Towards a Flexible User Simulation for Evaluating Spoken Dialogue Systems

Dmitry Butenkov
Advisor: Sebastian Möller
Quality & Usability Lab, Deutsche Telekom Laboratories, TU Berlin,
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{dmitry.butenkov, sebastian.moeller}@telekom.de

Abstract. The main aim of research is to introduce a new data-driven user simulation approach for the quality and usability evaluation for spoken dialogue systems.

Keywords: Human-Computer Interaction, Human Factors, Statistical User Simulation, Usability Evaluation, Spoken Dialogue Systems, Experimental Design.

1 Problem Definition

Measuring the quality and usability of a spoken dialogue system (SDSs) is a complex continuous process requiring a series of expensive experiments to be established. In controlled laboratory experiments, participants very often tend to demonstrate unnatural communication behavior (e.g. no hang offs, barge-ins, or too much focused on the final goal). Also, a lot of test cases and subject groups have to be covered to provide meaningful evaluation results [1]. Thus, the costs of thorough usability tests in the design phase of spoken dialogue systems are often unreasonably high.

A potential solution is the evaluation of systems in their early development phase. The authors of [2] and [3] have shown that even at initial stages of SDS development usability problems can be discovered. The earlier problems can be identified, the cheaper their correction is. The authors of [2] and [4] have recently reported that semi-automated evaluation and testing can significantly decrease the expenses for quality and usability experiments. Following this idea, a simulation of user-system interactions may help to identify usability problems early and at low costs, and it might even be possible to base a tentative evaluation of the entire systems on simulated (instead of real user) interactions. In some cases time costs can be also be reduced significantly in comparison to a classical subjective evaluation approach. Of course, this solution is a deal between complexity and unification, but it must be able to recover complex enough usability problems.

Some attempts were made so far: the authors of [5] provided interesting, but very limited personality generation service, while authors of [3] and [4] reported first attempts to develop flexible user simulation systems with a focus on usability evaluation. Despite of these prospective results, user simulation is still used not very

frequently, except as an assessment tool to enable various machine learning algorithms by enlarging the original learning corpus [6]. Moreover, the authors of [2] and [7] have shown that all modern user simulators in the field of SDSs are either user- or system-centric. The idea of the present research is to develop a balanced concept for a data-driven simulation of user behaviour towards SDSs, covering several typical user patterns and scalable up to several common systems.

2 Research Hypotheses

A major part of commercial German SDSs has been evaluated in the Voice Awards competition being held since 2003 [8]. Although, there is no detailed information on every particular system available, an analysis indicates that the respective systems are relatively simple in comparison to current research systems, for instance from the smart home control field. The logical consequence is to try to reduce the simulation problem to standard slot-filling mechanisms. Thus, the first research hypothesis is that typical commercial German SDS can be reduced to a standard slot-filling problem.

A second research hypothesis follows from the previous one: since the problem scope is significantly reduced, the generalization and unification may be more straight-forward. This actually means that the same simulation approach can be applied to several systems or several configurations of one system. Of course this is not a panacea and in practice it turns into very different performances on every particular system under test. However, up to now the generalization problem in SDS field was considered to be very complex, so the authors of [2], [6], and [7] note the appropriate solutions are still of high demand.

A third hypothesis recalls typical SDS user patterns. The authors of [9] did a trustworthy marketing research and usability expert workshop. As a result, they have shown that "for a complex field, such as interaction behavior and usability requirements, a one-dimensional description with types might be an oversimplification". Therefore, we propose to make use of a three-dimensional descriptive model of a particular user instance covering memory, behavioral, and motivation (MBM) aspects. It is important to note that a prospective fourth dimension is the emotional state of the user, which might be considered in an eMBM model. However, due to the complex, latent nature of this phenomenon, and because of some technical issues (e.g. reliable emotion recognition techniques) it still stands a very complex task even for a separate scientific research. Thus, the final hypothesis claims the MBM model captures the major aspects of users' behavior for the target SDSs described in the two previous hypotheses.

3 Methods Proposed

The Mental Models (MeMo) project described in [4] was a recent pioneer attempt to develop a probabilistic rule-based user model for usability evaluation. A big work has been done so far, resulting in positive interest from the scientific community. Therefore, the present research proposal will take up some interesting ideas

introduced in MeMo. However, the practical experience has shown that a rule-based approach has some significant drawbacks: the derived rules are often not precise enough to encode all behavioral aspects of human users. On the other hand, due to inherent corpora limitations, a purely data-driven method can not be safely applied to derive the complete user model from the observed interactions. Thus, a hybrid solution of both data-driven and rule-based descriptive models is considered for implementation. In order to predict user behavior, interaction parameters are learned from data, and a top logic of decision-making is affected by rules. Coupled by adaptive motivation system able to react on dynamic environment changes, this paradigm should enable enough flexibility, and coverage of patterns unseen in learning data.

Behavioral aspects are divided into two following classes (both represented by rich set of features): personal user characteristics and observable performance parameters. At first, a cluster analysis is performed on observable performance parameters to detect user stereotypes from the learning data. The factor analysis is then done to match these stereotypes into the user taxonomy and define the appropriate set of personal user characteristics. Of course, this task might fail due to missing or inconsistent data. Gathering the domain-specific knowledge is also a challenge. Therefore, some semantic web search techniques will be used for this propose, supported by an ontology-based reasoning mechanism.

Besides that, learning the optimal task completion path is also an interesting problem for simulating realistic task-driven interactions. Therefore, an appropriate machine learning algorithm will be used to compute this path. Recently, various reinforcement learning (RL) techniques became widely popular in SDS development and related research fields. However, several important constraints make the pure RL not as attractive in practice as it looks in theory. For instance, high computational costs coupled by learning data requirements make it infeasible in practice to find the optimal solution. Moreover, once configured, many classic RL algorithms do not provide enough flexibility and scalability. Last but not least, there is an oversensitivity problem with a meaningful reward function that is very often hand-crafted empirically. Thus, we propose to implement advanced RL technique dealing with the named limitations efficiently. A prospective approach might be the fuzzy reinforcement learning (FRL) paradigm [10].

4 Sketch of Proposed Solution

The proposed solution is planned as a part of the SpeechEval project carried out in collaboration with the German Research Center for Artificial Intelligence (DFKI). The issuer simulation will be implemented on the Ontology-based Dialog Platform (ODP) supporting several standard voice interfaces. The solution itself is being implemented as a standard Java plug-in supported by ODP. The frontend ASR and speech synthesis services are high-quality ready-to-use business solutions by Nuance. A MySQL database engine is a backend transactions support.

The learning phase is based on the Voice Awards contest corpus [8] provided by DFKI. The development phase grounds on a research SDS, the Bochum Restaurant

Information System (BoRIS) and its support corpora gathered in several experimental series [1]. Besides that, Let's Go! Lab [11] and ICT [9] projects' results are used to study usability-related user characteristics.

5 Conclusions

User simulation for evaluation is a young growing field of research. It is still open for significant contributions through applied and fundamental research activities that surely are both of scientific interest and commercial relevance. Thus, the major goal of this research project is to introduce and verify an innovative hybrid approach to user simulation, joining the best results from the SDS field with special focus on usability. The authors hope the research will change the classic role of user simulation in the development and evaluation of SDSs.

A minor goal of this project is to introduce the special FRL technique, adapted for SDS and optimized for usability criteria. Moreover, the structure of this research encapsulates the opportunity to incorporate and validate the results from previous projects in the field, done at T-Labs in cooperation with various partners. Last, but not least is the generalization problem of user simulations across systems and user groups, that stays unsolved up to now. We would like to demonstrate that it is possible to generalize our solution at least within different configurations of the same system.

References

1. Möller, S.: Quality of Telephone-based Spoken Dialog Systems. Springer Science + Business Media Inc., New York, NY (2005)
2. Ai, H., Weng, F.: User Simulation as Testing for Spoken Dialog Systems. In Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, Ohio (2008)
3. Chung, G.: Developing a Flexible Spoken Dialog System Using Simulation. In Proceedings of ACL 04, Barcelona, Spain (2004)
4. Möller, S., Englert, R., Engelbrecht, K.-P., Hafner, V., Jameson, A., Oulasvirta, A. Raake, A., Reithinger, N.: MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations. In proceedings of InterSpeech-06, Pittsburgh (2006)
5. Mairesse, F., Walker, M.: PERSONAGE: Personality Generation for Dialogue. In Proceedings of the 45th Annual Meeting of the ACL, Prague (2007)
6. Pietquin, O.: User Simulation/User Modeling: State of the Art and Open Questions. CLASSiC Project Consortium Meeting, Issy Les Moulineaux (2008)
7. Schatzmann, J., Georgila, K., Young, S.: Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems. 6th SIGdial, Lisbon, Portugal (2005)
8. Steimel, B., Jameson, A., Jacobs, O., Pulke, S.: Voice Awards 2007: Die Besten deutschsprachigen Sprachapplikationen. Project Deliverables (2007)
9. Hermann, F., Niedermann, I., Peissner, M.: Development of Usability-Oriented User Taxonomy. Report on Procedure, Methods, and the Resulting User Taxonomy (2007)
10. Berenji, H.: Tutorial: Fuzzy Reinforcement Learning. IEEE International Conference on Fuzzy Systems, FUZZ-IEEE-07, Imperial College, London, UK (2007)
11. Eskenazi, M., Black, A., Raux, A., Langner, B.: Let's Go Lab: a platform for evaluation of spoken dialog systems with real world users. InterSpeech-08 proceedings (2008)