UPCOMING NEW ITU-T RECOMMENDATION ON THE EVALUATION OF TEXT-BASED CHATBOTS

Sebastian Möller[1,2], Stefan Hillmann[1], Thilo Michael[1], Jan Nehring[2] & Tim Polzehl[2]

[1]Quality and Usability Lab, TU Berlin
[2]Speech and Language Technology Lab, DFKI Berlin
sebastian.moeller@tu-berlin.de

Abstract: The evaluation of spoken dialog systems has been an object of scientific research for decades. Whereas standardized methods were made available by the International Telecommunication Union (ITU-T), a comparable level of maturity is still missing for the evaluation of text-based chatbots. This contribution presents ongoing work developing a new ITU-T Recommendation describing subjective evaluation methods to quantify the quality of services relying on text-based chatbots, as experienced by the users of such services. Chatbots addressed by the upcoming Recommendation enable a text-based natural language interaction with a human user via a text interface on a turn-by-turn basis. They possess natural language understanding, dialogue management, and natural language generation capabilities. The evaluation methods address different aspects of quality from a user's point of view, taking the chatbot as a black box. They are based on laboratory or remote experiments in which participants interact with the chatbot in order to perform a pre-defined, realistic task. The participant's opinion on perceptive quality dimensions is solicited with the help of questionnaires, and examples of such questionnaires are provided.

1 Introduction

In parallel to the rise of speech-based assistants such as Siri or Alexa, more and more text-based chatbots pop up, which allow for a text-based natural language interaction with some type of service, such as question answering, customer care portals, or upselling. The success of these chatbots seems to suggest that their users well accept them; however, a common methodology for assessing system performance andquantifying the quality of experience from a user's perspective is still missing. This is in contrast to evaluation methods for spoken dialog systems, for which a long tradition of research (see, e.g., [1] for an overview), but also agreed-upon international standards [2][3] are commonly available.

A new work item was started recently in Study Group 12 of the International Telecommunication Union (ITU-T) on the evaluation of text-based chatbots. The work should result in a new Recommendation P.SEC similar to ITU-T Rec. P.851 [2] for spoken dialog systems. In addition, there might be a need to define parametric descriptions of text-based interactions, in a way similar to the parameters defined in Suppl. 24 to P-Series Recommendations [3]. In this paper, we provide an overview of the future Recommendation P.SEC, based on the draft provided by the authors in [4].

2 Overview of the Recommendation

2.1 Overview

The Recommendation will cover three main topics, plus introductory formalities regarding references and terminology, and concluding remarks on the analysis and interpretation of information collected with the described methods. The three topics in focus are a detailed definition of chatbots addressed by the Recommendation, remarks about the experimental set-up for

evaluations, and questionnaires to collect user ratings. The following paragraphs outline the related structure of the Recommendation and hand in hand the structure of this paper.

Instead of traditional chatbots, which focus on maintaining a language-based informal interaction for its own sake (sometimes called chitchat bots), most commercial chatbots now focus on question answering or providing information through a multi-turn natural language interaction between user and computer. Compared to spoken language, this written language interaction partly follows different rules. The interaction can be asynchronous, the user can more easily review the history of the interaction, and speech recognition errors are bypassed. We then address tasks and components of text-based chatbots, such as natural language understanding, dialog management, and natural language generation components. Further, an overview of quality aspects is provided, such as effectiveness and efficiency, usability, user satisfaction, input quality, output quality, and cooperativity.

To quantify these aspects, we propose a protocol for subjective evaluation experiments in Section 4. The experimental set-up of such experiments is described, including the system set-up, possible test scenarios, requirements for test participants, and the layout of questionnaires.

We provide example questionnaires to be distributed after each interaction in Section 5. They address the user's background, the overall quality, the information provided by the system, the communication with the system and its interaction behavior, the user's impression of the system, and finally, its acceptability. A reduced set of questions is then proposed to focus on the overall impression of the chatbot, after several interactions with it.

In the following sections, we will highlight the main aspects of each of the three content chapters of the Recommendation, as outlined in the draft [4].

3 Chatbots

3.1 Tasks and Components of Chatbots

The task of a chatbot is the exchange of information between a user and a computer system through written language. For the chatbots addressed by the Recommendation, the exchange commonly serves a dedicated task, such as answering a question, performing a transaction, or alike. Because written language is a common medium for humans to interact with each other, its use for human-computer interaction does not require any particular knowledge from the user, except that the user needs to be sufficiently familiar with the language used (in case of non-native users), and that the user needs to read and type written messages without problems (such as illiteracy, or viewing or motor difficulties). In addition, as the interaction devices typically include a computer or on-screen keyboard and a pointing device (e.g., mouse or touchscreen), the user needs to be familiar with the operation of such devices.

The interaction in the chatbot is commonly organized in several components which form a pipeline of information processing:

(1) A component to analyse the meaning of what was written by the user (Natural Language Understanding, NLU), using, e.g., simple keyword matching or more sophisticated machine learning techniques. Semantic concepts represent the meaning, e.g., in terms of intents and attribute-value pairs (AVPs), more abstract dialog acts (cp. [5]), or knowledge representations such as knowledge graphs.

(2) A component for interaction management (Dialog Manager, DM). It can be implemented either with the help of rules (following a dialog grammar), or with statistical transitions learned from training data using supervised or reinforcement learning. Whereas a rule-based Dialog Manager might be relatively easy to implement and has the advantage that it can easily be verified with respect to misleading dialogs, a corpus-

based Dialog Manager may be more flexible and may cover a wider range of interaction phenomena.

(3) A component for outputting the next contribution to the user (Natural Language Generation, NLG). It can base on pre-defined templates which are filled and concatenated with textual information, or again on machine-learning techniques.

In addition to the mentioned text processing components, a chatbot is equipped with a graphical output device, as well as an input device allowing for text input (soft or hard keyboard).

3.2 Interactions with Chatbots

A task may require one or several interaction steps to be performed, called “turns”. Turns may consist of complete sentences, sentence fragments, or single words or expressions. Commonly, chatbots try to provide complete sentences as output, whereas they also accept text fragments or text commands as input. In addition, chatbots may provide suggestions to the user as to which type of input they expect at a certain state of the dialog; in that case, the textual output of the system might be augmented by radio buttons, selection boxes, or alike. Whereas these input options might be helpful to guide the user and to lead the dialog to an expected goal, they are rather similar to options in a graphical user interface. They require other evaluation techniques which are not discussed in the Recommendation.

3.3 Quality Aspects and Influencing Factors

Because the chatbots addressed by the Recommendation mainly serve to accomplish a task goal, effectiveness and efficiency are important quality aspects. Effectiveness describes the accuracy and completeness with which specified users can achieve specified goals in particular environments. A common metric for effectiveness is a task completion rate (task success). Efficiency relates goal achievement to the resources used, such as time, cognitive effort, or alike. Common metrics are the interaction duration, the number of turns written by the system or the user, or the cognitive demand put on the user to perform the interaction.

Effectiveness and efficiency are central constituents of the usability of a chatbot; see the taxonomy of quality aspects of spoken dialogue services [1][2]. Usability, however, is generally defined in a much broader sense, and describes the capability of the service to be understood, learned, and used by specified users under specified conditions. It indicates the suitability of the service to fulfil the user's requirements, includes effectiveness and efficiency of the system, and results in user satisfaction. User satisfaction indicates the service's perceived usefulness and usability for the intended user group. It includes whether a user receives the wanted information, is comfortable with the service, and receives the information within an acceptable elapsed time [2].

The quality of the written interaction largely determines its functional quality. This includes the input quality (i.e., whether the user feels understood by the system), the output quality, the cooperativity of system behaviour, and the symmetry of the written dialogic interaction. Cooperativity can be defined in the sense of applying the cooperative principles of conversational communication, as defined by Grice [3]. A system with high interaction quality may result in an efficient interaction, in terms of speed of the interaction, dialogue conciseness, and dialogue smoothness. On the other hand, task efficiency is linked to task success and task ease. Two additional quality aspects result from the usage of natural language as an interaction modality: the “personality” of the chatbot (politeness, friendliness, naturalness of behaviour) and the effort required from the human user for the interaction (ease of communication, stress/fluster, etc.). These aspects can be subsumed under the term comfort. Together with communication and task efficiency, they contribute to usability, for which user satisfaction can be seen as an

indicator. On the other hand, service efficiency indicates the adequacy of the service for the given task, in comparison to other services which might be used instead. Usability, service efficiency, and economic benefit result in the utility of the service, and finally in its acceptability.

The chatbot exercises an influence on the mentioned quality aspects in two ways: Regarding the task a user can carry out with its help (task factors such as how well the chatbot captures the task it has been designed for, the complexity of the task), and regarding the factors which influence the dialogic interaction (agent factors). In addition, the usage environment may be an influencing factor (physical environment such as light or moving conditions, social environment such as other persons being present during the interaction). Finally, the user's characteristics (aims, experience, expectations, typing behaviour, linguistic background, etc.) influence perceived quality.

The quality of a service finally results from the perceptions of its users in relation to what they expect or desire from the service. It is highly dependent on the situation in which perception and judgement occur. This fact has to be taken into account when carrying out subjective quality evaluation experiments, namely by creating a sufficiently natural test situation and a realistic test user motivation.

3.4 Subjective Evaluation Methods

To quantify performance and quality, subjective experiments with representative users are commonly carried out. These experiments serve two main purposes:

1. During the interaction, instrumentally measurable system parameters are collected, and the messages of the chatbot and of the user are logged. The log-files are submitted to an expert evaluation, which results in a set of parameters describing specific aspects of the interaction on the turn, dialogue,and task level.
2. Before and after the interaction, test participants are given a questionnaire that aims to collect information about the expectations and the perceived quality features experienced during the interaction.

In laboratory experiments, both types of information can be obtained in parallel. In a field test, however, instrumentally logged interaction parameters are often the unique source of information for the service provider to monitor the quality of the system.

4 Experimental Set-up

Subjective interaction experiments with chatbots are recommended to be set up according to the general rules for subjective quality tests established by the ITU-T in [6], [7] and [8]. Laboratory tests are commonly carried out in "neutral" environments, such as sound-shielded rooms with daylight imitation. Such a controlled environment may generate misleading results with respect to the impact of device and display size, e.g., on a mobile device. In such a case, it might be better to allow for more realistic – but less controlled – usage situations, to reach a better ecological validity.

4.1 System Set-up

A set-up providing the full functionality of the chatbot in real time has to be implemented to allow interactions with human users. The exact nature of the set-up will depend on the availability of system components and thus on the system development phase. If system components have not yet been implemented, or if an implementation would be unfeasible (e.g., due to the lack of data) or uneconomic, a simulation of the respective components is required.

The simulation of the interactive system by a human being (the so-called "wizard"), i.e., the Wizard-of-Oz (WoZ) simulation, is a well-accepted technique in the system development

phase. However, in such a case, any deviations from the real system (e.g., a different response time of a human wizard compared to an NLU component) should be reported and considered in the interpretation of the results.

4.2 Test Scenarios

Because of the lack of a real motivation, laboratory tests should make use of experimental tasks which the participants have to carry out. Such an experimental task provides an explicit goal, but this goal should not be confused with a goal which a user would like to reach in a real-life situation. Because of this discrepancy, valid user judgements on system helpfulness and acceptability cannot easily be obtained in a laboratory test set-up. The experimental task is commonly defined by a scenario description, providing control over the task carried out by the test participants, while at the same time covering a wide range of possible situations (and possible problems) in the interaction. Scenarios can be intentionally designed to test specific system functionalities (so-called development scenarios), or to cover a wide range of potential interaction situations which is desirable for a summative evaluation.

Unfortunately, pre-defined scenarios prime the users to interact with the system. Written scenarios may invite the test subjects to imitate the language given in the scenario, leading to copying the scenario text instead of using the own language. Consequently, it might be desirable to present scenarios with the least amount of priming language and to use graphics, pictograms, or alike.

4.3 Test Participants

The purpose of the test should guide the choice of test participants. For example, analytical assessment of specific system characteristics will only be possible for trained test participants who are experts of the system under consideration. However, this group will not be able to judge overall aspects of system quality in a way which their knowledge of the system would not influence.. Valid overall quality judgements can only be expected from test participants who match the group of future service users as close as possible.

User factors that are expected to influence the behaviour and perception of the user should be taken into account in selecting test participants in a representative way. Some of these factors are related to the vision, motor capacity and language usage which may be expected, such as age, individual body characteristics, physical status, or native language. Other factors relate to the experience and expertise with the system, including the input and output devices, the task, and the domain. It can be expected that these factors will impact language usage, and thus the interaction with the system.

Users seem to develop specific interaction patterns, so-called practices, when they get familiar with a new system. These practices may reflect a “cognitive model” the user develops of the system and depend on the user’s former technology acquisition processes. Such a model is partly determined by the messages given to the system, and partly by the messages coming from the system. The user generally assumes that their utterances should be well understood by the system. In case of misunderstandings, the user gets confused, and dialogue flow problems are likely to occur.

5 Questionnaires

To obtain quantitative information about quality features perceived by the user, subjective judgements are collected on rating scales presented on questionnaires. In addition, open text answer options allow to collect subjective experiences which are not covered by the rating scales.

Scaling tasks can be carried out relatively easily, and untrained participants often prefer this method of judgement. For rating the quality of chatbots, judgements on discrete or continuous rating scales are usually solicited from the test participants. The rating task is sometimes described in terms of a statement (e.g., "The system was easy to understand."), and test participants have to express their agreement on the statement by marking the respective tick or category of the scale. Numbers are attributed to the categories or to the scale positions, depending on whether the statement is positive (from 1 for "strongly disagree" to 5 for "strongly agree") or negative (from 5 for "strongly disagree" to 1 for "strongly agree"), and the individual ratings are summed up for all participants.

5.1 Questions Related to the User's Background

Before the first interaction with the chatbot under test, participants are commonly asked to respond to a number of questions related to the user and his background, which is relevant to the experiment. These questions address the following items:

- Personal information: Age, gender, area or birth, current residence, language proficiency, visual impairments, motor impairments.
- Task-related information: Frequency of task, usual approach when resolving the task (alternative interfaces), motivation, other important task- and domain-related aspects.
- System-related information: Experience with chatbots, experience with speech technology devices (speech recognition, synthesized speech, etc.), experience with the haptic input device (smartphone, tablet, laptop, etc.).

5.2 Questions Related to an Individual Interaction

After each interaction with the system, questions are solicited which relate to the just-finished experience. These questions may address e.g.

- Overall quality
- Information provided by the system
- Communication with the system
- System behaviour
- User's impression of the system
- Acceptability

Table 1 gives a list of exemplary questions/statements.

5.3 Questions Related to the User's Overall Impression of the System

After carrying out interactions with the chatbot under test, an additional questionnaire can address the user's overall impression of the system, aggregating over positive and negative experiences. The questions on this questionnaire may address e.g.

- Overall quality
- System behaviour
- User's impression of the system
- Usability
- Acceptability

A list of exemplary questions/statements is given in Table 2.

6 Conclusions and Future Work

At the time of the ESSV 2022 conference, the Recommendation text is still in a draft format, and may be amended and corrected based on the input of conference participants. A stable draft is expected to be approved at the upcoming ITU-T Study Group 12 meeting in June 2022.

Feedback on the draft is explicitly welcome. Parallel to the work on the Recommendation draft, the methods described were and are being applied in various chatbot evaluation studies. The experiences and results gained in these studies have been submitted for publication and will also be incorporated into the final Recommendation.

7 List of References

[1] MÖLLER, S.: *Quality of Telephone-based Spoken Dialogue Systems*, Springer, New York NY, 2005.

[2] ITU-T RECOMMENDATION P.851: *Subjective quality evaluation of telephone services based on spoken dialogue systems*, International Telecommunication Union, Geneva, 2003.

[3] ITU-T SUPPL. 24 TO P-SERIES RECOMMENDATIONS: *Parameters describing the interaction with spoken dialogue systems*, International Telecommunication Union, Geneva, 2005.

[4] ITU-T CONTRIBUTION COM 12-590: *Draft Recommendation P.SEC on the evaluation of text-based chatbots*, Source: TU Berlin (Authors: S. Möller, S. Hillmann, T. Michael, J. Nehring and T. Polzehl), ITU-T Study Group 12 Meeting, Geneva, 12-21 October 2021.

[5] WEISS, B., HILLMANN, S., MÖLLER, S.: Intents in Sprachdialogen: Eine Praxisperspektive. In: Elektronische Sprachsignalverarbeitung 2021. Studientexte zu Sprachkommunikation, Bd. 99, S. 192-199, TUDpress, Dresden, 2021.

[6] ITU-T RECOMMENDATION P.800: *Methods for subjective determination of transmission quality,* International Telecommunication Union, Geneva, 1996.

[7] ITU-T RECOMMENDATION P.910: *Subjective video quality assessment methods for multimedia applications,* International Telecommunication Union, Geneva, 2008.

[8] ITU-T RECOMMENDATION P.911: *Subjective audiovisual quality assessment methods for multimedia applications,* International Telecommunication Union, Geneva, 1998.

Table 1: Quality aspects related to an individual interaction.

Quality aspect	Item
Overall quality	Overall impression
Information provided by the system	The system has provided you the desired information. The system's answers and proposed solutions were clear. You would rate the provided information as true. The information provided by the system was complete.
Communication with the system	The system always understood you well. You had to concentrate in order to understand what the system expected from you. The system's responses were well readable. You had to put high effort into typing in your messages. You were able to interact efficiently with the system.
System behaviour	You knew at each point of the interaction what the system expected from you. In your opinion, the system processed your specifications correctly. The system's behaviour was always as expected. The system often failed to understand you. The system reacted naturally. The system reacted flexibly. You were able to control the interaction in the desired way. The system reacted too slowly. The system reacted politely. The system's responses were too long.
Dialogue	You perceived the dialogue as natural. It was easy to follow the flow of the dialogue.

Quality aspect	Item
	The dialogue was too long.
	The course of the dialogue was smooth.
	You and the system could clear misunderstandings easily.
	You would have expected more help from the system
User's impression of the system	Overall, you were satisfied with the dialogue.
	The dialogue with the system was useful.
	It was easy for you to obtain the information you wanted.
	You have perceived the dialogue as pleasant.
	You felt relaxed during the dialogue.
	Using the system was fun.
Acceptability	In the future, you would use the system again.
	You would recommend the system to a friend.
	You were satisfied with the solution offered by the system.

Table 2: Quality aspects related to the overall impression of the system.

Quality aspect	Item
Overall quality	Overall impression
System behavior	The system's way of expression was clear/unclear.
	The system reacted politely/impolitely.
	You would have expected more help from the system.
	The system was able to answer all of your questions.
	Misunderstandings could be cleared easily.
	The system controlled the flow of the dialogue.
	You were able to handle the system without any problems.
User's impression of the system Usability	You enjoyed the dialogues.
	The handling of the system was easy/complicated.
	The system appropriately informed you about its capabilities.
	The interactions with the system were worthwhile.
	You perceived this possibility to obtain information as helpful/not helpful.
	You rate the system as reliable/unreliable.
Acceptability	You prefer to use another source of information.
	You prefer a human operator.
	In the future, you would use the system again.
	Which characteristics of the system did you like most?
	Which characteristics of the system disturbed you mostly?
	Do you have any proposals for system improvement?