Modeling the User's Belief about the State of a Spoken Dialog System

Klaus-Peter Engelbrecht (klaus-peter.engelbrecht@telekom.de)
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany

Keywords: Automatic usability evaluation; mental model, belief state.

Introduction

Spoken dialog systems (SDSs) have to respond adequately in many different situations to a multitude of different, partly misrecognized user inputs. Thus, user simulation is a valuable means to test such systems during design time. Although the user models used for the simulation are often incomplete and not always accurate, the simulated data contain much of the information found in a user test (Engelbrecht, 2012). Thus, next to reducing the effort to adapt the models to new systems, an interesting research question is how to analyze large amounts of generated data efficiently. This paper contributes to two types of analysis, namely design error detection, and prediction of perceived system quality.

Both tasks can be improved by modeling the point-of-view of the user on the dialog. One aspect of this is what the user believes to be the current system state. A wrong belief may point to concrete interface problems. On the other hand, we may be interested in how the user perceives the dialog to progress. From such data, it may be possible to derive good predictors of user judgments.

This paper presents ongoing work on this topic. We do not use a general model of cognition, but rather model this specific aspect of cognition on a conceptual level. The model is used to annotate real user interactions with an estimate of the believed system state at each dialog exchange. From this, several parameters are derived and correlated with design problems annotated in the corpus and with judgments by the users.

Belief Model

The believed system state is structured in the same way as the real system state. It consists of a set of slots (or variables) for each type of input, e.g. price range or food type. These slots are filled with values provided in the user utterances. E.g., if the utterance "I'm looking for a cheap Italian restaurant" is observed, the system would add the value "cheap" to the *price* slot, and "italian" to the *food* slot. Later, these values are used in the database query to find a matching restaurant. Contrary to the system state, the *believed* system state is not updated based on the concepts mentioned by the user, but based on the system feedback.

Recent work circling around POMDP-based, self-learning SDSs has discussed how a system may track several concurring hypotheses about the previous user inputs in a probabilistic representation of the "believed" user tasks (e.g. Thompson et al., 2010). Although a probabilistic model would be more powerful, we use a deterministic, rule-based belief model. The reason is that users exhibit far more competencies than systems in extracting context information, which are difficult to model statistically. In addition, the parameterization and model structure are not as straightforwardly specified.

In order to exemplify the resolution level of the system, some example rules for the belief update are presented in Table 1. It can be noted that rules can refer to many, and completely different, aspects of the dialog history, which complicates the efficient probabilistic representation in a Bayesian network. In addition, processing these rules requires some annotations of the prompts, mainly with the confirmed concepts and explicit or implicit information they carry about the system state.

Table 1: Belief update rules (examples).

- Add concepts explicitly confirmed by the system.
- In case affirmation of the confirmation by the user is required, and the user does not affirm or the system asks for any of the confirmed values in the next exchange, remove all confirmed values.
- Empty slots queried by the system; however, if the system asks to repeat the value, filled the slot with an unknown value ("XXX").
- If the system provides no feedback, add all values of the previous user utterance, as long as the system continues consistently (e.g. not asking for one of the provided slots)

Use Case

In order to analyze how the belief model can support the analysis of experimental data, we use a database collected with the BoRIS restaurant information system (Möller, 2005). 40 Users (29m, 11f; M = 29.0y, SD=9.7) performed five different tasks. Three dialogs could not be used in the analysis, resulting in 197 dialogs (2001 exchanges) in the entire dataset.

Each dialog was judged on a SDS usability questionnaire. Factor analysis revealed a scale related to the overall acceptance of the system (for details, see Möller, Engelbrecht & Schleicher, 2008). In addition, log files are available, listing transcripts of each user and system turn along with speech understanding errors and task success annotations. Finally, a list of design problems was compiled and annotated at all dialog exchanges where they manifest in interaction problems.

Results

First, it is analyzed how well problematic exchanges can be predicted by the occurrence of mismatches between actual and believed system state. Intuitively, situations where the user has a wrong belief about the system state are problematic by themselves. However, we try to provide some quantification with respect to the problems annotated in the database. This is usually measured by *recall* and *precision*, where *recall* measures how many of the exchanges where a problem is annotated are also annotated with a wrong belief state. *Precision*, in turn, measures how many of the wrong belief state exchanges also have a problem annotation. We measure a *recall* of 0.50 and a *precision* of 0.66. In other words, checking the 893 exchanges where a wrong belief was annotated, half of the problematic situations are found, and 304 exchanges are analyzed in vain.

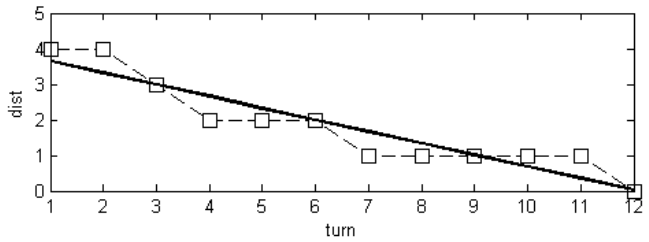

Figure 1: Distance between believed and desired system state through the 12 exchanges of an example dialog, and regression line through the points.

Next, features for the prediction of the user ratings were created. User judgments have previously been predicted from interaction data using trained classifiers (Walker, Litman, Kamm & Abella, 1997). A main problem remains to find good predictors generalizing across different systems. Analyzing the interactions from the user's perspective may be a key factor to achieve this.

First, the edit distance between the believed system state and the user goal can be determined as the number of unfilled slots plus twice the number of wrongly set slots. As illustrated in Figure 1, this distance can be specified for each exchange in a dialog. Via linear regression, the progress towards the goal can then be specified as the gradient of the regression line. Correlation analysis shows that the gradient is a fair predictor of system acceptance, compared to the standard predictors *dialog duration* and *task success* (Table 2). If we only consider the gradient over the last three exchanges, the correlation is even higher, which could be interpreted as *recency effect* (cf. Hassenzahl & Sandweg, 2004).

Table 2: Correlations of performance and judgments.

Performance	*r*	*p*
task success	0.26	0.00
dialog duration	0.31	0.00
gradient(dist.), all	0.26	0.00
gradient(dist.), last3	0.34	0.00
CER	0.00	0.96
perceived CER	-0.21	0.00

Furthermore, the perceived concept error rate (*CER*) can be calculated by comparing what the user said at each exchange to what she believes was understood. Table 2 shows that, contrary to the true *CER*, the *perceived CER* is significantly correlated with the judgments.

Conclusion

This paper showed, using an example SDS, that modeling the belief users have about the system state over the course of a dialog can provide valuable information for data analysis. Differences in the believed and desired system state (vaguely) hinted to system design errors. In the future, more qualified indicators may be derived from the belief annotations. Furthermore, new parameters for the prediction of user judgments were derived and showed correlations with the judgments in the range of *task success* and *dialog duration*. Subsequent research will show if the new parameters are independent from previous ones and thus useful as *additional* predictors.

Unfortunately, as many different parameters and complex relations between the dialog acts of user and system need to be exploited to update the believed system state, no sound probabilistic model could be presented at this stage. In addition, the generalization of the model to other SDSs has to be tested. Finally, other knowledge users collect about the system during a dialog could be tracked to analyze the data more comprehensively and run user simulations with the models. All this will be dealt with in future work.

References

Engelbrecht, K.-P. (2012). Estimating Spoken Dialog System Quality with User Models. TU Berlin.

Evanini, K., Hunter, P., Liscombe, J., Suendermann, D., Dayanidhi, K. & Pieraccini, R. (2008). Caller Experience: A Method for Evaluating Dialog Systems and its Automatic Prediction. *Proc. of SLT* (pp. 129-132). Goa, India.

Hassenzahl, M., Sandweg, N. (2004). From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments. *Proc. of CHI*, Vienna, Austria.

Hastie, H. W., Prasad, R., Walker, M. (2002). Automatic Evaluation: Using a Date Dialogue Act Tagger for User Satisfaction and Task Completion Prediction. *Proc. of LREC 2002* (pp. 641–648), Las Palmas de Gran Canaria.

Möller, S. (2005). Quality of Telephone-based Spoken Dialog Systems. New York, USA: Springer.

Möller, S., Engelbrecht, K.-P., Schleicher, R. (2008). Predicting the Quality and Usability of Spoken Dialogue Services. *Speech Communication*, 50, 730-744.

Thomson, B., Jurcicek, F., Gašić, M., Keizer, S., Mairesse, F., Yu, K., & Young, S. (2010). Parameter learning for POMDP spoken dialogue models. Proc. of SLT (pp. 271-276), Berkeley, California.

Walker, M., Litman, D., Kamm, C., Abella, A. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc. of ACL/EACL* (pp. 271–280), Madrid, Spain.