Usability Engineering for Spoken Dialogue Systems via Statistical User Models

Sebastian Möller[1], Robert Schleicher[1], Dmitry Butenkov[1],
Klaus-Peter Engelbrecht[1], Florian Gödde[1],
Tatjana Scheffler[2], Roland Roller[2], and Norbert Reithinger[2]

[1] Quality and Usability Lab, Deutsche Telekom Labs, TU Berlin
{sebastian.moeller,robert.schleicher,dmitry.butenkov,
klaus-peter.engelbrecht,florian.goedde}@telekom.de
[2] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI),
{tatjana.scheffler,roland.roller,norbert.reithinger}@dfki.de

Abstract. We describe ongoing work to integrate statistical user models in the usability engineering process of spoken dialogue systems. The idea is to generate user dialogue actions up to the spoken utterances in response to system utterances and directly feed them to the system under test. The underlying user models are derived semi-automatically from dialogue corpora. All simulated interactions are logged and a usability profile is derived from the log files, using HMMs to continuously estimate user judgments.

1 Introduction

The usability of many commercial spoken dialogue systems (SDSs) is still limited, as developers frequently do not have the time to perform user tests during the development cycle. This severely limits customer acceptance. To overcome this problem, we propose a software tool which helps to test SDSs with automatically generated user utterances. These artificial dialogues are logged and a usability profile is derived by estimating user quality judgments on the basis of the loggings.

Statistical models trained on data from real users have been used before to improve dialogue management. For example, [1] modeled user and system actions on the intentional level. Possible user actions and their probabilities were learned from real user data, and additional parameters describing the users' initiative or patience were used to manipulate action probabilities. [2] formalized the user model as a Bayesian Network and extended the status description by counting the number of times the system asked for an AVP. Other statistical approaches have been proposed by [3], [4], or [5]. There are also a few rule-based approaches which have been used for the purpose of evaluating dialog systems. E.g. [6] collected a corpus of possible utterances for expected user behavior, which is input to the automatic speech recognition (ASR) during the simulation.

The knowledge of user behavior alone, however, does not provide enough information for system adjustment. Although effectiveness (e.g. in terms of task success) and efficiency (e.g. in terms of the number of turns) can be estimated, there are other aspects which determine user satisfaction and finally acceptability. [7] introduced the PARADISE model to predict “user satisfaction”. An extensive evaluation of PARADISE and its application can be found in [8]. One of the main drawbacks is that they still require manual annotation of task success.

In the following, we describe a new integrated approach to statistically model user behavior and predict user satisfaction by letting a user simulation interact directly with the SDS under test. The project is called SpeechEval and funded by the European Regional Development Fund (ERDF).

2 SpeechEval User Simulation

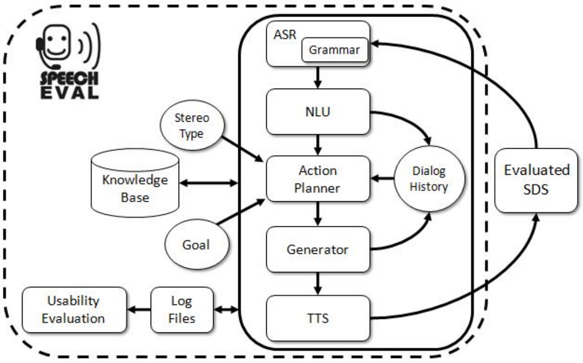

Fig. 1. Architecture of the SpeechEval user simulation.

SpeechEval's architecture largely follows a standard pipeline model (Fig. 1). The central knowledge bases for the user simulation are acquired semi-automatically from corpora. First, we extract domain grammars to be used in ASR and NLU, as well as as the basis of domain ontologies. Second, we derive utterance templates for understanding and for template-based generation. Finally, we learn dialog act classifiers based on the work by [9]. In addition, we can extract dialog models in the shape of automata from previous interactions with an SDS during run-time, which guide the user simulation in subsequent steps (for example, allowing for barge-ins).

In the action planner, the user model chooses a reply action based on the transition probabilities for similar dialog states observed in the corpus. Since some states have never been seen in the corpus, we choose a vector of features as the representation of each dialog state. These features include properties of the dialog history (such as the previous dialog act, the number of errors), the current user characteristics (expert vs. novice, for example), as well as other features such as the ASR confidence score. We estimate from the corpus the amount that each feature in the vector contributes to the choice of the next action. Thus, unseen states can be easily mapped onto the known state space as they lead to similar behavior as closely related seen states would.

The chosen next action is then enriched with content based on the goal and user characteristics. General heuristics are used to perform this operation of tying in the user simulation with the domain- and system-specific ontology. The utterance plan is then generated using a template-based approach.

All interactions are logged for analysis by the usability prediction algorithm, which is presented in the next section.

3 Usability Prediction

The usability prediction module aims at predicting the user satisfaction and to provide a detailed, continuous usability profile by modelling the user's current satisfaction judgment via a Hidden Markov Model (HMM). Each state of the model represents a possible judgment, e.g. in terms of a category number between 1 ("bad") and 5 ("excellent"). The course through the dialogue is modelled in terms of state transitions. During each transition, the state emits several dialogue features such as the current user speech act, the duration of each turn, the response delay, and the number and type of error messages produced by the system. These features are compared to the ones of the current dialogue, and the most probable path through the HMM is calculated using forward recursion. The most probable path being found, the predicted user judgment at the end of the dialogue, as well as the average user judgment during the dialogue, are taken as indicators of user satisfaction. In addition, effectiveness and efficiency estimates are calculated from the log files, determining task success and the average number and duration of the interaction. The usability profile is further amended by counting the number and types of meta-communication events (help requests, error messages, correction turns, etc.) which occurred during the dialogue.

The HMM has been trained on data which were obtained in a subjective interaction experiment with a prototypical system. During this experiment, users had to judge the dialogue up to the current state on a 5-point category scale, after each system turn. An experiment that provided first evidence for the performance of this approach is described in more detail in [10].

4 Current State and Future Work

SpeechEval has been implemented on the ODP platform [11], which makes use of commercial ASR + TTS modules (Nuance Recognizer 9.0 + SVOX 4.2) and allows us to implement the speech understanding, action planning and generation modules, as well as the different knowledge sources in a flexible way.

For training, we currently make use of three training corpora: Data from the Voice Award, a German contest of commercial SDSs which is carried out on a yearly basis with experts and naïve users (www.voiceaward.de); data from the BoRIS restaurant information system [12], and data from the INSPIRE smart-home system as a representative of a more complex, research dialogue system [13].

The verification of the approach is ongoing and performed using two types of metrics: First, we determine the degree of realism of the simulation by comparing characteristics of simulated dialogues to the real corpora. Second, we make use of the simulated corpora in order to identify usability problems. The rationale is that even a non-realistic simulation may be useful in the usability engineering process, e.g. by generating rare dialogues which point at important usability issues. In addition, we will focus on the performance of the usability prediction module and investigate how the current predicted user state can optimally be used for influencing the simulation behavior.

References

1. Eckert, W., Levin, E., Pieraccini, R.: User Modeling for Spoken Dialogue System Evaluation. In: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara CA (1997)
2. Pietquin O.: Machine Learning Methods for Spoken Dialog Simulation and Optimization. In: Mellouk, A., Chebira, A. (Eds.), Machine Learning, In-The (2009) 167-184
3. Scheffler, K., Young, S.: Corpus-based Dialogue Simulation for Automatic Strategy Learning and Evaluation. In: Proc of the NAACL-2001 Workshop on Adaptation in Dialogue Systems, Pittsburgh PA (2001)
4. Schatzmann, J., Thomson, B., Young, S.: Statistical User Simulation with a Hidden Agenda. In: Proc. of the 8th SIGDial Workshop on Discourse and Dialogue, Antwerp (2007)
5. Ai, H., Weng, F.: User Simulation as Testing for Spoken Dialog Systems. In: Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus OH (2008)
6. López-Cózar, R., de la Torre, A., Segura, J. C., Rubio, A. J.: Assessment of dialogue systems by means of a new simulation technique. Speech Communication 40 (2003), 387-407
7. Walker, M. A., Litman, D. J., Kamm, C. A., Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In: Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics, Madrid (1997)
8. Möller, S., Engelbrecht, K.-P., Schleicher, R.: Predicting the Quality and Usability of Spoken Dialogue Services, Speech Communication 50 (2008), 730-744
9. Germesin, S: Determining latency for on-line dialog act classification. In Proc. Machine Learning for Multimodal Interaction. (MLMI-08), September 8-10, Utrecht (2008)
10. Engelbrecht, K.-P., Hartard, F., Gödde, F., Möller, S.: A Closer Look at Quality Judgments of Spoken Dialog Systems. In: Proc. 10th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2009), Brighton (2009)
11. Pfalzgraf, A., Pfleger, N., Schehl, J., Steigner, J.: ODP: ontology-based dialogue platform, Technical report, SemVox GmbH, 2008
12. Möller, S.: Quality of Telephone-Based Spoken Dialogue Systems, Springer, New York NY (2005)
13. Möller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, A., Schuchardt, D., Fakotakis, N., Ganchev, T., Potamitis, I.: INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control. In: Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004), Lisbon (2004), 1603-1606