Pragmatic Usage of Linear Regression Models for the Prediction of User Judgments

Klaus-Peter Engelbrecht
Quality and Usability Lab
Deutsche Telekom Laboratories
Berlin University of Technology
Germany
Klaus-Peter.Engelbrecht@telekom.de

Sebastian Möller
Quality and Usability Lab
Deutsche Telekom Laboratories
Berlin University of Technology
Germany
Sebastian.Moeller@telekom.de

Abstract

Automatic evaluation of spoken dialog systems has gained interest among researchers in the past years. In the PARADISE framework (Walker et al. 1997), a linear regression function is trained on a dialog corpus to predict user ratings of satisfaction from interaction parameters. The accuracy of such predictions is generally measured with R^2, which usually is rather low. In this paper, it is shown that predictions according to PARADISE can lead to accurate test results despite the low R^2.

1 Introduction

Automatic usability testing of spoken dialog systems (SDSs) has gained interest among researchers in the past years. According to ISO 9241-11 (1998), usability of a system is compound of its effectiveness in doing typical tasks, the efficiency with which the task can be done and the satisfaction of the user with the system. Because user satisfaction is a subjective issue, usability testing involves humans who conduct typical tasks with the system and state their satisfaction with it afterwards. A key issue in automatic usability testing is the estimation of the expected user satisfaction without human involvement.

In the PARADISE framework (Walker et al. 1997) it is proposed to predict user satisfaction on the basis of interaction parameters captured in system log files. A linear regression (LR) model is trained on the parameters as predictors of user judgments of the corresponding dialogs as target. The percentage of the variance of the target that can be explained by the model is measured with R^2, which is based on the comparison of predicted and measured values for each dialog. When applying PARADISE, R^2 usually is below 0.6 for the prediction of the training data themselves (e.g. Walker et al. 2000), while the prediction of independent data is a stronger criterion and typically results in an even lower R^2.

Various steps have been taken to improve the predictive power of such equations. On the one hand, more and better predictor parameters have been searched for (Möller, 2005; Oulasvirta et al. 2006, Hastie et al. 2002), on the other hand, other prediction algorithms, e.g. classification and regression trees (CARTs) have been explored (Compagnoni 2006). However, R^2 values obtained remain unsatisfactory low.

While the standard accuracy measure for LR models, R^2, is based on a comparison of pairs of predicted and measured values for each dialog, in subjective measurement usually single ratings are not looked at. Instead, the researcher examines the overall distribution of all judgments for each questionnaire item. In fact, the very nature of subjective measurement involves joining ratings by multiple test subjects in order to minimize effects of inter-subject rating differences and by this maximizing the reproducibility of the findings. In other words, single ratings are tainted with different kinds of measurement errors (Annett, 2002). Consequently, an accurate prediction of single judgments is a Sisyphus task: it involves the difficult task to estimate the measurement errors, while at the same time the level of detail achieved by this is undesirable for the test result.

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 291–294,
Antwerp, September 2007.

In the best case, the detail lost in LR predictions would be congruent with the detail deliberately eliminated during test evaluation. If this was true, the pragmatic value of PARADISE models would be higher than the R^2 values suggest. This paper discusses the application of PARADISE predictions in a pragmatic context, in order to estimate the severity of loss of detail in PARADISE predictions for their practical application.

Corpora of two different experiments serving the evaluation of spoken dialog systems have been analyzed with respect to how well test results can be reproduced by predictions with LR equations. In the following section, the databases used will be described. In Section 3, the application of the PARADISE approach to the data is explained, and in Section 4 examples are given to illustrate how prediction results can be used to reconstruct specific test results.

2 Data

Experiment 1 has been carried out during the EU-funded INSPIRE project (IST 2001-32746). The SDS tested in the experiment is capable of controlling domestic devices such as lamps and a video recorder, leading a mixed-initiative dialog with the user. For the experiment, the speech recognition (ASR) was replaced by a Wizard-of-Oz, transcribing the users' utterances. The aim of the experiment was to test the impact of ASR accuracy on user satisfaction by adding different degrees of word substitutions, deletions and insertions to the wizard's transcription. 28 users took part in the experiment. Test participants were required to carry out three scenarios, each with 9-11 tasks and covering all devices which can be operated with the system. This results in 84 dialogs in this database. Further details can be found in (Möller et al. 2007).

In experiment 2, the BoRIS restaurant information system (Möller 2005; see this also for a detailed description of the experiment) was tested, which allows the user to search for a restaurant in Bochum, Germany, by specifying constraints for type of food, restaurant location etc. In the experiment, ten system configurations have been compared which differed with respect to the prompt quality (TTS or recorded natural language), the confirmation strategy (explicit or implicit) and the ASR performance, modeled in a similar way as in experiment 1. Each of the 40 participants did five telephone calls to the system, following instructive scenarios. 197 dialogs are available in this database.

Both experiments were executed in test labs. From the system log files, a vast number of interaction parameters was computed, including efficiency measures (such as dialog duration), qualitative measures (such as contextual appropriateness) and a classification of user errors (Oulasvirta et al. 2006). A complete list of the qualitative and efficiency measures can be found in (Möller 2005).

After each interaction, the participants filled out a questionnaire designed according to ITU-T Rec. P.851 (2003). The first rating of the questionnaire is on the systems overall quality (OQ), which was collected on a continuous scale with five equidistant and labeled points. The scale margins were extended to encourage the use of the full scale.

3 Prediction of subjective ratings

LR models were calculated with the interaction parameters as predictor variables and OQ as target variable. From the equations found, predictions of the respective ratings were made and compared to the true ratings. Two methods have been applied for the prediction: in the `useall` method, the whole database is used for training and prediction, while in the `leave-one-out` (`l1o`) method, successively each user is predicted from the function trained on the other users. While the `useall` method indicates how well the data can be described with such a function, the `l1o` method gives a more reliable estimation of the predictive power of the model.

Exp.	R^2 `useall`	R^2 `l1o`
1	0,580	0,202
2	0,466	0,235

Table 1. R^2 for predictions of Overall Quality ratings in exp. 1 and 2.

In both cases, in opposition to what is foreseen in the PARADISE approach, the variables have not been z-transformed before the training. In PARADISE, standardization of predictors and targets allows to read the importance of the predictors for the prediction from the coefficients of the equation, which, however, is not relevant for this study. Instead, the mean and STD values should be preserved here to allow an estimation of how well they can be predicted with the function obtained from the training.

Table 1 shows the R^2 values of the prediction models for the two databases. While the values are generally low, R^2 is considerably lower for the `l1o` predictions than for the `useall` predictions. The numbers reported here for the `useall` method lie in the range of those observed by other researchers for other systems, while Walker et al. (2000) achieved better results for tests on independent test data than those reported here for the `l1o` method.

4 Predicting test results

As stated above, we suspected that the R^2 values are not a good indicator of the usefulness of the predictions in a practical context. We therefore applied the same type of analysis to the predictions as has been applied to the real data in the studies the data stem from. In the following, four examples of this are given.

Exp. 1 aimed at detecting the level of ASR performance necessary for system acceptance by simulating four different target word accuracy (WA) rates (60, 73, 86, 100%). The means for each configuration were plotted and connected by straight lines. Then, the threshold of the positive user judgment was located.

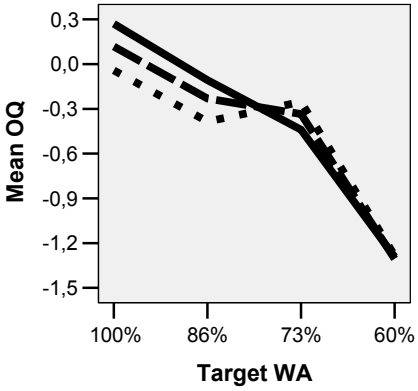

Figure 1. Overall Quality ratings for different WA in exp. 1; the solid line represents true ratings, the dashed line the `useall` predictions and the dotted one the `l1o` predictions.

Figure 1 shows how the results can be reproduced with data gained from predictions made with the LR equation. Displayed are the mean values of measured and predicted ratings for the four WA rates. While the predicted and measured means do not exactly agree with respect to the minimum WA leading to a positive judgment, the relation between WA and ratings is well reproduced by the prediction. The common conclusion that could be drawn from the predicted results as well as the true ratings would be that the WA should not fall below 73%, because from there on judgments decrease rapidly. Above 73%, the effect of WA is less drastic than below this value.

Similarly, results from experiment 2 can be predicted with the `l1o` and the `useall` method. In this experiment, again the users' judgment of the system for different target WA rates was tested. Figure 2 shows the means of measured and predicted values. Although the predicted values are slightly higher than the measured ones, the overall picture looks similar for the prediction and the actual measurement. The conclusion that can be drawn from both is the same: for recognition rates above 80 percent, an improvement of the target recognition rate is not reflected in improved ratings anymore, while for less than 80 percent, ratings drop to a lower range quickly.

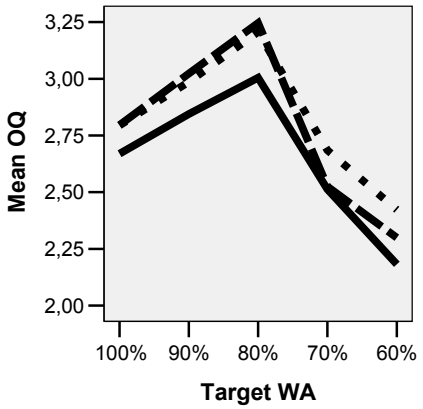

Figure 2. Overall Quality ratings for different target recognition rates in exp. 2; the solid line represents true ratings, the dashed line the `useall` predictions and the dotted one the `l1o` predictions.

In this experiment, also the impact of different system voices on the user judgment was tested. Figure 3 shows that both prediction methods reproduce the dramatic fall of ratings for the synthesized voice as compared to prerecorded human voices, however, the difference among the human speakers would not be detected with the prediction. Remarkably, `useall` and `l1o` method are comparably accurate despite the difference in their R^2*'s*.

Finally, the impact of the confirmation strategy on the judgments was tested with an ANOVA analysis (Table 2). While there is a bigger difference between the two confirmation strategies predicted with the LR equations than was actually

measured in the experiment, in all cases the difference is not significant (p>0.05, although F increases for the predicted values). Thus, the prediction leads to the same conclusion as the subjective ratings, namely that the confirmation strategy does not matter for the users' satisfaction with the BoRIS system.

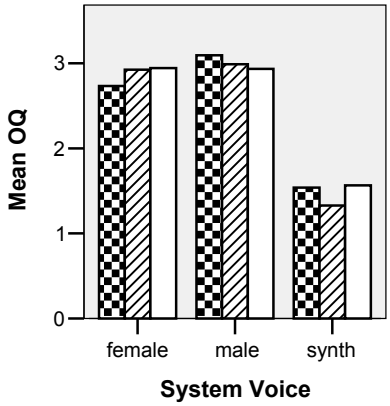

Figure 3. Overall Quality ratings for different Voices in exp. 2; the checkered bar represents true ratings, the shaded bar the `useall` predictions and the white one the `llo` predictions.

Confirmation Strategy (explicit/implicit)			
	measured	**`Useall`**	**`llo`**
F	(1,193) 0.02	(1,195) 2.88	(1,195) 3.10
p	0.89	0.09	0.08

Table 2. ANOVA results for explicit and implicit confirmation strategies. Differences in ratings for both strategies are not significant (p>0.05), neither in the measurement nor in the predictions.

5 Conclusion

In this paper, it was proposed to compare the mean values of predicted and true ratings rather than values for single dialogs. It was shown how LR models can be utilized for the automatic prediction of experimental results based on the observation of mean values. Although the predictions still lack some accuracy, the prediction models are more valuable in practical applications than their R^2 values suggest. In particular, the prediction of unseen data does not cause a dramatic drop of the model accuracy, as was indicated by the R^2 values. This is a particularly valuable finding since most applications intended for PARADISE involve the prediction of unseen data.

A further implication of the findings is that the improvement of usability prediction models on the basis of LR should not be based on changes in R^2 alone. While better methods for the evaluation of the models still have to be found, they might lead to significant progress in the models' development. This includes selection of appropriate modeling techniques (CARTs, Neural Networks etc.) and training methods for the algorithm, as well as the estimation of the usefulness of interaction or system parameters as usability predictors.

References

John Annett. 2002. Subjective Rating Scales. Science or Art? *Ergonomics*, 45(14):966-987

Bernardo Compagnoni. 2006. *Development of Prediction Models for the Quality of Spoken Dialog Systems*. Diploma thesis, Deutsche Telekom Laboratories/Institute for Telecommunications Technology, TU Braunschweig, Germany.

International Organization for Standardization. 1998. *ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*, Geneva, Switzerland.

Sebastian Möller, Paula Smeele, H. Boland, and Jan Krebber. 2007. Evaluating Spoken Dialog Systems According to De-facto Standards: A Case Study. *Computer Speech and Language* 21:26-53.

Sebastian Möller. 2005. *Quality of Telephone-based Spoken Dialog Systems*, Springer Science+Business Media, Inc., New York, NY.

Antti Oulasvirta, Klaus-Peter Engelbrecht, Anthony Jameson, and Sebastian Möller. 2006. The Relationship Between User Errors and Perceived Usability of a Spoken Dialog System. *ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, Germany:61-67.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability. *Natural Language Engineering* 6(3-4):363-377.

Marilyn Walker, Diane Litman, Candace Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, Madrid, Spain:271–280.

Helen Wright Hastie, Rashmi Prasad, and Marilyn Walker. 2002. Automatic Evaluation: Using a Dialogue Act Tagger for User Satisfaction and Task Completion Prediction. *Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC)*, 2:641–648, Las Palmas, Spain.