Automatic Learning and Evaluation of User-Centered Objective Functions for Dialogue System Optimisation

Verena Rieser and Oliver Lemon

School of Informatics, University of Edinburgh, UK
vrieser,olemon@inf.ed.ac.uk

Abstract

The ultimate goal when building dialogue systems is to satisfy the needs of real users, but quality assurance for dialogue strategies is a non-trivial problem. The applied evaluation metrics and resulting design principles are often obscure, emerge by trial-and-error, and are highly context dependent. This paper introduces data-driven methods for obtaining reliable objective functions for system design. In particular, we test whether an objective function obtained from Wizard-of-Oz (WOZ) data is a valid estimate of real users' preferences. We test this in a test-retest comparison between the model obtained from the WOZ study and the models obtained when testing with real users. We can show that, despite a low fit to the initial data, the objective function obtained from WOZ data makes accurate predictions for automatic dialogue evaluation, and, when automatically optimising a policy using these predictions, the improvement over a strategy simply mimicking the data becomes clear from an error analysis.

1. Introduction

The ultimate goal when building dialogue systems is to satisfy the needs of real users, but quality assurance for dialogue strategies is a non-trivial problem. In conventional dialogue design the dialogue often is designed following 'best practises' which are often obscure and emerge by trial-and-error (Paek, 2007). In addition, user preferences are highly context dependent (Hu et al., 2007). This is why dialogue strategy design is often referred to as being more of an art than a science (Jones and Galliers, 1996; Pieraccini, 2002)

Over recent years, data-driven statistical optimisation methods (e.g. Reinforcement Learning (RL)) for dialogue strategy design have become more and more popular (Lemon and Pietquin, 2007). One major advantage of RL-based dialogue strategy development is that the dialogue strategy can be automatically trained and evaluated using the same objective function (Walker, 2005). In the context of RL the objective function is also called the "reward" (Sutton and Barto, 1998). Despite its central aspect for RL, quality assurance for objective functions has received little attention so far. In fact, the reward function is one of the most hand-coded aspects in RL (Paek, 2006).

In this paper we propose a new method for meta-evaluation of the objective function. We bring together two strands of research: one strand uses Reinforcement Learning to automatically optimise dialogue strategies, e.g. (Singh et al., 2002), (Henderson et al., 2008), (Rieser and Lemon, 2008a; Rieser and Lemon, 2008b); the other other focuses on automatic evaluation of dialogue strategies, e.g. the PARADISE framework (Walker et al., 1997), and meta-evaluation of dialogue metrics, e.g. (Engelbrecht and Möller, 2007; Paek, 2007). Clearly, automatic optimisation and evaluation of dialogue policies, as well as quality control of the objective function, are closely inter-related problems: how can we make sure that we optimise a system according to real users' preferences?

In particular, we construct a data-driven objective function using the PARADISE framework, and use it for automatic dialogue strategy optimisation following pioneering work by (Walker et al., 1998). However, it is not clear how reliable such a predictive model is, i.e. if it indeed estimates real user preferences. The models obtained with PARADISE usually fit the data poorly (Engelbrecht and Möller, 2007). It is also not clear how general they are across different systems and user groups (Walker et al., 2000), (Paek, 2007). Furthermore, it is not clear how they perform when being used for automatic strategy optimisation within the RL framework.

In the following we evaluate different aspects of an objective function obtained from Wizard-of-Oz (WOZ) data (Rieser and Lemon, 2008b). We proceed as follows: The next Section shortly summarises the overall dialogue system design. In Section 3. we test the model stability in a test-retest comparison across different user populations and data sets. In Section 4. we measure prediction accuracy. In Section 5. we conduct a detailed error analysis where we test the relationship between improved user ratings and dialogue behaviour, i.e. we investigate which factors lead the users to give higher scores, and whether this was correctly reflected in the original objective function.

2. Overall framework

2.1. Dialogue System Design

Our application domains are multimodal information seeking dialogue systems as an interface to an in-car MP3 player. The structure of information seeking dialogues consists of an information acquisition dialogue and an information presentation sub-dialogue (see Figure 1).

For information acquisition the task of the dialogue policy is to gather 'enough' search constraints from the user, and then, 'at the right time', to start the information presentation phase where the task is to present 'the right amount' of information – either on the screen or listing the items verbally. What this actually means depends on the dialogue context and the preferences of our users as reflected in the objective function. We therefore formulate dialogue learning as a hierarchical optimisation problem (Rieser and Lemon, 2008b). The applied objective function follows this structure as well.

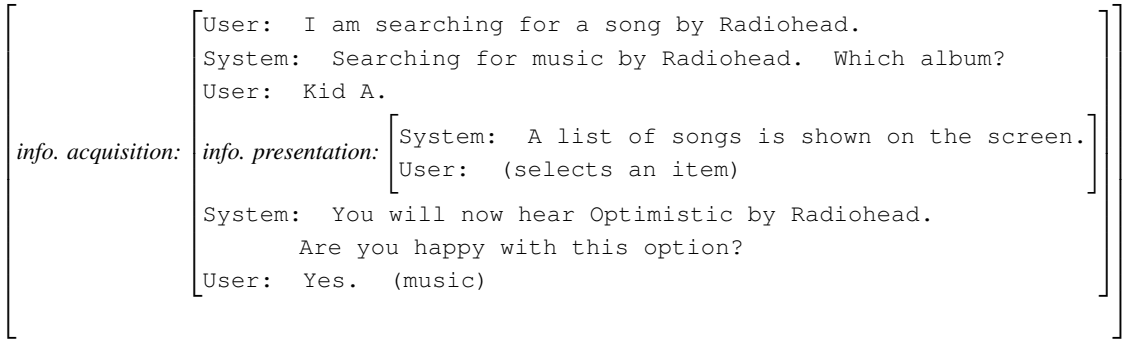


Figure 1: Hierarchical dialogue structure for information seeking multimodal systems.

2.2. Method

In the following the overall method is shortly summarised. Please see (Rieser and Lemon, 2008b; Rieser, 2008) for details.

1. We obtain an objective function from the WOZ data of (Rieser et al., 2005) according to the PARADISE framework. In PARADISE multivariate linear regression is applied to experimental dialogue data in order to develop predictive models of user preferences (obtained from questionnaires) as a linear weighted function of dialogue performance measures (such as dialogue length). This predictive model is used to automatically evaluate dialogues. For RL this function is used as the "reward" for training.

2. We train an RL-based dialogue system with the obtained model. The hypothesis is that, by using the obtained quality measures as a reward function for RL, we will be able to learn an improved strategy over a policy which simply mimics observed patterns (i.e. the human wizard behaviour) in the data. The baseline policy is therefore constructed using Supervised Learning (SL) on the WOZ data. We then test both strategies (SL and RL) with real users using the same objective/evaluation function.

3. Since the objective function plays such a central role in automatic dialogue design, we need to find methods that ensure its quality. In this paper, we evaluate the obtained function in a test-retest comparison between the model obtained from the WOZ study and the one obtained when testing the real system as described in the following.

3. Model Stability

For the information acquisition phase we applied stepwise multivariate linear regression to select the dialogue features which are most predictive for perceived Task Ease. Task Ease is a measure from the user questionnaires obtained by taking the average of two user ratings on a 5-point Likert scale.

1. The task was easy to solve.

2. I had no problems finding the information I wanted.

We choose Task Ease as the ultimate measure to be optimised following (Clark, 1996)'s *principle of the least effort* which says: "All things being equal, agents try to minimize their effort in doing what they intend to do".

The PARADISE regression model is constructed from 3 different corpora: the SAMMIE WOZ experiment (Rieser et al., 2005), and the *iTalk* system used for the user tests (Rieser and Lemon, 2008b) running the supervised baseline policy and the RL-based policy. By replicating the regression model on different data sets we test whether the automatic estimate of Task Ease generalises beyond the conditions and assumptions of a particular experimental design. The resulting models are shown in Equations 1-3 , where $TaskEase_{WOZ}$ is the regression model obtained from the WOZ data, $TaskEase_{SL}$ is obtained from the user test data running the supervised policy, and $TaskEase_{RL}$ is obtained from the user test data running the RL-based policy. They all reflect the same trends: longer dialogues (measured in turns) predict a lower Task Ease, whereas a good performance in the multimodal information presentation phase (multimodal score) will positively influence Task Ease. For the *iTalk* user tests almost all the tasks were completed; therefore task completion was only chosen to be a predictive factor for the WOZ model.

$$TaskEase_{WOZ} = 1.58 + .12 * taskCompl + .09 * mmScore - .20 * dialogueLength \quad (1)$$

$$TaskEase_{SL} = 3.50 + .54 * mmScore - .34 * dialogueLength; \quad (2)$$

$$TaskEase_{RL} = 3.80 + .49 * mmScore - .36 * dialogueLength; \quad (3)$$

To evaluate the obtained regression models we use two measures: how well they fit the data (goodness-of-fit) and how close the functions are to each other (model replicability). For the WOZ model the data fit was rather low ($R^2_{WOZ} = .03$), [1] whereas for the models obtained from the *iTalk* system the fit has improved ($R^2_{RL} = .48$, and $R^2_{SL} = .55$).

To directly compare the functions we plotted them in 3D space (the 4th dimension for $TaskEase_{WOZ}$ was omitted), see Figure 2. While the models obtained with the *iTalk* system show almost perfect overlap ($R^2 = .98$), the (reduced) WOZ model differs ($R^2 = .22$) in the sense that it assigns

[1] for R^2 we use the adjusted values.

less weight to dialogue length and the multimodal presentation score.

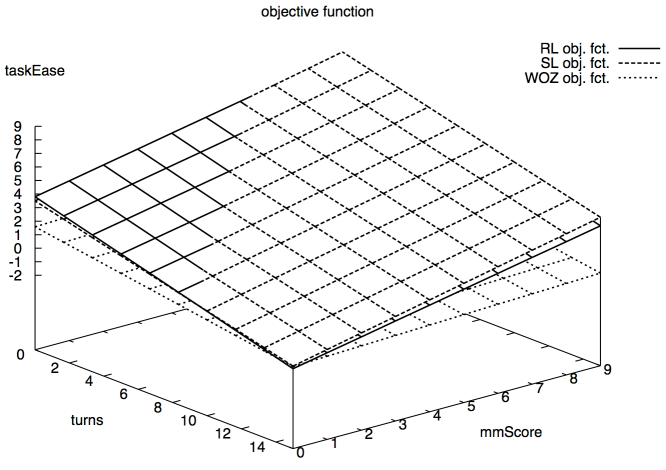


Figure 2: 3D Visualisation of the objective functions obtained from WOZ data and real user data using a SL and RL-based strategy.

4. Model Performance: Prediction Accuracy

We now investigate how well these models generalise by testing their prediction accuracy. Previous research evaluated two aspects: how well a given objective function is able to predict unseen events from the original system (Engelbrecht and Möller, 2007), and how well it is able to predict unseen events of a new/different system (Walker et al., 2000). We evaluate these two aspects as well, the only difference is that we use the Root Mean Standard Error (RMSE) instead of R^2 for measuring the models prediction accuracy. RMSE is (as we argue) more robust for small data sets. In particular, we argue that, by correcting for variance, R^2 can lead to artificially good results when using small tests sets (which typically vary more) and is sensitive to outliers (see Equation 4). RMSE instead measures the (root) mean difference between actual and predicted values (see Equation 5).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y - \overline{y})} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \quad (5)$$

First, we measure the predictive power of our models within the same data set using 10-fold cross validation, and across the different systems by testing models trained on one system to predict perceived Task Ease for another system, following a method introduced by (Walker et al., 2000).

The results for comparing the RMSE (max.7 for SL/RL, and max.5 for WOZ) for training and testing within data sets (ID 1-3) and across data sets (ID 4,5) are shown in Table 1. In order to present results from different scales we also report the percentage of the RMSE of the maximum error (% error). The results show that predictions according to PARADISE can lead to accurate test results despite the low data fit. While for the regression model obtained from the WOZ data the fit was 10-times lower than for SL/RL, the prediction performance is comparably good (see Table 1, ID 1–3). The models also generalise well across systems (see Table 1, ID 4–5).

ID	train	test	RMSE	% error
1	WOZ SAMMIE	WOZ SAMMIE	0.82	16.42
2	SL *iTalk*	SL *iTalk*	1.27	18.14
3	RL *iTalk*	RL *iTalk*	1.06	15.14
4	RL *iTalk*	SL *iTalk*	1.23	17.57
5	SL *iTalk*	RL *iTalk*	1.03	14.71

Table 1: Prediction accuracy for models within (1-3) and across data sets (4,5).

In addition, we evaluate model accuracy following a method introduced by (Engelbrecht and Möller, 2007). They suggest to compare model performance by plotting mean values for predicted and true ratings by averaging over conditions. We replicate this method, averaging mean ratings for observed and predicted Task Ease over number of turns. The resulting graphs in Table 2 show that the predicted mean values per turn are fairly accurate for the SL and RL objective functions (first two graphs from the left). For the WOZ data, the predictions are less accurate especially for low numbers of turns (graph on the right). This is due to the fact that for low numbers of turns only very few observations are in the training set: 25% of the dialogues are between 5 and 6 turns long (where the predictions are close to the observations) and 42% of dialogue are over 14 turns long (where the curves converge again). Only 33% covers the span between 7-13 turns, where the graphical comparison indicates low prediction performance. However, these results are misleading for small data sets (as we argue). Quite the contrary is the case: the predicted values show that the linear model does well for the majority of the cases and is not sensitive to outliers, i.e. the graph only diverges if there are too little obeservations. It therefore generalises well.

5. Error Analysis

In previous work we showed that the RL-based policy significantly outperforms the supervised policy in terms of improved user ratings and dialogue performance measures (Rieser and Lemon, 2008b). Here, we test the relationship between improved user ratings and dialogue behaviour, i.e. we investigate which factors lead the users to give higher scores, and whether this was correctly reflected in the original reward function.

We concentrate on the information presentation phase, since there is a simple two-way relationship between user scores and the number of presented items. To estimate this relationship we use curve fitting, which is used as an alternative model to linear regression in cases where the relationship between two variables can also be non-linear. For each presentation mode (verbal vs. multimodal) we select the (simplest) model with the closest fit to the data (R^2).

5.1. Training

We first use this method to construct the reward function for policy learning from the WOZ data. Figure 3 shows

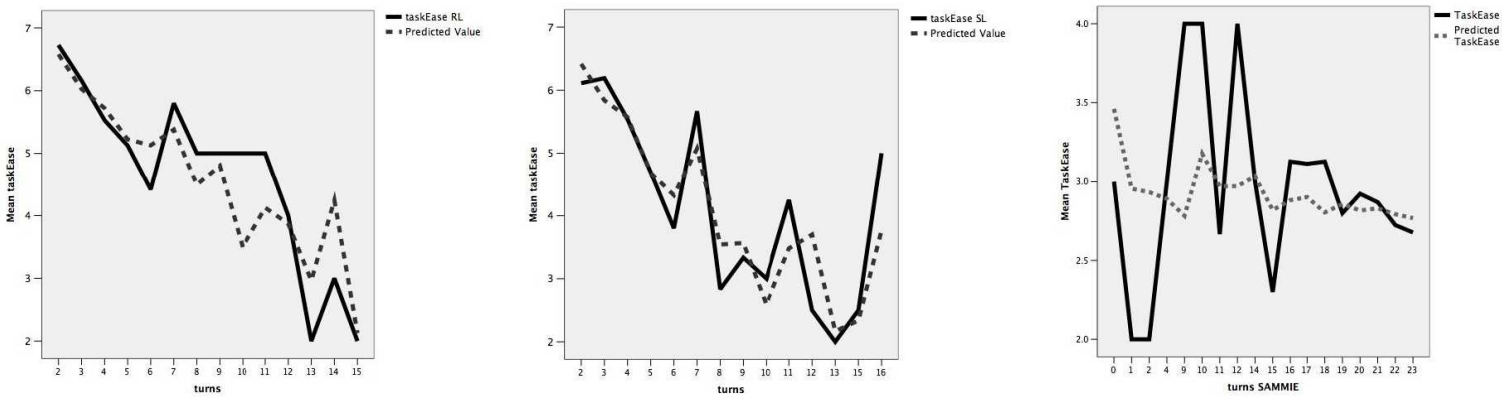


Table 2: Average Task Ease ratings for dialogues of different length (in turns); the solid lines are the true ratings and the dashed line the predicted values.

the employed reward function for information presentation modelled from the WOZ data. The straight line presents the objective function for verbal presentation and the quadratic curve the one for multimodal presentation.

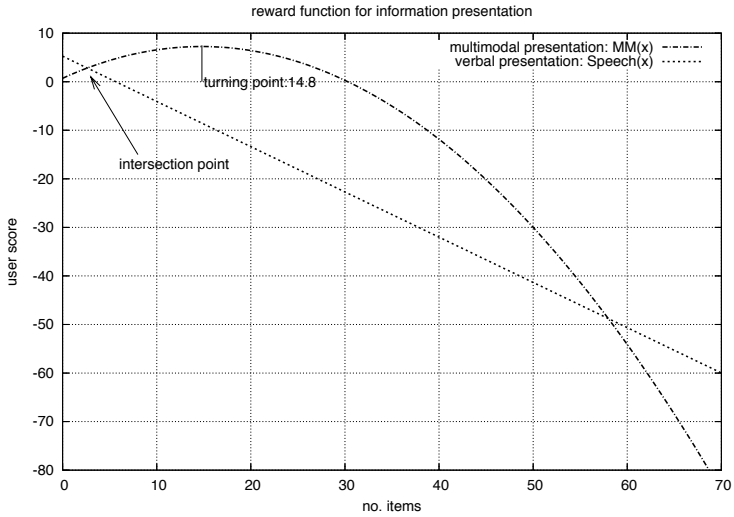


Figure 3: WOZ objective function for the information presentation phase

In the WOZ experiments wizards never presented more than 3 items using speech, resulting in a linearly decreasing line. This fact was captured by the learning schemes in different ways. SL extracted the rule "never present more than 3 items using speech". For RL the extrapolated line assigns negative values to more than 5 verbally presented items and intersects with the multimodal reward at 2.62, i.e. for more than 3 items the returned reward is higher when presenting multimodally. Therefore the RL-based strategy learns to present up to 3 items verbally (on average not more than 2.4 items per dialogue).

5.2. Testing

We now apply the same curve-fitting method on the *iTalk* user test data in order to test whether the policy optimisation had been successful. We therefore compare the curve fitting model obtained from the system running the RL policy against the model obtained from the SL policy. The hypothesis is that if the policy is good (i.e. consistently making the right decisions), this will result in equally high scores for all presented items, represented by a straight line; whereas if the curve is not linear, this indicates that the policy was sometimes making the right decision and sometimes not.

The estimated relationship between the average number of items presented verbally and the verbal presentation score from the user questionnaire is shown in the left column of Table 3. The straight, slightly declining line indicates that the policies in general make the right decision, although the fewer items they present the better. For verbal presentation both learning schemes (RL and SL) were able to learn a policy from the WOZ data which received consistently good ratings from the users (between 6–5 for RL, and 5–4 for SL on a 7-point Likert scale).

For multimodal presentation the WOZ objective function has a turning point at 14.8 (see Figure 3). The RL-based policy learned to maximise the returned reward by displaying no more than 15 items. The SL policy, in contrast, did not learn an upper boundary for when to show items on the screen (since the wizards did not follow a specific pattern, (Rieser and Lemon, 2008b)). When relating number of items to user scores, the RL policy produces a linear (slightly declining) line between 7 and 6 (Table 3, bottom right), indicating that the applied policy reflected the users' preferences. Hence, we conclude that the objective function derived from the WOZ data gave the right feedback to the learner.

For the SL policy the Logarithmic function best describes the data.It function indicates that the multimodal presentation strategy received the highest scores if the number of items presented were just under 15 (Table 3, top right), which is the turning point of the WOZ objective function. This again indicates that, for the *iTalk* users the preferred multimodal policy was indeed the one reflected in the WOZ objective function.

6. Conclusion

This paper introduces data-driven methods for obtaining reliable objective functions for dialogue system design, and so steers dialogue design towards science rather than art. We applied data-driven methods to build objective func-

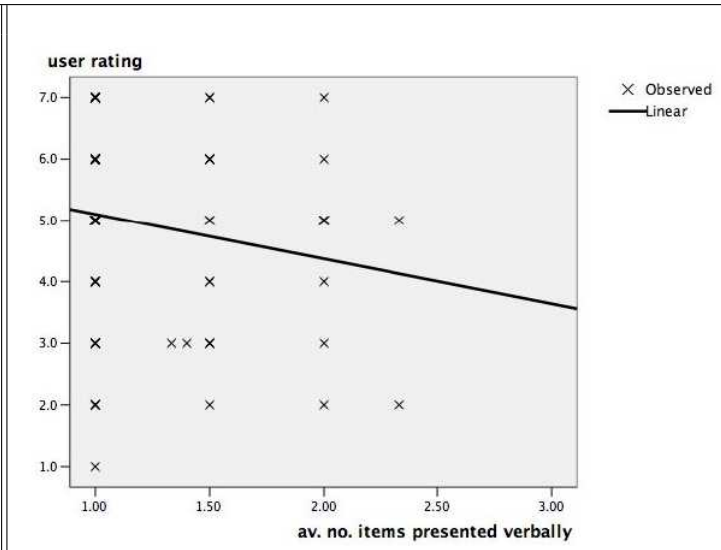
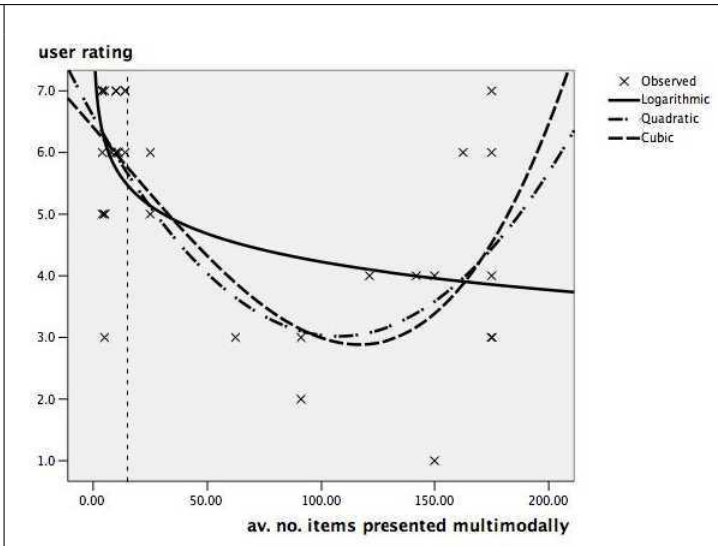
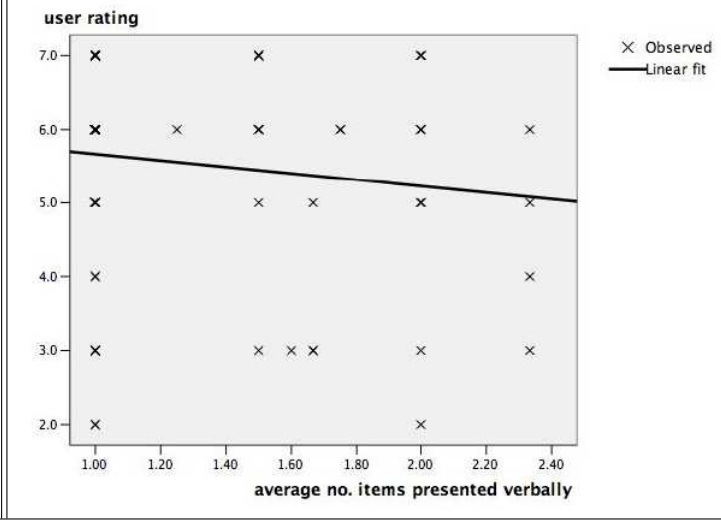
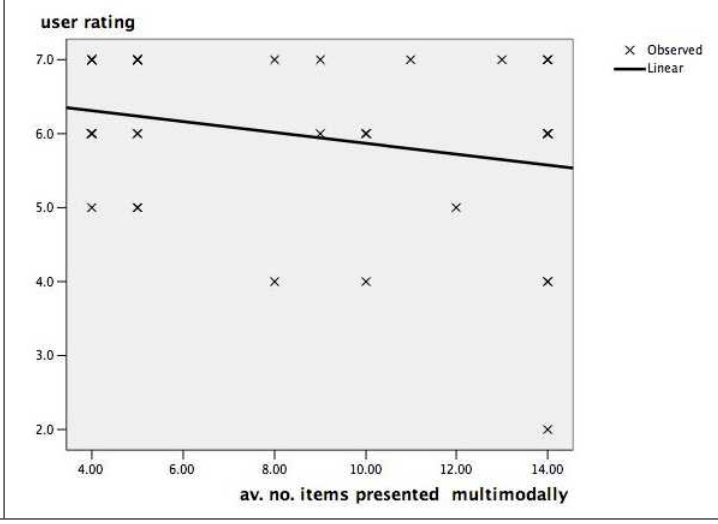
corpus	verbal	multimodal
SL		
RL		

Table 3: Objective functions for information presentation

tions (for both dialogue policy learning and evaluation) reflecting the needs of real users. In particular, we derived a non-linear objective function from Wizard-of-Oz data which is used to automatically train a Reinforcement Learning-based dialogue strategy, which was then evaluated with real users.

To ensure the quality of the applied objective function we evaluated its stability, predictive power, and strategy improvements in a test-retest comparison. We also conduct a detailed error analysis.

In sum, according to our measures, an objective function obtained from WOZ data is a valid first estimate of real users' preferences. Despite a low fit to the initial data, the objective function obtained from WOZ data makes accurate predictions for automatic dialogue evaluation, and, when automatically optimising a policy using these predictions, the improvement over a strategy just mimicking the data becomes clear from an error analysis. The models obtained from the tests with a real system follow the same trends, but can be seen as more reliable estimates of the objective function in this domain. In future work we will explore incrementally training a system according to improved representations of real user preferences, for example gathered online from a deployed spoken dialogue system.

This work also introduces non-linear objective functions for dialogue optimization, which merit further exploration in future work.

Acknowledgements

This work was partially funded by the International Research Training Group Language Technology and Cognitive Systems, Saarland University, and by EPSRC project number EP/E019501/1. The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 216594 (CLASSIC project: www.classic-project.org)

7. References

Herbert Clark. 1996. *Using Language*. Cambridge University Press.

Klaus-Peter Engelbrecht and Sebastian Möller. 2007. Pragmatic usage of linear regression models for the predictions of user judegments. In *Proc. of SIGdial*.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. Hybrid reinforcement / supervised learning of dialogue policies from fixed datasets. *Computational Linguistics (to appear)*.

Jiang Hu, Andi Winterboer, Clifford Nass, Johanna D. Moore, and Rebecca Illowsky. 2007. Context & usability testing: user-modeled information presentation in easy and difficult driving condition. In *Proc. CHI*, pages 1343–1346.

Karen Sparck Jones and Julia Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag.

Oliver Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. In *Proc. of Interspeech*.

Tim Paek. 2006. Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In *Proc. Dialog-on-Dialog Workshop, Interspeech*.

Tim Paek. 2007. Toward evaluation that leads to best practices: Reconciling dialogue evaluation in research and industry. In *Proc. of the NAACL-HLT Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*.

Roberto Pieraccini. 2002. The art and science of spoken dialog systems. Invited talk, The Center for Language and Speech Processing, Johns Hopkins University.

Verena Rieser and Oliver Lemon. 2008a. Does this list contain what you were searching for? Learning adaptive dialogue strategies for interactive question answering. *Natural Language Engineering (to appear)*.

Verena Rieser and Oliver Lemon. 2008b. Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL/HLT-08)*.

Verena Rieser, Ivana Kruijff-Korbayová, and Oliver Lemon. 2005. A corpus collection and annotation framework for learning multimodal clarification strategies. In *Proc. SIGdial*.

Verena Rieser. 2008. *Bootstrapping Reinforcement Learning-based Dialogue Strategies from Wizard-of-Oz data (to appear)*. Ph.D. thesis, International Research Training Group Language Technology and Cognitive Systems, Saarland University.

Satinder Singh, Diane Litman, Micheal Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *JAIR*, 16.

Richard Sutton and Andrew Barto. 1998. *Reinforcement Learning*. MIT Press.

Marilyn Walker, Diane Litman, Candance Kamm, and Alicia Abella. 1997. PARADISE: a general framework for evaluating spoken dialogue agents. In *ACL/EACL*.

Marilyn Walker, Jeanne Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proceedings of ACL/COLING*.

Marilyn Walker, Candance Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3).

Marilyn Walker. 2005. Can we talk? Methods for evaluation and training of spoken dialogue system. *Language Resources and Evaluation*, 39(1):65–75.