*Healthcare of the Future*
*T. Bürkle et al. (Eds.)*


*doi:10.3233/978-1-61499-961-4-101*

# Improving and Evaluating eMMA's Communication Skills: A Chatbot for Managing Medication

Gabriel Immanuel HESS[a,1], Guillaume FRICKER[a,1], Kerstin DENECKE[a,2]
[a]*Bern University of Applied Sciences, Biel, Switzerland*

**Abstract.** In previous work, a mobile application for medication self-management (eMMA) was introduced. It contained a basic conversational user interface (CUI). In this work, we extended the CUI by integrating the chatbot framework RiveScript and an instruction interface. To study task success, dialog quality and efficiency, we performed a theoretical and a quantitative evaluation as well as a usability test. The results show that the technical extensions of eMMA were useful to improve the chatbot's quality. However, the underlying knowledge base still requires substantial extensions before the system can be used in practice.

**Keywords.** Chatbot, medication self-management, mHealth, conversational user interface

## 1. Introduction

Many mobile applications exist for patients managing their prescribed medication. Within the "Hospital of the Future live" Project [1], the mobile application eMMA (referred to as eMMA 1.0) was introduced as an electronic medication management assistant for persons prescribed to medications within an age range between 18 and 85 [2]. The goal was to address the problems of improving patient's medication adherence and communicating medication data with health care providers, as well as serving patients as an educational source for drug information. Unlike other electronical medication diaries, eMMA uses a standardized format for medication data and is built with a conversational user interface (CUI) to simulate the interaction with a human assistant. A CUI is not only expected to be handled easier by elderly people, the assumption is that the illusion of interacting with an actual assistant could also improve medication adherence. eMMA 1.0 relied on a CUI with restricted knowledge base, only able to respond to several key words and a selection of drug names [2]. In this paper, we describe the extension of eMMA's CUI by integrating a rule-based chatbot engine and an extended knowledge base (referred to as eMMA 2.0) for improving the quality of the CUI. Furthermore, we conducted a three-stage analysis for evaluating these extensions. This analysis consists of a theoretical analysis using a feature checklist, a quantitative analysis evaluating chatlogs from test persons using eMMA 2.0 and a usability test.

[1] Contributed equally

[2] Corresponding Author: Kerstin Denecke, Berner Fachhochschule, Quellgasse 21, 2502 Biel, Switzerland, kerstin.denecke@bfh.ch

## 2. Methods

### *2.1. Chatbot implementation*

To decide on a technology stack, different chatbot frameworks were evaluated. Criteria for the evaluation included: the capabilities of the chatbot service (in particular the ability of handling conversation context), the possibility of integrating it into eMMA 1.0 and the ability of working with medication data. Services running on external servers were excluded to ensure data privacy and continuous availability. Once a chatbot engine was chosen, the corresponding rule set that defines the system's knowledge base was defined. For this purpose, first user tests with six test persons, recruited from the author's personal background, were conducted, to clarify the weaknesses of the CUI of eMMA 1.0. All test persons had already made first experiences in using chatbots. In additional iterations, the rule set was extended.

### *2.2. Chatbot evaluation*

The chatbot was evaluated in three stages. First, a theoretical analysis using the TRINDI framework was made to get a benchmark of the enhanced CUI. TRINDI is a checklist comprising 3 groups with 16 questions addressing a dialog-based system's competences [3]. The first group (9 questions) refer to the flexibility in dialogue handling. The second group (5 questions), addresses the overall functionality of the system. The third group (2 questions) deals with the system's ability of context awareness. The checklist was independently filled out by two of the authors, with the answering options of "yes", "partially", "in theory", "no" and "unknown", as suggested by [3]. The resulting checklists where compared, and divergences were discussed until consent was achieved.

For the quantitative analysis, eight logs from eMMA 2.0 were compared to five logs from the tests with eMMA 1.0, where one log was lost due to technical problems. The analysis was done along the categories suggested from the PARADISE framework [4], namely task success, dialogue efficiency, dialog quality and user satisfaction. While user satisfaction was measured by a questionnaire accompanying the usability tests, the other categories could be derived from transcribed conversation logs. In order to analyze task success, individual tasks were identified in the logs and graded successful (coded TRUE) or unsuccessful (coded FALSE). For dialog efficiency, the number of dialogue steps used to complete such a task was counted. Dialog quality was measured by two variables: 1) the systems answering time, and 2) the adequacy of the systems immediate reply.

Furthermore, a usability test of the CUI of eMMA 2.0 was conducted with a sample of eight test persons. Since these test persons were not prescribed to medications at the time of the test, a scenario was created for getting them in a context where a medication management assistant is applicable. For surveying and comparing user satisfaction among the two versions of eMMA, the questionnaire from the usability test of eMMA 1.0 was used, consisting of 12 questions that were answered on a Likert scale from -2 to +2. Thus, the questionnaire could be completed within a range from -24 to +24 points.

To put the results from the quantitative analysis into context, the test persons additionally answered a short survey to what extend they would tolerate the lack of task success, dialog efficiency and dialog quality from a chatbot in a medical context like eMMA.

## 3. Results

### *3.1. Implementation*

The chatbot framework RiveScript was chosen for reasons of privacy, ease of implementation and availability of German language foundations. It was included into eMMA 2.0 using the RiveScript 1.19 node module as an interpreter, running directly on the device. This interpreter module works as a black box, generating an answer to a user utterance, based on multiple rules specified in RiveScript files. The knowledge base of eMMA 2.0 was built on multiple pillars: First, basic German language understanding was added by a static script from the ALICE chatbot [5]. The file written in AIML could be translated into RiveScript syntax and had to be adapted to be adequate for the context. Answering patterns with medication context were included in another RiveScript file, context-awareness in mind. These patterns are based on the knowledge of eMMA 1.0, but also include results from the first usability tests and were improved iteratively during the entire development process. The Specialty List (http://www.spezialitaeten-liste.ch), a list containing all drug names of approved medications in Switzerland, was parsed to the RiveScript syntax and imported into the chatbot as external knowledge.

Since the RiveScript interpreter has no direct access to the application's memory, a dynamic context file is generated at every launch of the chatbot service, containing for example the user's medication or the name of the general practitioner. Additionally, we implemented an instruction interface that allows the chatbot to control the application through the interpreter. Specific keywords in the returned text string are caught before the answer is displayed to the user and trigger the corresponding action, e.g. displaying the user YES / NO buttons instead of a free text answering field. Other use cases for the instruction interface are adding a medication to the plan or looking up medication details online. With dynamically generated RiveScript rule files and the instruction interface, we enabled a two-way communication between the application and the black-boxed RiveScript interpreter.

### *3.2. Evaluation results*

The theoretical analysis with the TRINDI checklist shows that the implemented chatbot still has room for improvement. None of the sixteen checklist items could be answered with *yes*, five points are fulfilled *partially*. Four other features were assessed with *theoretically*, meaning that the RiveScript syntax would enable them, but not the current implementation. The remaining checklist items were evaluated with *no* (see Table 1). The most important failed item is if the system checks its understanding of the user's utterance and can thus react accordingly.

**Table 1.** Evaluation results of the TRINDI categories

| | **Flexibility** | **Overall functionality** | **Context awareness** |
|---|---|---|---|
| Yes | 0 | 0 | 0 |
| Partially | 2 | 1 | 0 |
| Theoretically | 4 | 1 | 1 |
| No | 3 | 3 | 1 |
| **Total** | 9 | 5 | 2 |

As Table 2 shows, task success and the adequacy aspect of dialog quality were improved within eMMA 2.0. The slightly slower response time can be explained by the more complex pattern matching given the extended rule base. The average number of steps to

complete a task went up within eMMA 2.0. Contrary to expectations, eMMA 1.0 achieved a slightly better user satisfaction, although both versions are in the center of the scale ranging from -24 to 24. The usability test showed that users were able to interact with eMMA 2.0 and successfully finish complex tasks that need to hold the conversation context over several messages.

**Table 2.** Results of the quantitative evaluation

| | Task success | Dialog efficiency | Dialog quality (response time) | Dialog quality (adequacy) | User satisfaction |
|---|---|---|---|---|---|
| eMMA 1.0 | 7.7% | 5.3 steps | 20 ms | 16.9% | 2.6 pt |
| eMMA 2.0 | 62.4% | 9.7 steps | 20 – 50 ms | 59.3% | -3.3 pt |

For context, we asked our test persons in the usability test what percentages of task success or dialog efficiency and quality they would consider acceptable. The resulting 84.2% for task success and 74% of dialog adequacy could not been reached by either of the evaluated versions. Also, the regarded acceptable number of steps of four for a simple and up to 9.5 for a complex task was missed by eMMA 2.0. Unaltered response time was in the range of milliseconds for both versions. Early usability tests showed that these are considered too fast, leading the original developers of eMMA 1.0 to artificially slow down the answering speed to the scale of seconds, which was kept for eMMA 2.0.

## 4. Discussion

In this paper, we introduced a rule-based chatbot to enhance the CUI of an existing medication management application by internal and external context information. The technical extensions led to better values in task success as well as in dialog quality. It turned out that the extended ability for more complex tasks led to a poorer dialog efficiency. To address this, graphical user interface elements could be contextually brought up inside the CUI for quicker handling of complex tasks. In general, the evaluations and user test shows, that, besides fixing technical bugs, the eMMA 2.0 chatbot still needs an enhanced knowledge base and a better context management. This can be achieved with the existing technology stack, but needs massive enhancement of the RiveScript rule files, based on more conversation logs and possibly assisted by machine learning. The introduced instruction interface could also be used to implement a function that allows the chatbot react adequately when it can't understand a user's utterance. Once these extensions have been realized, eMMA 2.0 will have the potential of being released on the market.

## References

[1] T. Bürkle, K. Denecke, M. Lehmann, E. Zetz, J. Holm. Integrated Care Process Designed for the Future Healthcare System. *Studies in Healthcare and Informatics* **245** (2017), 20–24.

[2] M. Tschanz, T.L. Dorner, J. Holm, K. Denecke. Using eMMA to Manage Medication. *Computer* **51** (2018), 18–25.

[3] M. Beveridge, J. Fox. Automatic generation of spoken dialogue from medical plans and ontologies. *Biomed Inform.* **39 (5)** (2006), 482–99.

[4] M.A. Walker, D.J. Litman, C.A. Kamm, A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proceedings of the 8th EACL conference, Madrid, Spain. Jul 7* (2005), 271–80.

[5] ALICE AI Foundation [Internet]. German_1.aiml - Alicebot. [updated 2008 Aug 16; cited 2018 May 6]. Available from: http://alicebot.wikidot.com/aiml:de-de:cdrossman:alice:german-1-aiml.