Domain-independent User Satisfaction Reward Estimation for Dialogue Policy Learning

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić and Steve Young

Engineering Department, University of Cambridge, Cambridge, United Kingdom

{su259,pfb30,ic340,nm480,lmr46,phs26,thw28,mg436,sjy11}@cam.ac.uk

Abstract

Learning suitable and well-performing dialogue behaviour in statistical spoken dialogue systems has been in the focus of research for many years. While most work which is based on reinforcement learning employs an objective measure like task success for modelling the reward signal, we propose to use a reward based on user satisfaction. We will show in simulated experiments that a live user satisfaction estimation model may be applied resulting in higher estimated satisfaction whilst achieving similar success rates. Moreover, we will show that one satisfaction estimation model which has been trained on one domain may be applied in many other domains which cover a similar task. We will verify our findings by employing the model to one of the domains for learning a policy from real users and compare its performance to policies using the user satisfaction and task success acquired directly from the users as reward.

Index Terms: spoken dialogue systems, statistical dialogue management, interaction quality, reinforcement learning

1. Introduction

For modelling the decision-making component of a spoken dialogue system (SDS), the dialogue policy, different approaches exist. A very prominent one is to model the problem as a (partially observable) Markov decision process ((PO)MDP) using reinforcement learning (RL) to learn the optimal system behaviour. In RL, the policy π is trained to make decisions so that a potentially delayed objective (the reward function) is maximised. For information-seeking dialogues, most existing work uses task success as principal component of the reward function.

The goal of this paper is to demonstrate that user satisfaction (US) may be used as a reward to learn dialogue policies which not only maximise US but also lead to high task success rates. We argue that training a system to maximise US is a good alternative to task success (TS) for the following reasons:

1. User satisfaction represents the user's view of the interaction while task success represents the system's view. However, it is the user who ultimately decides whether to continue using the system or not. In fact, task success has been used in prior work precisely because it does correlate well with user satisfaction [1].
2. User satisfaction—in contrast to TS—may be linked to interaction phenomena which are independent of the user's goal [2] and hence, no prior knowledge of the goal or any other domain dependent information is required.
3. As user satisfaction is independent of application domain information, the use of an estimator of user satisfaction has the potential to generalize well across domains. Thus, learning dialogue policies for new, previously unseen domains becomes much easier.

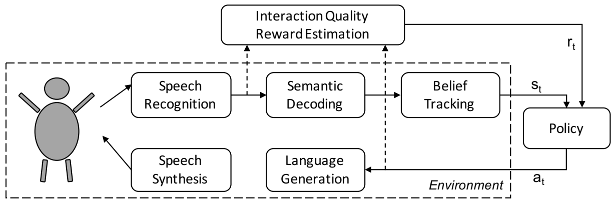

Figure 1: *The proposed RL framework integrating an interaction quality reward estimator. The policy learns to take action a_t at time t while being in state s_t and receiving reward r_t.*

In this contribution, the emphasis lies on the second and third item. Following up on previous work [3, 4, 5, 6, 7], interaction quality (IQ)—a less subjective version of user satisfaction[1]—will be used as reward estimator. The estimation model is based on domain-independent, interaction-related features which do not have any information available about the goal of the dialogue. The model has been trained on manually annotated dialogue turns of a bus information system achieving an accuracy of 0.89[2]. The proposed RL framework is shown in Figure 1. It has previously been applied for in-domain experiments and simulated evaluation [7]. In this paper, we will complete the work by using a modified reward function to show its domain-independence and the resulting high potential to be applicable for learning in unseen domains. Moreover, the estimator is used in an experiment where the policy is learned through interaction with real humans.

Most of previous work focuses on employing task success as the main reward signal [9, 10, 11, 12, 13, 14, 15, 16]. However, task success is usually only computable for predefined tasks e.g., through interactions with simulated or recruited users, where the underlying goal is known in advance. To overcome this, the required information can be requested directly from users at the end of each dialogue [17]. However, this can be intrusive, and users may not always cooperate.

An alternative is to use a task success estimator [18, 15, 16]. With the right choice of features, these can also be applied to new and unseen domains [19]. However, these models still attempt to estimate completion of the underlying task, whereas our model evaluates the overall user experience.

In this paper, we show that an interaction quality reward estimator trained on dialogues from a bus information system will result in well-performing dialogues both in terms of success rate and user satisfaction on five other domains, while only using interaction-related, domain-independent information, i.e., not knowing anything about the task of the domain.

Others have previously introduced user satisfaction into

[1]The relation of US and IQ has been closely investigated in [2, 8].

[2]taking into account neighbouring values, cf. Sec. 2

the reward [20, 21, 22, 23] by using the PARADISE framework [24]. However, to derive user ratings within that framework, users have to answer a questionnaire which is usually not feasible in real world settings. To overcome this, PARADISE has been used in conjunction with expert judges instead [25, 26] to enable unintrusive acquisition of dialogues. 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 whereas a non-linear dependency might be more appropriate [27]. Therefore, we use interaction quality [2] in this work because it 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: in Section 2, the interaction quality reward estimation module is presented in detail. Section 3 contains the simulated experiments on several domains as well as an experiment with paid subjects. The findings are discussed in Section 4 and conclusions are drawn in Section 5.

2. Interaction Quality Reward Estimation

In this work, we propose to use interaction quality (IQ) [2] as a reward estimator for learning information-seeking dialogue policies. IQ represents a less subjective variant of user satisfaction: instead of being acquired from users directly, experts annotate pre-recorded dialogues to avoid the large variance that is often encountered when users rate their dialogues directly [2].

IQ is defined on a five-point scale from five (satisfied) down to one (extremely unsatisfied). To derive a reward from this value, the equation

$$R_{IQ} = T \cdot (-1) + (iq - 1) \cdot 5 \qquad (1)$$

is used where R_{IQ} describes the final reward. It is applied to the final turn of the dialogue of length T with a final IQ value of iq. Thus, a per-turn penalty of -1 is added to the dialogue outcome. This results in a reward range of 19 down to $-T$ which is consistent with related work [9, 19, 16, e.g.] in which binary task success (TS) was used to define the reward as:

$$R_{TS} = T \cdot (-1) + \mathbb{1}_{TS} \cdot 20 \; , \qquad (2)$$

where $\mathbb{1}_{TS} = 1$ only if the dialogue was successful, $\mathbb{1}_{TS} = 0$ otherwise. R_{TS} will be used as a baseline.

The problem of estimating IQ is cast as a classification problem where the target classes are the distinct IQ values. The input consists of domain-independent variables called interaction parameters. These parameters incorporate information from the automatic speech recognition (ASR) output and the preceding system action. Based on this information, which is available at every turn, temporal features are computed taking sums, means or counts from the turn-based information for a window of the last 3 system-user-exchanges[3] and the complete dialogue (see Fig. 2). This results in a feature set of 16 parameters as shown in Table 1.

As training data, the LEGO corpus [28] is used which consists of 200 dialogues (4,885 turns) from the Let's Go bus information system [29]. There, users with real needs are able to call the system to get information about the bus schedule. Each turn of these 200 dialogues has been annotated with IQ (representing the quality of the dialogue up to the current turn) by three

[3]a system turn followed by a user turn

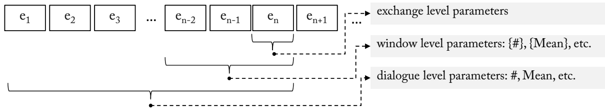

Figure 2: *Modelling of temporal information in the interaction parameters used as input to the IQ estimator.*

Table 1: *The parameters used for IQ estimation extracted on the exchange level from each user input plus counts, sums and rates for the whole dialogue (#,%,Mean) and for a window of the last 3 turns ($\{\cdot\}$).*

Parameter	Description
ASRRecognitionStatus	ASR status: *success*, *no match*, *no input*
ASRConfidence	confidence of top ASR results
RePrompt?	is the system question the same as in the previous turn?
ActivityType	general type of system action: *statement*, *question*
Confirmation?	is system action confirm?
MeanASRConfidence	mean ASR confidence if ASR is success
#Exchanges	number of exchanges (turns)
#ASRSuccess	count of ASR status is success
%ASRSuccess	rate of ASR status is success
#ASRRejections	count of ASR status is reject
%ASRRejections	rate of ASR status is reject
{Mean}ASRConfidence	mean ASR confidence if ASR is success
{#}ASRSuccess	count of ASR is success
{#}ASRRejections	count of ASR status is reject
{#}RePrompts	count of times RePromt? is true
{#}SystemQuestions	count of ActivityType is question

experts. The final IQ label has been assigned using the median of the three individual labels.

The estimation model was trained using a support vector machine [30, 31] achieving an unweighted average recall[4] (UAR) of 0.55 with 10-fold cross-validation. However, missing the correct estimated IQ value by only one has little impact for modelling the reward, and if neighbouring values are taken into account, the model achieves an accuracy of 0.89.

As a comparison, previous work has used the LEGO corpus with a full IQ feature set (which includes additional partly domain-related information) and this achieves a UAR of 0.55 using ordinal regression [32], 0.53 using a two-level SVM approach [33], and 0.51 using a hybrid-HMM [34]. Human performance on the same task is 0.69 UAR [2].

3. Experiments and Results

The proposed IQ reward estimation framework (Fig. 1) was evaluated on several domains within a simulated environment. Furthermore, the simulated results were validated by applying the framework to one of the domains and learning the policy directly through interaction with real humans.

3.1. Experimental Setup

To train and evaluate the proposed framework, a policy model based on the GP-SARSA algorithm [9] is used. This is a value-based method that uses a Gaussian process to approximate the state-value function. As it takes into account the uncertainty of the approximation, it is very sample efficient and may even be used to learn a policy directly through real human interaction [17]. The decisions of the policy are based on a summary

[4]UAR is the arithmetic average of all class-wise recalls.

Table 2: *Statistics of the domain the IQ reward estimator is trained on (LetsGo) and the domains it is applied to.*

Domain	*Code*	*# constraints*	*# DB items*
LetsGo		4	-
CamRestaurants	CR	3	110
CamHotels	CH	5	33
SFRestaurants	SR	6	271
SFHotels	SH	6	182
Laptops	L	6	126

Table 3: *Results of the simulated experiments for all domains showing task success rate (TSR), average interaction quality (AIQ), and average dialogue length (ADL) in number of turns. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials. * marks statistically significant difference between R_{TS} and R_{IQ} ($p < 0.05$, T-test).*

Domain	*SER*	*TSR*		*AIQ*		*ADL*	
		R_{TS}	R_{IQ}	R_{TS}	R_{IQ}	R_{TS}	R_{IQ}
CR	0%	0.98	0.98	3.88	3.96	4.37	4.34
	15%	0.86	0.85	3.51*	3.76*	5.21	4.93
	30%	0.84*	0.76*	3.34	3.46	5.73	5.54
CH	0%	0.97	0.96	3.02*	3.32*	5.74	5.79
	15%	0.79*	0.66*	2.69*	3.21*	7.27*	6.53*
	30%	0.62	0.55	2.13*	2.72*	8.81*	7.87*
SR	0%	0.93	0.93	2.88*	3.36*	6.31*	5.57*
	15%	0.58	0.65	2.5*	3.25*	8.03*	6.62*
	30%	0.46	0.41	2.17*	2.71*	9.13*	7.95*
SH	0%	0.94	0.93	3.1*	3.36*	5.66	5.92
	15%	0.71	0.67	2.61*	3.07*	7	6.73
	30%	0.51	0.5	2.29*	2.77*	8.94	8.64
L	0%	0.85	0.89	2.68*	3.11*	7.01*	6.15*
	15%	0.59	0.63	2.12*	2.97*	9.04*	6.72*
	30%	0.45	0.41	2.1*	2.52*	9.11*	8.09*
TV	0%	0.92*	0.86*	3.08*	3.42*	5.84	5.76
	15%	0.85*	0.78*	2.85*	3.44*	6.78*	5.88*
	30%	0.69	0.68	2.77*	3.06*	7.2	6.75

space representation of the dialogue state tracker. In this work, the focus tracker [35]—an effective rule-based tracker—is used. The policy may choose out of a set of summary actions which are based on general intents like *request*, *confirm* or *inform*. The exact number of system actions varies for the domains and ranges from 16 to 25.

The IQ reward estimator is evaluated against the baseline of using the traditional reward function based on task success (TS). While IQ needs to be estimated, TS can be computed by comparing the outcome of each dialogue with the pre-defined goal. Of course, this is only possible in simulation and when evaluating with paid subjects. This goal information is not available to the IQ estimator, nor is it required.

To measure the dialogue performance, the task success rate (TSR) and the average interaction quality (AIQ) are measured: the TSR represents the ratio of dialogues for which the system was able to provide the correct result. AIQ is calculated based on the estimated IQ at the end of each dialogue.

3.2. Domain-independent Learning from Simulation

For the simulation experiments, the performance of the trained polices on five different domains was evaluated: Cambridge Hotels and Restaurants, San Francisco Hotels and Restaurants, and Laptops. The complexity of each domain is shown in Table 2 and compared to the LetsGo domain (the domain the estimator has been trained on).

The dialogues were created using PyDial [36] which contains an implementation of the agenda-based user simulator [37] with an additional error model to simulate the required semantic error rate (SER) caused in the real system by the noisy speech channel. For each domain, both reward models are compared on three SERs: 0%, 15%, and 30%. Hence, for each domain and for each SER, policies have been trained using 1,000 dialogues followed by an evaluation step of 100 dialogues. The results in Table 3 were computed based on the evaluation step averaged over three train/evaluation cycles with different random seeds.

The results nicely show the successful evaluation of the policies using the IQ reward estimator in terms of TSR and AIQ. For the domains CR, CH, SR, and SH, the TSRs of R_{IQ} are very similar to the TSRs of R_{TS} for an SER of 0%. This slightly degrades for higher SERs. This behaviour may be attributed to the following two reasons: the more the source domain (LetsGo) of the estimator and the target domain differ, the more the results differ in terms of TSR. Furthermore, the higher the noise, the more the policy has to focus on success[5].

Naturally, as only the IQ-based model is aware of the IQ concept and indeed is trained to optimise it, the results show that the AIQs are better throughout the experiments.

In comparison to the task success estimator proposed by Vandyke et al. [19] who trained the estimator on CR and applied it to SF and SH achieving comparable results, the model proposed here does not require the maximum number of slots to be defined, i.e., the features which have been used for the user satisfaction estimator are independent of the slots.

3.3. Learning from Real Humans

For learning a policy directly from the interaction with real humans, the CR domain was used. Using the Amazon Mechanical Turk, subjects were recruited to talk to the telephone-based dialogue system. At the end of each dialogue, users were asked two questions. The first was a yes/no question targeting the dialogue success ("Have you found all the information you were looking for?") which has been used as the baseline for R_{TS}. As this label is noisy, only the dialogues where this *subjective* success label matches the *objective* success were used for policy training [17] ($obj = subj$).

A second baseline was also included: directly acquiring a user satisfaction (US) rating from the users after each dialogue. For this, the second question posed was: "How satisfied are you with the interaction?" The users were able to respond on a six-point scale: 6=very satisfied, 5=satisfied, 4=generally ok, 3=unsatisfied, 2=very unsatisfied or 1=extremely unsatisfied. This rating was converted to a reward in correspondence with R_{IQ}:

$$R_{US} = T \cdot (-1) + (US - 1) \cdot 5 \ . \qquad (3)$$

Hence, each dialogue was also evaluated using the average user satisfaction (AUS).

Two policies were trained for each reward function. The learning curves show moving TSR, moving AIQ and moving AUS and are presented in Figure 3. Each value in the graphs is

[5]The satisfaction will be less different in noisy channels between a dialogue which is successful but has a lot of ASR non-understandings compared to a dialogue that is not successful. Thus, there might be an upper bound for the satisfaction if there is noise in the channel with the consequence that satisfaction plays a reduced role in training.

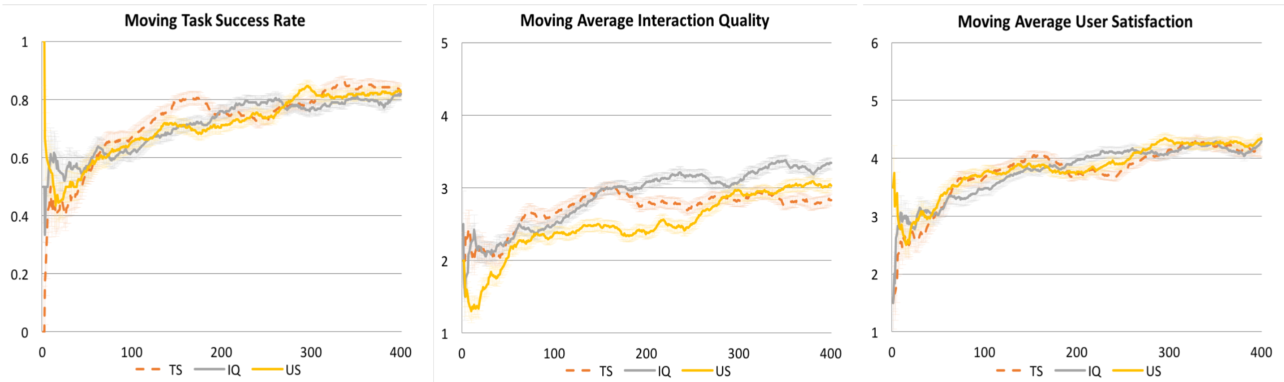

Figure 3: *Moving TSR (left), moving AIQ (middle) and moving AUS (right) for using either TS, IQ, or US as reward averaged over two policies respectively, computed on windows consisting of 120 dialogues.*

Sys	Welcome to the Cambridge restaurant dialogue system. How may I help you?
Usr	i'm looking for european food in the expensive price range
Sys	Welcome to the Cambridge restaurant dialogue system. How may I help you?
Usr	i want to european restaurant
Sys	Would you like something in the cheap, moderate, or expensive price range?
Usr	i don't care the area
Sys	Would you like something in the cheap, moderate, [. . .] ?
Usr	european expensive
Sys	Would you like something in the cheap, moderate, [. . .] ?
Usr	european food
Sys	What part of town do you have in mind?
Usr	i don't care
Sys	michaelhouse cafe is a nice place [. . .] .
Usr	thank you goodbye

Figure 4: *Example dialogue trained using R_{TS} after 387 training dialogues resulting in $iq = 1$.*

calculated using the respective previous 120 training dialogues averaged over the three trials.

The results show that R_{IQ} may successfully be used for learning through interactions with real users. It achieves the same TSR compared to the baseline of R_{TS} while resulting in better results for AIQ. Furthermore, the results for moving AUS indicate a slightly better user satisfaction compared to R_{TS}.

The second baseline of using R_{US} also resulted in competitive results showing a similar TSR as R_{IQ} and R_{TS} whilst resulting in slightly better AUS. This indicates that for a problem as defined by the CR domain, using user satisfaction directly as a learning signal is also a viable option for learning policies.

Figures 4 and 5 show two successful example dialogues for the models trained with R_{TS} and R_{IQ}, respectively. One effect of training with R_{IQ} was a reduced number of system repetitions (which may be linked to the RePrompt? feature).

4. Discussion

A key aspect of this work to emphasise is that the estimator works without any knowledge about the domain. So, in contrast to task success estimators [15, 19], it does not use the dialogue state as input. Simply by using parameters encoding interaction characteristics, a dialogue policy was trained to achieve not only good US (which it is optimised on) but also a good TSR.

One limitation of the proposed IQ reward estimator is that it requires manual annotation of dialogues with interaction quality

Sys	Welcome to the Cambridge restaurant dialogue system. How may I help you?
Usr	international food
Sys	You are looking for a international venue right?
Usr	international
Sys	Would you like something in the cheap, moderate, or expensive price range?
Usr	don't care
Sys	You are looking for a international venue right?
Usr	yes
Sys	What part of town do you have in mind?
Usr	center
Sys	bloomsbury restaurant is a nice place [. . .] .
Usr	thank you goodbye

Figure 5: *Example dialogue trained using R_{IQ} after 365 training dialogues resulting in $iq = 4$.*

labels. These labels incur a higher annotation cost than success labels. However, for this work, a total of only 200 annotated dialogue were sufficient to create a model that was able to be used on several different domains. In contrast, training a task success estimator based on recurrent neural networks typically requires 1,000 annotated dialogues [15, 19].

5. Conclusion

This work has shown that employing a user satisfaction reward estimator for learning dialogue policies without any knowledge about the domain can yield good performance in terms of both task success rate and (estimated) user satisfaction. This has been demonstrated by training the reward estimator on a bus information domain and applying it to learn dialogue policies in five different domains (Cambridge restaurants and hotels, San Francisco restaurants and hotels, Laptops) in a simulated experiment. Moreover, the estimator has successfully been applied to learning dialogue policies in the domain of finding a restaurant in Cambridge through interaction with real users.

For future work, the problem of degrading performance if the noise level increases should be tackled. One possible solution would be to have a combination of success and satisfaction as the reward. In addition, active learning will be investigated to mitigate the requirement for IQ annotated training data.

6. Acknowledgements

This research was funded by the EPSRC grant EP/M018946/1 *Open Domain Statistical Spoken Dialogue Systems*.

7. References

[1] J. D. Williams and S. J. Young, “Characterizing task-oriented dialog using a simulated asr chanel.” in *Proc. of the 8th Interspeech*, 2004, pp. 185–188.

[2] A. Schmitt and S. Ultes, “Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction,” *Speech Communication*, vol. 74, pp. 12–36, Nov. 2015.

[3] S. Ultes, T. Heinroth, A. Schmitt, and W. Minker, “A theoretical framework for a user-centered spoken dialog manager,” in *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, R. López-Cózar and T. Kobayashi, Eds. New York, NY: Springer New York, Sep. 2011, pp. 241–246.

[4] S. Ultes, A. Schmitt, and W. Minker, “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)*. Montréal, Canada: ACL, Jun. 2012, pp. 49–52.

[5] S. Ultes, H. Dikme, and W. Minker, “Dialogue Management for User-Centered Adaptive Dialogue,” in *Situated Dialog in Speech-Based Human-Computer Interaction*, A. I. Rudnicky, A. Raux, I. Lane, and T. Misu, Eds. Cham: Springer International Publishing, 2016, pp. 51–61.

[6] S. Ultes, M. Kraus, A. Schmitt, and W. Minker, “Quality-adaptive spoken dialogue initiative selection and implications on reward modelling,” in *Proc. of the 16th SIGDial Conference*. ACL, Sep. 2015, pp. 374–383.

[7] S. Ultes, J. Miehle, and W. Minker, “On the applicability of a user satisfaction-based reward for dialogue policy learning,” in *Proc. of the 9th IWSDS)*, Jun. 2017.

[8] S. Ultes, A. Schmitt, and W. Minker, “On quality ratings for spoken dialogue systems – experts vs. users,” in *Proc. of NAACL-HLT*. ACL, Jun. 2013, pp. 569–578.

[9] M. Gašić and S. J. Young, “Gaussian processes for POMDP-based dialogue manager optimization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 28–40, 2014.

[10] M. Gašić, D. Kim, P. Tsiakoulis, C. Breslin, M. Henderson, M. Szummer, B. Thomson, and S. J. Young, “Incremental on-line adaptation of POMDP-based dialogue managers to extended domains,” in *Proc. of the 15th Interspeech*. ISCA, Aug. 2014, pp. 140–144.

[11] O. Lemon and O. Pietquin, “Machine learning for spoken dialogue systems,” in *Proc. of the 8th Interspeech)*, 2007, pp. 2685–2688.

[12] L. Daubigney, M. Geist, and O. Pietquin, “Off-policy Learning in Large-scale POMDP-based Dialogue Systems,” in *Proc. of the 37th ICASSP*. Kyoto (Japan): IEEE, 2012, pp. 4989–4992.

[13] E. Levin and R. Pieraccini, “A stochastic model of computer-human interaction for learning dialogue strategies.” in *Proc. of the 5th Eurospeech*, vol. 97, 1997, pp. 1883–1886.

[14] S. J. Young, M. Gašić, B. Thomson, and J. D. Williams, “POMDP-based statistical spoken dialog systems: A review,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[15] P.-H. Su, D. Vandyke, M. Gašić, D. Kim, N. Mrkšić, T.-H. Wen, and S. J. Young, “Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems,” in *Proc. of the 16th Interspeech*. ISCA, 2015, pp. 2007–2011.

[16] P.-H. Su, M. Gašić, N. Mrkšić, L. Rojas-Barahona, S. Ultes, D. Vandyke, T. H. Wen, and S. Young, “On-line active reward learning for policy optimisation in spoken dialogue systems,” in *Proc. of the 54th ACL*. ACL, Aug. 2016, pp. 2431–2441.

[17] M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. J. Young, “On-line policy optimisation of Bayesian spoken dialogue systems via human interaction,” in *Proc. of the 23th ICASSP*. IEEE, 2013, pp. 8367–8371.

[18] L. El Asri, R. Laroche, and O. Pietquin, “Task completion transfer learning for reward inference,” *Proc of MLIS*, 2014.

[19] D. Vandyke, P.-H. Su, M. Gašić, N. Mrkšić, T.-H. Wen, and S. Young, “Multi-domain dialogue success classifiers for policy training,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 763–770.

[20] M. Walker, J. C. Fromer, and S. S. Narayanan, “Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email,” in *Proc. of the 36th ACL and 17th CoLing-Volume 2*. ACL, 1998, pp. 1345–1351.

[21] M. Walker, “An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email,” *Journal of Artificial Intelligence Research*, vol. 12, pp. 387–416, 2000.

[22] V. Rieser and O. Lemon, “Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation.” in *Proc. of the 46th ACL-HLT*. ACL, Jun. 2008, pp. 638–646.

[23] ——, “Automatic learning and evaluation of user-centered objective functions for dialogue system optimisation,” in *Proc. of 6th LREC*, N. Calzolari, K. Choukri, and B. Maegaard, Eds. Marrakech, Morocco: ELRA, May 2008, pp. 2356–2361.

[24] M. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “PARADISE: a framework for evaluating spoken dialogue agents,” in *Proc. of the 8th EACL*. Morristown, NJ, USA: ACL, 1997, pp. 271–280.

[25] L. El Asri, R. Laroche, and O. Pietquin, “Reward Function Learning for Dialogue Management,” in *Proc. of the 6t Starting AI Researchers' Symposium (STAIRS)*. IOS Press, 2012, pp. 95–106.

[26] ——, “Reward shaping for statistical optimisation of dialogue management,” in *Statistical Language and Speech Processing*. Springer, 2013, pp. 93–101.

[27] A. Schmitt, B. Schatz, and W. Minker, “Modeling and predicting quality in spoken human-computer interaction,” in *Proc. of the 12th SIGDial Conference*. Portland, Oregon, USA: ACL, Jun. 2011, pp. 173–184.

[28] A. Schmitt, S. Ultes, and W. Minker, “A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system,” in *International Conference on Language Resources and Evaluation (LREC)*, May 2012, pp. 3369–337.

[29] A. Raux, D. Bohus, B. Langner, A. W. Black, and M. Eskenazi, “Doing research on a deployed spoken dialogue system: One year of let's go! experience,” in *Proc. of the 7th Interspeech*, Sep. 2006.

[30] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[31] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[32] L. El Asri, H. Khouzaimi, R. Laroche, and O. Pietquin, “Ordinal regression for interaction quality prediction,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2014, pp. 3245–3249.

[33] S. Ultes and W. Minker, “Improving Interaction Quality Recognition Using Error Correction,” in *Proc. of the 14th SIGDial Conference*. ACL, Aug. 2013, pp. 122–126.

[34] ——, “Interaction Quality Estimation in Spoken Dialogue Systems Using Hybrid-HMMs,” in *Proc. of the 15th SIGDial Conference*. ACL, Jun. 2014, pp. 208–217.

[35] M. Henderson, B. Thomson, and J. Williams, “The second dialog state tracking challenge,” in *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, vol. 263, 2014.

[36] S. Ultes, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gašić, and S. J. Young, “Pydial: A multi-domain statistical dialogue system toolkit,” in *Proc. of the 55th ACL–Demos*. ACL, 2017, software available at http://www.pydial.org.

[37] J. Schatzmann and S. J. Young, “The hidden agenda user simulation model,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 733–747, 2009.