Blending Task Success and User Satisfaction: Analysis of Learned Dialogue Behaviour with Multiple Rewards

Stefan Ultes and **Wolfgang Maier**
Mercedes-Benz Research & Development
Sindelfingen, Germany
{stefan.ultes,wolfgang.mw.maier}@daimler.com

Abstract

Recently, principal reward components for dialogue policy reinforcement learning use task success and user satisfaction independently and neither the resulting learned behaviour has been analysed nor a suitable proper analysis method even existed. In this work, we employ both principal reward components jointly and propose a method to analyse the resulting behaviour through a structured way of probing the learned policy. We show that blending both reward components increases user satisfaction without sacrificing task success even in more hostile environments and provide insight about actions chosen by the learned policies.

1 Introduction and Related Work

The core task of a spoken dialogue systems is to select the next system response to a given user input utterance. Modular systems divide this problem into the sub-problems natural language understanding, dialogue state tracking, dialogue policy execution, and natural language generation. For many years, research on modular spoken dialogue systems has rendered this decision making task of finding the optimal policy as a reinforcement learning (RL) problem that optimises an expected long-term future reward. The principal reward component has previously been either task success (TS) (Gašić and Young, 2014; Daubigney et al., 2012; Levin and Pieraccini, 1997; Young et al., 2013; Su et al., 2016, 2015; Lemon and Pietquin, 2007; Ultes et al., 2018) or user satisfaction (US) (e.g. Walker, 2000; Ultes, 2019) independently.

The goal of this paper is to apply both, TS and US, as principal reward components at the same time and to gain insights into the learned dialogue behaviour. This requires a learning setup that allows multiple principle reward components simultaneously and an analysis method with a structured procedure to probe learned dialog policies. This is achieved through a multi-objective reinforcement learning (MORL) setup (Ultes et al., 2017b) and an analysis method that builds upon work from Ultes and Maier (2020). The chosen MORL setup employs a linear reward scalarisation that combines the principal reward components TS and interaction quality (IQ) (Schmitt and Ultes, 2015)—a more objective measure for modelling US.

The two main contributions of this work are (1) a universal behaviour analysis method that aims at investigating the influence of multiple learning objectives on the learned dialog policy and (2) analysing the performance and learned behaviour when blending TS and IQ as principal reward components.

Previous work on RL-based dialogue policy learning focused either on TS or US as the principal reward component. Task success can be computed (Schatzmann and Young, 2009; Gašić et al., 2013, e.g.) or estimated (El Asri et al., 2014b; Su et al., 2015; Vandyke et al., 2015; Su et al., 2016) only when information about the task and underlying goal are known in advance. Integrating US into the reward by using the PARADISE (Walker et al., 1997) framework (Walker, 2000; Rieser and Lemon, 2008; El Asri et al., 2013, e.g.) or through a measure called response quality (Bodigutla et al., 2020, e.g.). Both are not suitable for this research as PARADISE directly incorporates task knowledge and response quality incorporates functionality of back-end services.

Ultes et al. (2017a; 2019) showed that a pre-trained interaction quality reward estimator can lead to a policy that is able to produce successful dialogues while achieving higher user satisfaction. This has been shown across different domains, including the domain that is used in this work. However, success declines with increasing noise in the communication channel, increasing differences in domain structure, and less co-operative users. Combining TS and IQ poses one viable way of learning

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 403–410
July 29–31, 2021.

dialogue policies that lead to a good task success rate while still achieving good user satisfaction.

Section 2 presents the employed MORL algorithm and interaction quality estimation method that are both used together with different ways of reward modelling (Sec. 3) for learning dialogue policies. The experiments and their results and analysis are presented in Sections 5 and 6.

2 Preliminaries

The presented work builds upon previously published approaches on multi-objective reinforcement learning and interaction quality modelling:

Interaction Quality Estimation The interaction quality (IQ) (Schmitt and Ultes, 2015) 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 (Schmitt and Ultes, 2015). Interaction quality shows a good correlation with user satisfaction (Ultes et al., 2013) and fulfils the requirements necessary for its application in dialog systems (Ultes et al., 2012, 2016).

Estimating IQ has been cast as a turn-level classification problem where the target classes are the distinct IQ values ranging from 5 (satisfied) down to 1 (extremely unsatisfied). The input consists of domain-independent interaction parameters that incorporate turn-level information from the automatic speech recognition (ASR) output and the preceding system action. Furthermore, temporal features are computed by taking sums, means or counts of the turn-based information for a window of the last three system-user-exchanges[1] and the complete dialogue. Ultes et al. (2017a, 2015) use a feature set of 16 parameters to train a support vector machine (SVM) (Vapnik, 1995; Chang and Lin, 2011) with linear kernel using the LEGO corpus (Schmitt et al., 2012) achieving an unweighted average recall[2] (UAR) of 0.44 in a dialog-wise cross-validation setup. The LEGO corpus consists of 200 dialogues with a total of 4,885 annotated system-user-exchanges from the Let's Go bus information system (Raux et al., 2006; Eskenazi et al., 2008) of Carnegie Mellon University in Pittsburgh, PA. The system provided information about bus schedules and connections to actual users with real needs and was live from 2006 until 2016. Each turn of these 200 dialogues has been annotated with IQ (representing the quality of the dialogue up to the current turn) by three experts. The final IQ label has been assigned using the median of the three individual labels. Subsequent work applied deep neural networks achieving an UAR of 0.45 (Rach et al., 2017) and a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) achieving an UAR of 0.54 (Ultes, 2019).

Previous work has used the LEGO corpus with a full IQ feature set (which includes additional partly domain-related information) achieving an UAR in a turn-wise cross-validation setup of 0.55 using ordinal regression (El Asri et al., 2014a), 0.53 using a two-level SVM approach (Ultes and Minker, 2013), and 0.51 using a hybrid-HMM (Ultes and Minker, 2014). Human performance on the same task is 0.69 UAR (Schmitt and Ultes, 2015).

Multi-objective Reinforcement Learning The task of reinforcement Learning (RL) is to find the optimal policy π^* that maximises a potentially delayed objective (the reward function r) (Sutton and Barto, 1998). In multi-objective reinforcement learning (MORL), the objective function consist of multiple dimensions so that a reward r becomes a vector $\mathbf{r} = (r^1, r^2, \ldots, r^m)$, where m is the number of objectives. A scalarisation function f uses weights $\mathbf{w}$ for the different objectives to map the vector representation to a scalar value.

Ultes et al. (2017b) successfully applied the multi-objective GPSARSA algorithm for dialogue policy learning which will be used in this work. It builds upon the GPSARSA (Gašić and Young, 2014) and directly models the expectation of the scalarised reward vector.

For practical solutions, a MORL setup is only reasonable if the ideal weight configuration is not known during learning time. However, for analysing and comparing different weight settings, MORL offers consistent comparisons between any two different weight configurations as all make use of the same learned policy (and thus all have seen the same data during learning).

3 Reward Modelling

One core contribution of this work is to model the reward using both principal reward components, task success and interaction quality. To remain consistent with related work, an penalty term is added to discount long dialogues.

[1] A system-user-exchange consist of a system turn followed by a user turn.

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

The multi-objective reward function R_w is applied a the end of a dialogue and defined as

$$R_w = w_{ts} \cdot r_{ts} + w_{iq} \cdot r_{iq} - T \ , \qquad (1)$$

where T is the number of dialogue turns, w_{ts} and w_{iq} are the weights for the TS and IQ reward components, $w_{iq} = 1 - w_{ts}$,

$$r_{ts} = \mathbb{1}_{ts} \cdot 20 \qquad (2)$$

is the task success reward component, and

$$r_{iq} = (iq - 1) \cdot 5 \qquad (3)$$

the interaction quality reward component. $\mathbb{1}_{ts} = 1$ iff a dialogue was successful, 0 otherwise. iq is the final estimated IQ score at the end of the dialogue. It is scaled to the range between 0 and 20 to match the values of the TS reward component. A positive reward of 20 has been selected in accordance with related work (e.g. Young et al., 2013; Gašić and Young, 2014; Su et al., 2016).

With this definition of R_w, a weight configuration of $w_{ts} = 1.0, w_{iq} = 0.0$ results in a reward model that only uses TS as the principal reward component and matches exactly the reward model of previous work. Likewise, a weight configuration of $w_{ts} = 0.0, w_{iq} = 1.0$ results in a reward model that only uses the IQ as principal reward component, also matching related work.

One additional scalarisation function is proposed based on a task success gate:

$$R_g = \mathbb{1}_{ts} \cdot (w_{ts} \cdot r_{ts} + w_{iq} \cdot r_{iq}) - T \ . \qquad (4)$$

The main reward component is only non-zero for successful dialogues. Hence, even for $w_{iq} = 1.0$, a positive reward is only possible if the task has been achieved successfully.

4 Behaviour Analysis Method

The second core contribution of this work is to propose and apply a universal behaviour analysis method that is used to gain deeper insight into the behaviour that was learned by applying different reward models. The proposed analysis method builds on the analysis methodology proposed by Ultes and Maier (2020), extending it to the context of MORL. It contains the following main steps:

1. Use MORL to learn ***one*** unified policy for all possible weight configurations.
2. Use a pre-defined and fixed set of generated dialog states to probe the learned policy for each weight configuration of interest.
3. Analyse the resulting system actions, e.g., by quantifying the differences or by visualising the actions for different weight configurations.

This method will be used in this work to gain insights into the behaviour learned from applying different principal reward components.

Table 1: Results of the multi-objective learning setup for R_w and R_g with different weight configurations, $w_{iq} = 1 - w_{ts}$.

	TSR		AIQ		ADL	
w_{ts}	R_w	R_g	R_w	R_g	R_w	R_g
0.0	0.78	0.80	2.58	2.75	7.65	7.67
0.1	0.79	0.80	**2.60**	2.73	7.79	7.79
0.2	0.81	0.81	2.57	2.78	7.66	7.63
0.3	0.83	0.85	2.50	**2.79**	7.89	7.66
0.4	0.85	0.83	2.39	2.59	7.80	7.94
0.5	0.86	0.86	2.28	2.66	7.68	7.43
0.6	0.88	**0.88**	2.34	2.54	7.48	7.63
0.7	0.88	0.87	2.26	2.49	7.50	7.54
0.8	**0.89**	0.86	2.08	2.31	7.54	7.62
0.9	**0.89**	**0.88**	2.08	2.28	**7.44**	**7.40**
1.0	0.90	0.87	1.96	2.31	7.48	7.52

5 Experiments and Results

The experiments are conducted with the publicly available PyDial dialog system toolkit (Ultes et al., 2017c). It contains an agenda-based user simulator (Schatzmann and Young, 2009) with an additional error model to simulate the required semantic error rate (SER) caused in the real system by the noisy speech channel.

For both reward models, five multi-objective GPSARSA policies with different random seeds are trained with 3,000 simulated dialogues each in the Cambridge Restaurants domain[3]. As using interaction quality and task success rewards are both known to perform similar in a setup with co-operative users and low noise, we use a semantic error rate of 15% and a less co-operative simulated user configuration (mostly reflected by the probabilities with which the simulated user voluntarily provides additional information) which corresponds to Task 5.1 of Casanueva et al. (2017).

[3]The experiments do not build upon an existing data set like MultiWOZ (Budzianowski et al., 2018) but generate new dialogues through simulation. However, the domain definitions of PyDial are the ones that produced the ontologies of MultiWOZ.

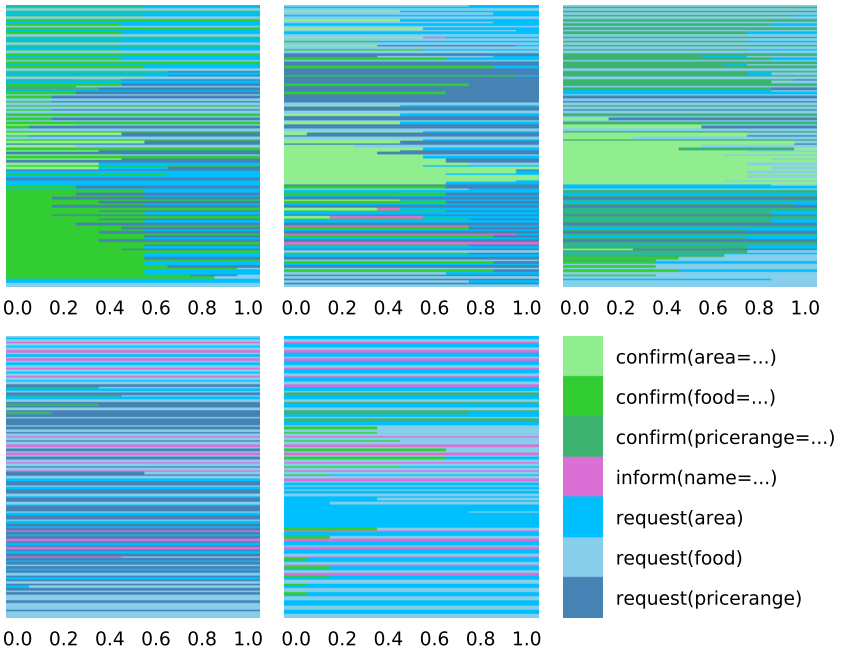

Figure 1: Colour-coding of the resulting system actions of the five trained R_g policies based on the weight configuration having only interaction quality on the left ($w_{ts} = 0.0$) and only task success on the right ($w_{ts} = 1.0$). One line in each graph represents the same state for all policies. The corresponding results of each individual policy and weight configuration are shown in Table 2.

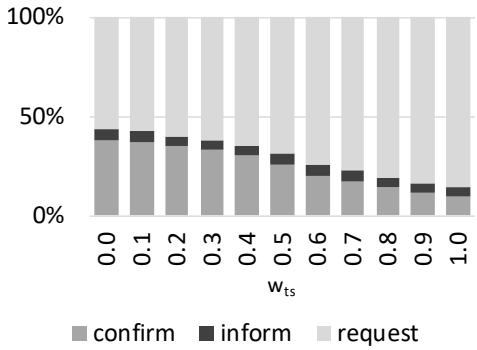

Figure 2: Distribution of the learned dialogue act types based for R_g computed over all random seeds.

The interaction quality reward estimator uses a linear SVM (Ultes et al., 2017a) pre-trained on the LEGO corpus (Schmitt et al., 2012) as described in Section 2. Even though the BiLSTM-based estimator achieved better performance in the experiments (Ultes, 2019), its performance degrades drastically if the user behaviour differs more substantially from the training data. The SVM has already shown its good applicability for the task as it achieves an extended accuracy[4] of 0.89.

Each of the five policies was evaluated for each of the weight configurations (w_{iq}, w_{ts}) in $[(0.0, 1.0), (0.1, 0.9), \ldots, (1.0, 0.0)]$ with 200 dialogues. Absolute results in task success rate (TSR), average dialogue length (ADL) and average interaction quality (AIQ) are shown in Table 1 for R_w and R_g. AIQ uses the estimated interaction quality at the end of each dialogue and computes the average over all dialogues.

The results clearly show the successful application of the learning setup: weight configurations with a high w_{iq} achieve a higher AIQ and weight configurations with a high w_{iq} achieve a high TSR, both for R_w and R_g. Intermediate weight configurations result in AIQ and TSR that lay between the extremes. Another finding is that R_g results in higher AIQ than the non-gated R_w. We speculate that this is due to the removed noise of non-successful training dialogues.

Based on the results, the weight configuration of $(w_{iq} = 0.4, w_{ts} = 0.6)$ is selected as a good compromise between interaction quality and task success reward components both for R_w and R_g.[5]

6 Behaviour Analysis

To gain a deeper understanding about the learned behaviour, 252 states have been generated based on different probabilities of the constraint slots *food-type*, *area*, and *pricerange* ranging from 0.0 to 1.0

[4] taking into account neighbouring values

[5] The question how well this weight balance generalises to other domains and systems is left for future work.

Table 2: Individual results of the five trained R_g policies corresponding to Figure 1 with different weight configurations, $w_{ts} = 1 - w_{iq}$.

	0			1			2			3			4		
w_{ts}	TSR	AIQ	ADL	TSR	AIQ	ADL	TSR	AIQ	ADL	TSR	AIQ	ADL	TSR	AIQ	ADL
0.0	0.79	2.8	7.7	0.84	2.9	7.8	0.79	2.6	7.7	0.83	2.6	7.5	0.74	2.8	7.6
0.1	0.78	2.7	8.1	0.79	2.9	7.5	0.84	2.7	7.7	0.85	2.6	7.4	0.76	2.7	8.3
0.2	0.78	2.7	7.7	0.85	3.0	7.6	0.83	2.8	7.3	0.81	2.6	7.6	0.80	2.8	8.0
0.3	0.90	2.8	7.8	0.88	3.0	7.6	0.91	2.9	7.3	0.82	2.6	7.9	0.77	2.7	7.7
0.4	0.85	2.6	7.9	0.84	2.9	7.7	0.89	2.7	7.7	0.84	2.3	7.8	0.74	2.4	8.7
0.5	0.85	2.5	7.9	0.86	2.8	7.5	0.94	2.9	6.8	0.84	2.3	7.3	0.82	2.8	7.7
0.6	0.86	2.4	7.6	0.92	2.9	7.4	0.89	2.4	7.7	0.88	2.2	7.3	0.85	2.8	8.1
0.7	0.89	2.4	7.7	0.91	2.6	7.7	0.93	2.5	7.2	0.85	2.2	7.5	0.78	2.8	7.7
0.8	0.85	2.3	8.1	0.90	2.4	7.5	0.87	2.3	7.3	0.90	2.1	7.3	0.78	2.5	7.9
0.9	0.91	2.0	7.2	0.91	2.5	7.0	0.87	2.2	7.7	0.91	2.2	7.0	0.82	2.5	8.1
1.0	0.89	2.2	7.7	0.89	2.5	6.9	0.82	2.0	8.1	0.90	2.2	7.2	0.84	2.6	7.6

in steps of 0.05. Each of these was paired with probabilities for the other two slots with $(0.0, 0.0)$, $(0.0, 1.0)$, $(1.0, 0.0)$, and $(1.0, 1.0)$. Each of the five trained multi-objective policies and weight configurations has been probed with these states and the resulting actions have been recorded.

Figure 2 shows a distribution over the dialogue act types of the selected system actions for R_g demonstrating that a high w_{iq} results in a higher percentage of *confirm* dialog acts indicating that a proper grounding strategy increases user satisfaction. R_w shows a similar distribution.

The learned system actions for R_g are shown in Figure 1 with the corresponding performance measures in Table 2: the system actions for the different states are shown for each weight configuration of the five learned policies. Each line in each chart corresponds to the same probing state. This visualisation gives more insight into the selected actions showing that many of the states that produce a *confirm* action for a high w_{iq} produce a *request* action with a high w_{ts}. States that produce *inform* are mostly the same for each w_{ts}[6]. The findings for R_w are similar. Note that this type of visualisation is only possible through the application of MORL where all weight configurations originate in the same policy.

Differences in learned behaviour are quantified by computing the total match rate (TMR) (Ultes and Maier, 2020) between each weight configuration and the extreme configurations of $w_{ts} = 0$ and $w_{ts} = 1$. The results are shown in Figure 3 for R_g demonstrating that TMR decreases with the increased weight differences in a stable fashion with a minimum TAR of 0.69. The proposed optimal weight configuration of $(w_{iq} = 0.4, w_{ts} = 0.6)$ is still quite similar to the extremes with TMRs of 0.87 and 0.81. The findings for R_w are similar.

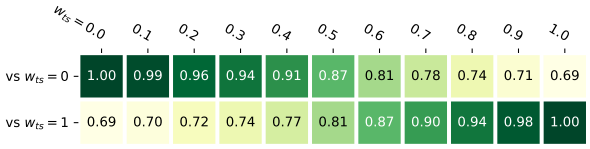

Figure 3: Similarity scores computed between the different weight configurations and $w_{ts} = 0$ and $w_{ts} = 1$.

7 Conclusion

In this work, we presented a universal method for analysing the interplay of multiple principal reward components on the learned dialogue behaviour using multi-objective reinforcement learning and a strategy for probing the resulting policies. This analysis method has been applied successfully to the task of blending task success and user satisfaction rewards. Two findings are that a user satisfaction reward favours *confirmation* system actions and that these confirmations are transformed into requests for task success rewards. Furthermore, an optimal blend was selected for a gated multi-objective reward function supported by similarity scores leading to a good balance between user satisfaction and task success.

In future work, the proposed universal analysis method will be applied to new setups with additional and less complementing principal reward components, e.g., emotions or sentiment. Furthermore, we plan to conduct a human evaluation which compares our proposed model with a model that uses only TS or only IQ.

[6]Some policies do not show any *inform* which means that none of the states, that are used for probing, results in an *inform* action. This emphasises the importance selecting a suitable state set used for probing.

References

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3897–3909, Online. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Iñigo Casanueva, Paweł Budzianowski, Pei-Hao Su, Nikola Mrkšić, Tsung-Hsien Wen, Stefan Ultes, Lina Rojas-Barahona, Steve Young, and Milica Gašić. 2017. A benchmarking environment for reinforcement learning based task oriented dialogue management. In *Deep Reinforcement Learning Symposium, 31st Conference on Neural Information Processing Systems (NIPS)*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Lucie Daubigney, Matthieu Geist, and Olivier Pietquin. 2012. Off-policy Learning in Large-scale POMDP-based Dialogue Systems. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, pages 4989–4992, Kyoto (Japan). IEEE.

Layla El Asri, Hatim Khouzaimi, Romain Laroche, and Olivier Pietquin. 2014a. Ordinal regression for interaction quality prediction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3245–3249. IEEE.

Layla El Asri, Romain Laroche, and Olivier Pietquin. 2013. Reward shaping for statistical optimisation of dialogue management. In *Statistical Language and Speech Processing*, pages 93–101. Springer.

Layla El Asri, Romain Laroche, and Olivier Pietquin. 2014b. Task completion transfer learning for reward inference. *Proc of MLIS*.

Maxine Eskenazi, Alan W Black, Antoine Raux, and Brian Langner. 2008. Let's go lab: a platform for evaluation of spoken dialog systems with real world users. In *Ninth Annual Conference of the International Speech Communication Association*.

Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve J. Young. 2013. On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8367–8371. IEEE.

Milica Gašić and Steve J. Young. 2014. Gaussian processes for POMDP-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Oliver Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. In *European Conference on Speech Communication and Technologies (Interspeech'07)*, pages 2685–2688.

Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Eurospeech*, volume 97, pages 1883–1886.

Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction quality estimation using long short-term memories. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 164–169. Association for Computational Linguistics.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*.

Verena Rieser and Oliver Lemon. 2008. Automatic learning and evaluation of user-centered objective functions for dialogue system optimisation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2356–2361, Marrakech, Morocco. European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2008/.

Jost Schatzmann and Steve J. Young. 2009. The hidden agenda user simulation model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):733–747.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey. European Language Resources Association (ELRA).

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Online active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany. Association for Computational Linguistics.

Pei-Hao Su, David Vandyke, Milica Gašić, Dongho Kim, Nikola Mrkšić, Tsung-Hsien Wen, and Steve J. Young. 2015. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *Interspeech*, pages 2007–2011. ISCA.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*, 1st edition. MIT Press, Cambridge, MA, USA.

Stefan Ultes. 2019. Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 11–20, Stockholm, Sweden. Association for Computational Linguistics.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017a. Domain-independent user satisfaction reward estimation for dialogue policy learning. In *Interspeech*, pages 1721–1725. ISCA.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017b. Reward-balancing for statistical spoken dialogue systems using multi-objective reinforcement learning. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 65–70, Saarbrücken, Germany. Association for Computational Linguistics.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina Rojas-Barahona, Bo-Hsiang Tseng, Yenchen Wu, Steve Young, and Milica Gašić. 2018. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.

Stefan Ultes, Hüseyin Dikme, and Wolfgang Minker. 2016. Dialogue Management for User-Centered Adaptive Dialogue. In Alexander I. Rudnicky, Antoine Raux, Ian Lane, and Teruhisa Misu, editors, *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 51–61. Springer International Publishing, Cham.

Stefan Ultes, Matthias Kraus, Alexander Schmitt, and Wolfgang Minker. 2015. Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 374–383, Prague, Czech Republic. Association for Computational Linguistics.

Stefan Ultes and Wolfgang Maier. 2020. Similarity scoring for dialogue behaviour comparison. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 311–322, 1st virtual meeting. Association for Computational Linguistics.

Stefan Ultes and Wolfgang Minker. 2013. Improving interaction quality recognition using error correction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 122–126, Metz, France. Association for Computational Linguistics.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 208–217, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017c. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2012. 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)*, pages 49–52, Montréal, Canada. Association for Computational Linguistics.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On quality ratings for spoken dialogue systems – experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578, Atlanta, Georgia. Association for Computational Linguistics.

David Vandyke, Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialogue success classifiers for policy training. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 763–770. IEEE.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Marilyn Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Marilyn Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 271–280, Morristown, NJ, USA. Association for Computational Linguistics.

Steve J. Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.