On the Applicability of a User Satisfaction-based Reward for Dialogue Policy Learning

Stefan Ultes, Juliana Miehle, and Wolfgang Minker

Abstract Finding a good dialogue policy using reinforcement learning usually relies on objective criteria for modelling the reward signal, e.g., task success. In this contribution, we propose to use user satisfaction instead represented by the metric Interaction Quality (IQ). Comparing the user satisfaction-based reward to the baseline of task success, we show that IQ is a real alternative for reward modelling: designing a reward function using IQ may result in a similar or even better performance than using task success. This is demonstrated in a user simulator evaluation using a live IQ estimation module.

1 Introduction and Related Work

Finding well-performing dialogue strategies for Spoken Dialogue Systems (SDSs) has been in the focus of research for many years. One possibility is to create handcrafted rules designed by experts. However, this approach is problematic: the resulting strategy is usually not very robust and strongly biased by the expert's view.

Instead of handcrafting the dialogue behaviour, more recent approaches aim at learning the optimal dialogue behaviour using reinforcement learning [6, 23]. Here, the dialogue strategy (called policy) is trained with a number of sample dialogues which are evaluated using a reward R. Traditional approaches incorporate the objective task success (TS) into the reward. This task success, though, only models the system view on the interaction. Incorporating the user view instead might be of

Stefan Ultes
University of Cambridge, Engineering Department, Trumpington Street, Cambridge, UK
e-mail: su259@cam.ac.uk
Research carried out while working at Ulm University

Juliana Miehle, Wolfgang Minker
Ulm University, Albert-Einstein-Allee 43, Ulm, Germany
e-mail: firstname.lastname@uni-ulm.de

more interest [14, 19] as it is the user who ultimately decides whether to use the system again or not.

For task-oriented dialogue systems, the user view may be captured with the user satisfaction. In fact, task success has only been used in previous work as it has been found to correlate well with user satisfaction [22]. While previously, user satisfaction was not accessible during learning, the recently proposed Interaction Quality (IQ) metric [10] has been designed to automatically predict the user's satisfaction level during the dialogue.

The main contribution of this work is to address the question how user satisfaction and task success relate with respect to their applicability on dialogue policy learning when being used as the main reward. We formulate the following expectations on their behaviour:

1. A policy which is optimised on user satisfaction should yield similar performance in task success compared to a policy optimised on task success. We assume that a user is not satisfied with the dialogue if the dialogue was not successful.
2. A policy which is optimised on task success will result in worse satisfaction rates than a policy which is optimised on user satisfaction.

Hence, for a satisfaction metric to be applicable for dialogue learning, it must show similar behaviour. In our previous work, we successfully showed that IQ is suitable for designing a rule-based policy and outlined an IQ-based dialogue-level reward function [15] and showed that using IQ for a binary decision over TS results in a precision of 84.5%. This indicates that similar performance in task success rate may be expected when using it as the main reward.

To follow up on this, we present a study in a simulated environment comparing IQ and TS as the main rewards for dialogue learning. We investigate the effects on the resulting TS rate of using IQ as the main reward and vice versa. Only if the expected behaviour is met, applying IQ for dialogue learning may be regarded as feasible.

Others have previously introduced user ratings into the reward. Gašić et al. [4] have successfully used the user's success rating directly during learning with real humans. Su et al. [12] extended the idea by using a similar setup plus having an additional task success estimator. While both use task success as measure, we will investigate whether user satisfaction may be used as reward in a similar fashion.

A prominent way to model user satisfaction is the PARADISE framework [21] which has also been used for reward modelling [20, 8, e.g.]. However, a questionnaire has to be answered by real users to derive user ratings with that framework. This is usually not feasible in real world setting. To overcome this, PARADISE has been used in conjunction with expert annotators [2, 3] which allows an unintrusive rating. However, the problem of mapping the results of the questionnaire to a scalar reward value still exists. Furthermore, PARADISE assumes a linear dependency between measurable parameters and user satisfaction. However, assuming a non-linear dependency might be more appropriate [9]. Therefore, we will use the Interaction Quality [10] in this work which uses scalar values applied by experts and assumes a non-linear dependency between measurable parameters and the target value.

The remainder of the paper is organized as follows: the IQ metric and the core contribution of modelling dialogue-level reward functions with IQ is described in Section 2 with its experiments and results in Section 3. Conclusions for future work are drawn in Section 4.

2 Interaction Quality Dialogue-level Reward

The Interaction Quality (IQ) [10], which will be used for modelling the dialogue-level reward, has been proposed as a turn-level metric for user satisfaction in spoken dialogue systems.

Interaction Quality Paradigm The general idea of the IQ paradigm is to use a set of measurable interaction parameters to create a non-linear statistical classification model. The target variable is a scalar value ranging from five (=satisfied) down to one (=completely unsatisfied). The input variables called interaction parameters encode information about the current turn as well as temporal information which is modelled on the window and the dialogue level (counts, means, sums and rates of turn-level parameters). The turn-level parameters are derived from the SDS modules Speech Recognition, Language Understanding, and Dialogue Management.

IQ meets the requirements for being used in an adaptive dialogue framework [19] and is the ideal candidate for our research. The IQ values are annotated by expert annotators yielding a high correlation to real user ratings [18].

Previously, a Support Vector Machine (SVM) has been applied for IQ classification achieving an unweighted average recall[1] (UAR) of 0.59 [9]. Furthermore, an ordinal regression approach achieved 0.55 UAR [1] and a hybrid-HMM 0.51 UAR [16]. As comparison, the human performance on that task is 0.69 UAR [10].

Reward Model In this work, we compare two different approaches for modelling the reward. Both have the same shape: for each turn, a small negative discount is added to favour shorter dialogues. Additionally, a final reward is defined based on the dialogue outcome. Both are combined for calculating the overall reward R for a complete dialogue of length T. For the baseline of using task success (TS), R is defined as

$$R_{TS} = T \cdot (-1) + \mathbb{1}_{TS} \cdot 20 \,,$$

where $\mathbb{1}_{TS} = 1$ only if the dialogue resulted in a successful task, $\mathbb{1}_{TS} = 0$ otherwise.

Based on the same binary decision principle, the IQ-based reward function has been defined:

$$R_{IQ_b} = T \cdot (-1) + \mathbb{1}_{IQ} \cdot 20 \,.$$

A final reward of 20 is assigned only if a high IQ has been achieved at the end of a dialogue, i.e., the final IQ value ($\mathbb{1}_{IQ} = 1$ only if $IQ \geq 4$, otherwise $\mathbb{1}_{IQ} = 0$).

[1] The arithmetic average over all class-wise recalls.

	TSR	ADL	AIQ
IQ_b	56.3% (±11.8)	13.3 (±0.9)	2.0 (±0.5)
TS	45.4% (±9.5)	14.2 (±1.4)	1.5 (±0.5)

Table 1 Final TSR, ADL, and AIQ computed out of the last 200 training dialogues for each reward function averaged over five policy trainings along with the respective confidence interval.

3 Experiments and Results

Evaluation of the reward functions presented in Section 2 is performed using an adaptive dialogue system interacting with a user simulator. A user simulator offers an easy and cost-effective way for training and evaluating dialogue policies and getting a basic impression about their performance.

Experimental Setup For creating the IQ estimation model as well as for training and evaluation of the dialogues, the Let's Go bus information domain [7] has been chosen as it represents a domain of suitable complexity. The Let's Go User Simulator (LGUS) [5] is used for policy training and evaluation to neutralise the need for human evaluators. LGUS has been trained on 1,275 real user dialogues collected with Let's Go. The simulator is set to converse for at least 5 turns as this is the minimum number to successfully complete the dialogue. To get bus information from the system, departure place, arrival place, travel time, and optionally the bus number may be provided.

In order to evaluate the reward functions, the adaptive statistical dialogue manager OwlSpeak [17] is used with a connected IQ estimation module [13]. The IQ estimation module uses a Support Vector Machine UAR of 0.55 on the training data [11] using 10-fold cross-validation. The policy of OwlSpeak operates on the summary level, i.e., it maps a summary space representation to a summary action, e.g., `request` or `confirm`. The summary action is mapped back to a system action using a heuristic.

For evaluation, the objective metrics task success rate (TSR, the ratio of dialogues for which the system was able to provide the correct result) and average dialogue length (ADL) have been chosen. In addition, the average IQ value (AIQ) is calculated for each policy based on the IQ value of the last exchange of each dialogue (which is also used within some of the reward functions). All AIQ values are based on the SVM estimates.

Results For each reward model, the results depicted in Table 1 are computed after 1,000 training dialogues based on the last 200 dialogues averaged over five trails. The corresponding learning curves for moving TSR and IQ are shown in Figure 1. Both show that for the respective reward, learning has mostly saturated for the last 200 dialogues.

Looking at the relation between IQ and TSR for the two reward models, both expectations presented in Section 1 are clearly met: R_{IQ} results in similar success

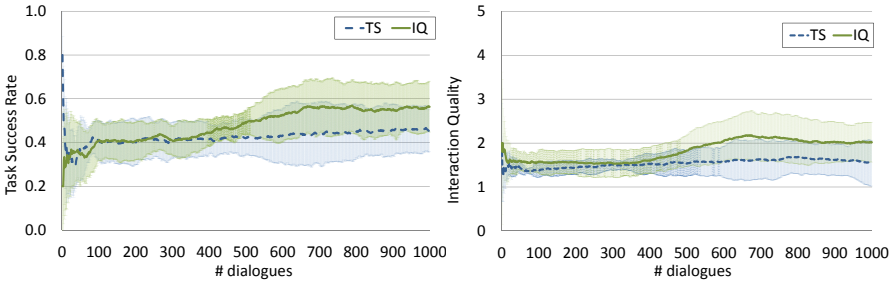

Fig. 1 Moving task success rate and moving interaction quality computed over a window of 200 turns for both reward models. The learning curves are averaged over five trained policies.

rates compared to R_{TS} (Exp. 1) and yields higher IQ values than R_{IQ} (Exp. 2) with 2.0 compared to 1.5. Even though the success rate for R_{IQ} with an TSR of 56.3% surpasses R_{TS} with a TSR of 45.4%, this may regarded as a statistical insignificant.

As both expectations have been met, the results clearly suggest that IQ is applicable for dialogue policy learning. Although the experiments have only been carried out in simulation, we would expect to see similar behaviour in real user experiments.

4 Conclusion

In this paper, we have presented user satisfaction represented by the Interaction Quality (IQ) metric as an alternative to task success for modelling the dialogue-level reward used for reinforcement learning of the dialogue policy. We have analysed its applicability by formulating two key expectations on the relation of task success and IQ and shown in simulated experiments that these expectations have been clearly met. Learning a dialogue policy using IQ in the reward results in similar performance of the resulting policy in terms of task success while achieving better results in terms of estimated user satisfaction.

Of course, to asses the impact on user satisfaction, experiments with real users are necessary which will be part of future work. Furthermore, while the difference in TSR is not significant, it needs to be investigated further. Finally, IQ is currently only regarded as being part of the reward. It may as well be part of the dialogue state which will also be in the focus of future work.

References

1. El Asri, L., Khouzaimi, H., Laroche, R., Pietquin, O.: Ordinal regression for interaction quality prediction. In: Proc. of ICASSP, pp. 3245–3249. IEEE (2014)
2. El Asri, L., Laroche, R., Pietquin, O.: Reward Function Learning for Dialogue Management. In: Proc. of the 6th STAIRS, pp. 95–106. IOS Press (2012)

3. El Asri, L., Laroche, R., Pietquin, O.: Reward shaping for statistical optimisation of dialogue management. In: Statistical Language and Speech Processing, pp. 93–101. Springer (2013)
4. Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., Young, S.J.: On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In: Proc. of ICASSP, pp. 8367–8371. IEEE (2013)
5. Lee, S., Eskenazi, M.: An unsupervised approach to user simulation: toward self-improving dialog systems. In: Proc. of 13th SIGDial, pp. 50–59. ACL (2012)
6. Levin, E., Pieraccini, R.: A stochastic model of computer-human interaction for learning dialogue strategies. In: Eurospeech, vol. 97, pp. 1883–1886 (1997)
7. Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M.: Doing research on a deployed spoken dialogue system: One year of let's go! experience. In: Proc. of ICSLP (2006)
8. Rieser, V., Lemon, O.: Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In: Proc. of 46th ACL-HLT, pp. 638–646. ACL (2008)
9. Schmitt, A., Schatz, B., Minker, W.: Modeling and predicting quality in spoken human-computer interaction. In: Proc. of 12th SIGDial, pp. 173–184. ACL, Portland, OR (2011)
10. Schmitt, A., Ultes, S.: Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. Speech Communication **74**, 12–36 (2015). DOI 10.1016/j.specom.2015.06.003
11. Schmitt, A., Ultes, S., Minker, W.: A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In: Proc. of LREC, pp. 3369–337 (2012)
12. Su, P.H., Gašić, M., Mrkšić, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Wen, T.H., Young, S.: On-line active reward learning for policy optimisation in spoken dialogue systems. In: Proc. of 54th ACL, pp. 2431–2441. ACL (2016)
13. Ultes, S., Dikme, H., Minker, W.: Dialogue Management for User-Centered Adaptive Dialogue. In: Situated Dialog in Speech-Based Human-Computer Interaction, pp. 51–61. Springer International Publishing, Cham (2016). DOI 10.1007/978-3-319-21834-2_5
14. Ultes, S., Heinroth, T., Schmitt, A., Minker, W.: A theoretical framework for a user-centered spoken dialog manager. In: Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, pp. 241–246. Springer New York, New York, NY (2011). DOI 10.1007/978-1-4614-1335-6_24
15. Ultes, S., Kraus, M., Schmitt, A., Minker, W.: Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In: Proc. of 16th SIGDIAL, pp. 374–383. ACL (2015)
16. Ultes, S., Minker, W.: Interaction Quality Estimation in Spoken Dialogue Systems Using Hybrid-HMMs. In: Proc. of 15th SIGDIAL, pp. 208–217. ACL (2014)
17. Ultes, S., Minker, W.: Managing adaptive spoken dialogue for intelligent environments. Journal of Ambient Intelligence and Smart Environments **6**(5), 523–539 (2014). DOI 10.3233/ais-140275
18. Ultes, S., Schmitt, A., Minker, W.: On quality ratings for spoken dialogue systems – experts vs. users. In: Proc. of the 2013 NAACL-HLT, year = 2013, pages = 569–578, publisher = ACL, location = Atlanta, Georgia
19. Ultes, S., Schmitt, A., Minker, W.: Towards quality-adaptive spoken dialogue management. In: NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012), pp. 49–52. ACL, Montréal, Canada (2012)
20. Walker, M.: An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. Journal of Artificial Intelligence Research **12**, 387–416 (2000)
21. Walker, M., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: a framework for evaluating spoken dialogue agents. In: Proc. of 8th EACL, pp. 271–280. ACL, Morristown, NJ, USA (1997). DOI 10.3115/979617.979652
22. Williams, J.D., Young, S.J.: Characterizing task-oriented dialog using a simulated asr chanel. In: Proc. of 8th Interspeech, pp. 185–188 (2004)
23. Young, S.J., Gašić, M., Thomson, B., Williams, J.D.: POMDP-based statistical spoken dialog systems: A review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)