to it but then quickly moves on thus showing that it is not a human. It explains its purpose and directs the conversation towards the actual domain if possible.

- **Explicit User Feedback** Our chatbot interface allows users to give feedback about its answers throughout the entire conversation. For this purpose, thumbs up and thumbs down symbols are placed next to each answer to evaluate its quality. In addition, the bot may suggest to start a “feedback dialog” in the scope of which the user can explicitly describe positive and negative
experiences with the bot conversation. This feedback gives us insights in the users’ perception of the dialog. It helps us to improve the methods for guiding the conversation and for detecting shortcomings.

The interpretation of the “weak” user signals should be always performed in their context. Studying logged chats, we find a wide spectrum of communication styles and expectations. Thus, the more frequent a user interacts with the bot, the more the bot can be trained on observed interaction patterns. Explicit hints to the capabilities of an agent might become annoying for users. Users need more time for the dialog and for reaching the goal; but these signals are effective to avoid misunderstandings. Explicit signals should be used carefully and dependent from the specific context.

3 Evaluation and Preliminary Results

In this section, we report our experiences with operating two chatbot instances answering questions related to the services offered by the administration of Berlin and Hamburg. We report observations and discuss our experiences. The quantitative evaluation is based on the chat logs collected in 2019 with the bot Bobbi.

**Bot Introduction and Small Talk** Our chatbot introduces itself whenever a customer starts a new conversation. The chatbot emphasizes that it is just a computer program that answers standard administrative questions and gives some examples. Still, the collected feedback shows that several users seem to skip the reading of the bot introduction and believe that they are chatting with a human.

About four percent of interactions relate to small talk. Therein, opening and closing phrases make up about forty percent. Another roughly fifteen percent concern the chatbot’s functions and skills. Few users ask ‘ordinary’ small talk questions. The lack of such small talk suggests that the system makes sufficiently clear that it has not been built for chit-chat.

**Grounding** Our chatbot operates as an official communication channel of the administration. Thus, the bot must provide reliably correct information. To ensure that users understand the answer’s context, the bot explains what intention has been extracted and to which problem the answer relates. The bot presents a list of matching interpretations of the user question from which the user can select the intended option.

Our data show that in thirty-six percent of dialogs, the system ranks the desired intention the highest. Users frequently appear to search for less popular services, which only partially match their questions. The explicit grounding ensures that the bot provides an answer matching the user intentions and avoids inadequate answers. In our system, grounding is a suitable method to prevent misunderstandings and to increase the comprehensibility of the answers even if it is sometimes perceived as irritating.
**Proactivity** The chatbot provides recommendations of additional, important aspects. We examine how customers respond to the recommendations. Users ignored recommendations in about twenty-two percent of cases. Ignoring recommendations could be due to unfavorable placement or mismatching users’ information needs. In general, the recommendations help new users to find the relevant information and to highlight the bot’s capabilities; but the recommendations must consider the context and the user experiences.

**Handling Difficult Questions** Looking at the chat logs, we failed to recognize the intention of a user question in about five percent of cases. Reasons for the system’s inability to respond adequately include off-domain questions, questions outside the bot’s responsibilities, as well as misspellings or unknown wording.

Off-Domain and off-responsibility questions remain a challenge for the bot. Lack of data leaves the chatbot no other choices but to reject the question and repeat its limitations. This kind of answer typically dissatisfies users. Assuring that users understand the systems limitations is crucial to avoid such disappointment.

The system can handle unusual wording. It asks users to rephrase their questions. This is particularly important as users and the administration use a different vocabulary. Our analysis shows that about seventy percent of cases lead to reformulation if the system asks for it. Reformulated questions tend to be shorter. Initial questions contain 7.0 words on average, whereas rephrased questions contain on average 4.3 words. The reformulated questions focus on keywords—this simplifies the bot the intent extraction. The good acceptance of the query reformulation shows that most users are interested in reaching their goal and that they try to adapt to the chat partner.

**Handling emotional, angry responses** Another challenge for chatbots comes in the form of angry or emotional user input. Sentiment classifiers [5] can detect anger. Evaluating the reasons for emotional user inputs, we found that users are angry that administrative processes are difficult to understand or that appointments are unavailable at the time. The chatbot itself appears not to be the source for their anger in most cases. Thus, the bot should clarify that it is not responsible for the trouble and provide general hints, how to solve the problem.

**Discussion** We have discussed approaches and metrics for measuring the ‘style’ of communication with a chatbot. The metrics allow us to optimize the bot with respect to specific objectives, e.g. imitate the chat behavior of humans or highlight the specific characteristics of a machine. In successful communication both partners should honestly show their capabilities but also try to adapt to each other. The balance between an extremely machine-like behavior and simulating a completely human behavior should be found.

**4 Conclusion and Future Work**

This work discusses methods for ensuring successful communication between a chatbot and its users based on the experiences obtained through running large
Chatbot systems. We have highlighted the value of showing users the system’s capabilities as well as the limitations. Using the proposed methods can help to adjust users’ expectations, steering clear of misunderstandings, and guide the conversation into fruitful territory (with respect to the bot’s domain and competences).

Chatbots can efficiently support users in a wide range of domains. They are always friendly and available. Users typically are interested in a fruitful conversion and reaching their goal. In order to avoid misunderstandings, both chat participants should understand the capabilities of the partner and adapt to the each other’s needs. Thus, pretending wrong facts is risky and often leads to problems. We have shown that explicit and implicit signals should be used if the communication partners cannot see each other and only interact in a chat.

This pattern is also very important, if the chat is handed over to a new communication partner, e.g. from a bot to a human whenever the bot cannot handle the problem. In case of a handover, the new participant should clearly explain their tasks and responsibilities, as specific experts for different domains and tasks exist. The often promoted idea that a chatbot should completely imitate a human agent can result in inflated expectations and anger about the failure to deliver the right information.

With the rapid developments in NLP and knowledge processing future bots will have improved capabilities, but it will take time for them to mature. Chatbots should continue to express their capabilities and limitation. In addition, we can expect humans to generally understand the bot “needs” with the growing number of chatbot instances.

References