Dialogue & Discourse 12(2) (2021) 81–114 doi: 10.5210/dad.2021.203

User Satisfaction Reward Estimation Across Domains: Domain-independent Dialogue Policy Learning

Stefan Ultes STEFAN.ULTES@DAIMLER.COM
Mercedes-Benz AG Research & Development
Sindelfingen, Germany

Wolfgang Maier WOLFGANG.MW.MAIER@DAIMLER.COM
Mercedes-Benz AG Research & Development
Sindelfingen, Germany

Editor: Kallirroi Georgila

Submitted 12/2020; Accepted 07/2021; Published online 09/2021

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 that is based on reinforcement learning employs an objective measure like task success for modelling the reward signal, we propose to use a reward signal based on user satisfaction. We propose a novel estimator and show that it outperforms all previous estimators while learning temporal dependencies implicitly. We 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 show that a satisfaction estimation model trained on one domain may be applied in many other domains that cover a similar task. We 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 user satisfaction and task success acquired directly from the users as reward.

Keywords: statistical dialogue management, reinforcement learning, spoken dialogue systems

1. Introduction

Spoken dialogue systems (SDSs) enable voice interaction between technical systems and humans. They have been advanced into our everyday lives. Prominent examples include Apple's Siri, Amazon Alexa, or Google Assistant as well as more specialised systems like the in-car voice assistant HeyMercedes. Spoken dialogue systems that target the fulfilment of a certain task are called task-oriented and are usually built using a modular pipeline architecture comprising speech recognition, semantic decoding, dialogue management (consisting of dialogue state tracking and deciding on the next system action), language generation, and speech synthesis (see Figure 1).

One prominent way of modelling the decision-making component of a spoken dialogue system is to use (partially observable) Markov decision processes ((PO)MDPs) (Lemon and Pietquin, 2012; Young et al., 2013). There, reinforcement learning (RL) (Sutton and Barto, 1998) is applied to find the optimal system behaviour represented by the policy π. Task-oriented dialogue systems model the reward r, which is used to guide the learning process, traditionally with task success as the principal reward component (Gašić and Young, 2014; Lemon and Pietquin, 2007; Daubigney et al., 2012; Levin and Pieraccini, 1997; Singh et al., 2002; Young et al., 2013; Su et al., 2015, 2016).

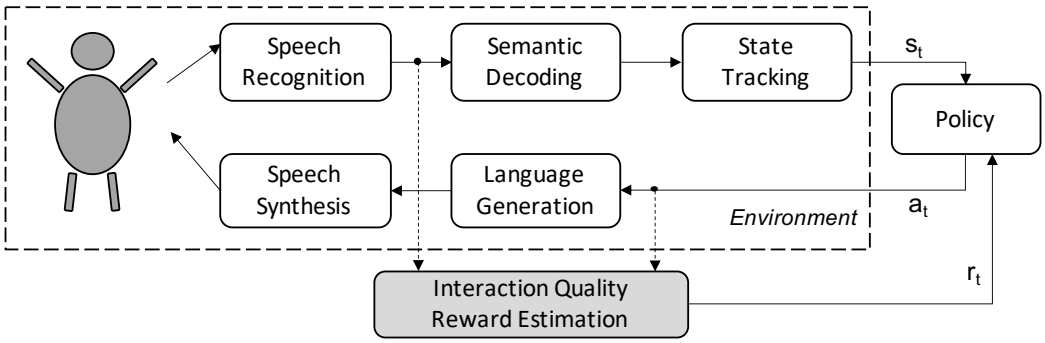

Figure 1: The modular pipeline architecture of a spoken dialogue system with the focus of its application in a reinforcement learning setup and the presented extension of integrating an interaction quality reward estimator as originally proposed by Ultes et al. (2017a). The policy learns to take action a_t at time t while being in state s_t and receiving reward r_t.

The goal of this article 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 (TS) rates. We apply two user satisfaction reward estimators (Ultes et al., 2017a; Ultes, 2019) where one uses a support vector machine (Vapnik, 1995) and relies on handcrafted temporal features. The other uses a deep learning network that uses long short-term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) to learn the temporal dependencies of a multi-turn dialogue interaction implicitly.

We argue that training a system to maximise US is a good alternative to TS for the following reasons:

1. User satisfaction is favourable over task success as it represents more accurately the user's view and thus whether the user is likely to use the system again in the future. In fact, task success has only been used as principal reward component as it has been shown to correlate well with user satisfaction (Williams and Young, 2004).

2. User satisfaction—in contrast to TS—may be linked to interaction phenomena that are independent of the user's goal (Schmitt and Ultes, 2015) 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 generalise well across domains. Thus, learning dialogue policies for new, previously unseen domains becomes much easier.

Following up on previous work (Ultes et al., 2011, 2012; Ultes and Minker, 2014; Ultes et al., 2015, 2019), interaction quality (IQ)—a less subjective version of user satisfaction[1]—will be used for estimating the reward. We apply a conventional IQ estimator (Ultes et al., 2017a) using domain-independent, interaction-related features that do not contain any information about the task or the

1. The relation of US and IQ has been closely investigated by Schmitt and Ultes (2015) and Ultes et al. (2013), see also Section 3.1.

goal of the dialogue. This allows the reward estimator to be applicable for learning in unseen domains. To circumvent the dependency of the convectional estimator on handcrafted temporal features, we will additionally present a deep learning-based IQ estimator (Ultes, 2019) that utilises the capabilities of recurrent neural networks to get rid of all handcrafted features that encode temporal effects. By that, these temporal dependencies may be learned instead.

The applied RL framework is shown in Figure 1. It has previously been applied for in-domain experiments and simulated evaluation (Ultes et al., 2019). Within this setup, both IQ estimators are used for learning dialogue policies in several domains to analyse their impact on general dialogue performance metrics. Moreover, one estimator is used in an experiment where the policy is learned through interaction with real humans.

The remainder of the paper is organised as follows: in Section 2, related work is presented focusing on dialogue learning and the type of reward that is applied. In Section 3, the interaction quality is presented and how it is used in the reward model. The deep learning-based interaction quality estimator is then described in detail in Section 3.3 followed by the experiments and results both of the estimator itself and the resulting dialogue policies in Section 4.

The work described in this article builds upon and extends work published by Ultes (2019) and Ultes et al. (2017a).

2. Relevant Related Work

Most of previous work on dialogue policy learning focuses on employing task success as the main reward signal (Gašić and Young, 2014; Gašić et al., 2014; Lemon and Pietquin, 2007; Daubigney et al., 2012; Levin and Pieraccini, 1997; Singh et al., 2002; Young et al., 2013; Su et al., 2015, 2016). 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 (Gašić et al., 2013). However, this can be intrusive, and users may not always cooperate.

An alternative is to use a task success estimator (El Asri et al., 2014b; Su et al., 2015, 2016). With the right choice of features, such a task success estimator can also be applied to new and unseen domains (Vandyke et al., 2015). 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 the reward (Walker et al., 1998; Walker, 2000; Henderson et al., 2008; Rieser and Lemon, 2008b,a) by using the PARADISE framework initially proposed by Walker et al. (1997). However, PARADISE relies on the existence of explicit task success information, which is usually hard to obtain. Furthermore, 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 (El Asri et al., 2012, 2013) to enable unintrusive acquisition of dialogues. However, the problem of mapping the results of the questionnaire to a scalar reward value still exists.

A similar measure called response quality has been proposed as a measure to capture user satisfaction as an alternative to interaction quality (Bodigutla et al., 2019b,a, 2020). In contrast to

the interaction quality, the response quality focuses more on the overal performance of a system, e.g., including the functionality of back-end services. Thus, it is hard to identify if a negative score should be attributed to the performance of a back-end service or the interaction itself. Thus, the response quality is not suitable for learning the behaviour of a dialogue system.

Other research uses different cues for reward estimation. For instance, Misu et al. (2012) investigate domain-independent dialogue policy learning in the context of question-answering dialogues. Users are asked to rate dialogues on a Likert scale, then regression is used to compute rewards for question-answer pairs. As there is no explicit task success metric (their dialogues are not strictly task-oriented and somewhat similar to chat), their approach is also domain-independent. Shi and Yu (2018) incorporate user sentiment information obtained from multi-modal cues in both a supervised and a reinforcement learning setup. Liu and Lane (2018) forego the need for explicit ratings and rely only on adversarial rewards. In the area of non-task-oriented dialogue, Cuayáhuitl et al. (2018) rely on deep reinforcement learning and investigates the amount of dialogue history that has to be taken into account for the estimation of reward, also without relying on explicit ratings.

In this work, we use interaction quality (Section 3) due to the following reasons: interaction quality models user satisfaction instead of task success that targets the measurement of task completion. Other than automatically derived measures such as adversarial signals, interaction quality provides expert ratings. These expert ratings reflect the quality of the interaction itself, with no dependency on the connected APIs as it is the case for response quality; furthermore, they are easier to obtain than ratings from mixed sources (Shi and Yu, 2018). Interaction quality furthermore uses scalar values applied by experts and only uses task-independent features that are valid across domains and easy to derive, which gives it an edge over the approaches relying on the PARADISE framework.

3. Interaction Quality Reward Estimation

In this work, the reward estimator is based on estimating the interaction quality (IQ) (Schmitt and Ultes, 2015) 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 (Schmitt and Ultes, 2015).

The IQ estimation model will be used as a reward estimator as depicted in Figure 1. With parameters that are collected from the dialogue system modules for each time step t, the reward estimator derives the reward r_t that is used for learning the dialogue policy π.

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

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

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. A per-turn penalty of -1 is added to the dialogue outcome. This results in a reward range of 19 (because there is always at least one turn) down to $-T$, which is consistent with related work (e.g., Gašić and Young, 2014; Vandyke et al., 2015; Su et al., 2016) in which binary task success (TS) was used to define the reward as:

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

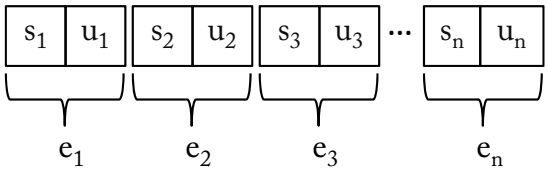

Figure 2: A dialogue may be separated into a sequence of system-user-exchanges where each exchange e_i consists of a system turn s_i followed by a user turn u_i.

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

3.1 Interaction Quality and the LEGO Corpus

Interaction Quality is defined similarly to user satisfaction: while the latter represents the true disposition of the user, IQ is the disposition of the user assumed by an expert annotator. Here, expert annotators are people who listen to recorded dialogues *after* the interactions and rate them by assuming the point of view of the actual person performing the dialogue. These experts are supposed to have some experience with dialogue systems. For the LEGO corpus—the data set used in this work—, expert annotators were "advanced students of computer science and engineering" (Schmitt and Ultes, 2015; Schmitt et al., 2011), i.e., grad students.

Interaction quality has been shown to be a suitable surrogate for user satisfaction by Ultes et al. (2013). Comparing user satisfaction ratings and interaction quality ratings for the same data showed a high correlation between the labels. Estimation models trained on one (user satisfaction or interaction quality) and evaluated on the other result in estimation performances clearly above chance. Furthermore, interaction quality estimation is much more reliable and accurate than user satisfaction estimation. Moreover, the interaction quality matches requirements identified by Ultes et al. (2012) for an estimation approach to be used for online adaptation: exchange-level ratings, automatically derivable and domain-independent input features, a consistent labelling process and reproducible and unbiased labels.

The LEGO corpus (Schmitt et al., 2012b) is based on 200 calls to the "Let's Go Bus Information System" of the Carnegie Mellon University in Pittsburgh (Raux et al., 2006) recorded in 2006. Labels for IQ have been assigned by three expert annotators to 200 calls consisting of 4,885 system-user-exchanges (see Figure 2) in total with an inter-annotator agreement of $\kappa = 0.54$ using Cohen's weighted kappa with a linear weighting function 13. This may be considered as a moderate agreement (cf. Landis and Koch's Kappa Benchmark Scale (1977)), which is quite good considering the difficulty of the task that required to rate each exchange. For instance, if one annotator reduces the IQ value only one exchange earlier than another annotator, both already disagree on two exchanges. The final label of each exchange was derived by using the median of all three individual ratings as the median showed the best overall agreement (Schmitt and Ultes, 2015).

IQ was labelled on a scale from 1 (extremely unsatisfied) to 5 (satisfied) considering the complete dialogue up to the current exchange. Thus, each exchange has been rated without regarding

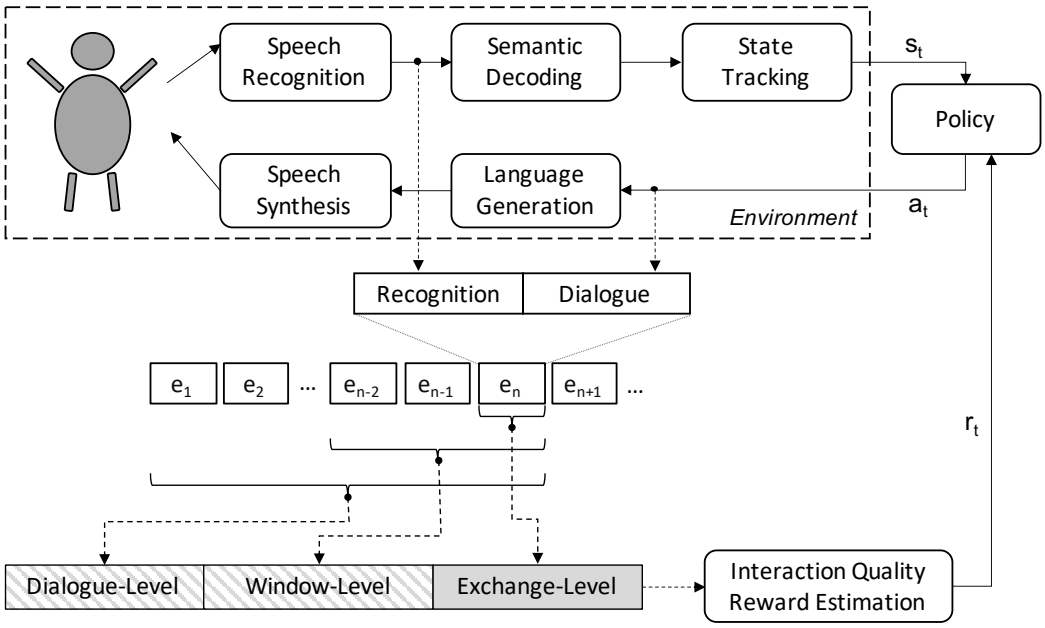

Figure 3: The overall architecture of deriving the interaction parameters from the spoken dialogue system that model the temporal information explicitly and are used as input to the interaction quality reward estimator.

any upcoming user utterance. As the users are expected to be satisfied at the beginning, each dialogue's initial rating is 5. In order to ensure consistent labelling, the expert annotators had to follow distinct labelling guidelines (Schmitt et al., 2012b).

An example of an annotated dialogue of the LEGO corpus is shown in Table 1. The dialogue starts off with a good IQ until the system first attempts to find results. There, the IQ falls drastically as the user input does not correspond to what the system expects and the system remains in a loop until the user reacts appropriately.

3.2 Conventional Interaction Quality Estimation

The problem of estimating IQ has been 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 (see Figure 3): the ASR status (one of *success*, *no match*, *no input*), the ASR confidence of the highest ranked result, the general type of the system action (one of *statement*, *question*), whether the system action is a repetition of the previous system action, and whether the role of the system action is to confirm previous user input. This information, which is available at every turn, is used to compute temporal features by taking sums, means or counts from the turn-based information for a window of the last three system-user-exchanges[2] and the complete

2. A system turn followed by a user turn.

Table 1: Example of a dialogue annotated with Interaction Quality for each system-user-exchange. As the system fails to interact properly with the user in the end, the IQ value drops down to one. Utterance represents either the system utterance or the recognised user utterance. (CallID 2061123000 of the LEGO corpus.)

	Utterance	IQ
S:	Welcome to the CMU Let's Go bus information system.	
U:		5
S:	To get help at any time, just say Help or press zero.	
U:		5
S:	What can I do for you?	
U:	PENN AT BUTLER TIME IS IT IS THE CENTRE AT HIGHLAND FROM RUNS THE 61A EIGHT THIRTY 61B CAN'T OAKLAND	5
S:	Leaving from OAKLAND . Is this correct?	
U:	OAKLAND	4
S:	Leaving from OAKLAND . Is this correct?	
U:	*YES*	4
S:	Right	
U:	-	4
S:	Where do you want to go?	
U:	*SQUIRREL HILL OF FIFTH AVENUE*	4
S:	Going to FIFTH AVENUE . Is this correct?	
U:	*YES*	5
S:	Alright	
U:	-	5
S:	When do you wanna travel?	
U:	*SUNDAY*	4
S:	At what time do you wanna travel?	
U:	AT FIVE FOUR VERY ABOUT ELEVEN O'CLOCK	4
S:	*Leaving at 11 a.m. . Is this correct?*	
U:	*YES*	5
S:	Right	
U:	-	5
S:	Hold on. Let me check that for you.	
U:	-	5
S:	OAKLAND and FIFTH AVENUE are both the same stop. Please provide a different start or end point.	
U:	-	3
S:	To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U:	*61A*	2
S:	You can say start a new query, or goodbye.	
U:	*FIVE DUQUESNE*	1
S:	To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U:	*LAST THE NEXT*	1
S:	You can say start a new query, or goodbye.	
U:	*START WH*	1
S:	To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U:	*THE*	1
S:	You can say start a new query, or goodbye.	
U:	*THANK YOU QUERY ELEVEN BACK*	1
S:	To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U:	*GOODBYE*	1
S:	Thank you for using the cmu Let's Go Bus Information System. Goodbye.	
U:	-	1

dialogue (see lower part of Figure 3). The baseline IQ estimation approach as applied by Ultes et al. (2017a) (and originating from Ultes et al. (2015)) uses a feature set of 16 parameters as shown in Table 2 to train a support vector machine (SVM) (Vapnik, 1995; Chang and Lin, 2011) with linear kernel.

Table 2: The interaction parameters extracted from each user input (exchange level) plus counts, sums and rates for the whole dialogue (#, %, Mean) and for a window ($\{\cdot\}$) of the last 3 turns.

	Parameter	Description
Exchange level	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?
Dialogue level	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
Window level	{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

The LEGO corpus (Schmitt et al., 2012a) provides data for training and testing and consists of 200 dialogues (4,885 turns) 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.

Previous work has used the LEGO corpus with a full IQ feature set (which includes additional partly domain-related information) achieving an unweighted average recall[3] (UAR) 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).

3.3 LSTM-based Interaction Quality Estimation

The architecture of the LSTM-based IQ estimation model is shown in Figure 4. It is based on the idea that the temporal information may be learned by using recurrent neural networks instead of encoding it explicitly with the window and dialogue interaction parameter levels as used by the conventional estimator. Thus, only the exchange level parameters $\mathbf{e}_t$ are considered (see Table 2). Long Short-Term Memory (LSTM) cells are at the core of the model and have originally been proposed by Hochreiter and Schmidhuber (1997) as a recurrent variant that remedies the vanishing gradient problem (Bengio et al., 1994).

3. UAR is the arithmetic average of all class-wise recalls.

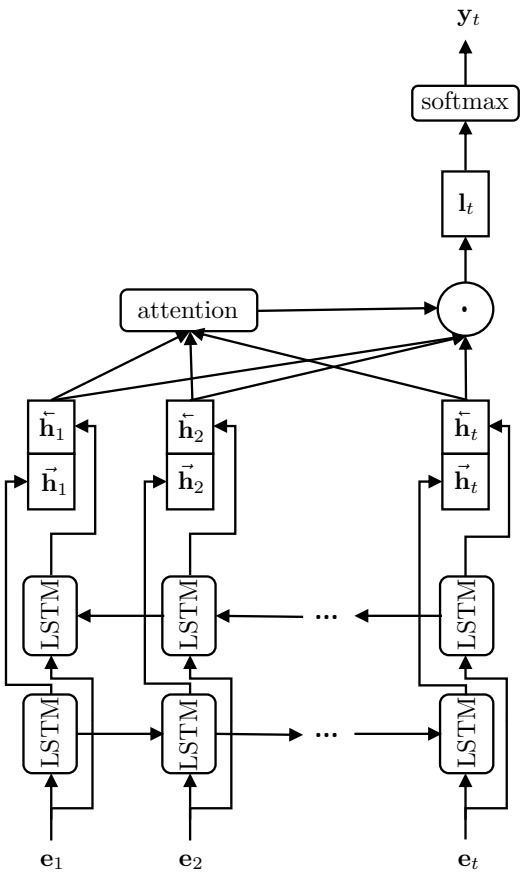

Figure 4: The architecture of the proposed BiLSTM model with self attention. For each time t, the exchange level parameter of all exchanges $\mathbf{e}_i$ of the sub-dialogue $i \in \{1 \dots t\}$ are encoded to their respective hidden representation $\mathbf{h}_i$ and are considered and weighted with the self attention mechanism to finally estimate the IQ value $\mathbf{y}_t$ at time t.

As shown in Figure 4, the exchange level parameters form the input vector $\mathbf{e}_t$ for each time step or turn t to a bi-directional LSTM (Graves et al., 2013) layer. The input vector $\mathbf{e}_t$ encodes the nominal parameters ASRRecognitionStatus, ActivityType, and Confirmation? as 1-hot representations. In the BiLSTM layer, two hidden states are computed: $\vec{\mathbf{h}}_t$ constitutes the forward pass through the current sub-dialogue and $\overleftarrow{\mathbf{h}}_t$ the backwards pass:

$$\vec{\mathbf{h}}_t = \mathrm{LSTM}(\mathbf{e}_t, \vec{\mathbf{h}}_{t-1}) \tag{3}$$

$$\overleftarrow{\mathbf{h}}_t = \mathrm{LSTM}(\mathbf{e}_t, \overleftarrow{\mathbf{h}}_{t+1}) \tag{4}$$

The final hidden layer is then computed by concatenating both hidden states:

$$\mathbf{h}_t = [\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t] \ . \tag{5}$$

Even though information from all time steps may contribute to the final IQ value, not all time steps may be equally important. Thus, an attention mechanism (Vaswani et al., 2017) is used that

evaluates the importance of each time step t' for estimating the IQ value at time t by calculating a weight vector $\alpha_{t,t'}$.

$$\mathbf{g}_{t,t'} = \tanh(\mathbf{h}_t^T \mathbf{W}_t + \mathbf{h}_{t'}^T \mathbf{W}_{t'} + \mathbf{b}_t) \tag{6}$$

$$\boldsymbol{\alpha}_{t,t'} = \mathrm{softmax}(\sigma(\mathbf{W}_a \mathbf{g}_{t,t'} + \mathbf{b}_a)) \tag{7}$$

$$\mathbf{l}_t = \sum_{t'} \boldsymbol{\alpha}_{t,t'} \mathbf{h}_{t'} \tag{8}$$

Zheng et al. (2018) describe this as follows: "The attention-focused hidden state representation $\mathbf{l}_t$ of an [exchange] at time step t is given by the weighted summation of the hidden state representation $\mathbf{h}_{t'}$ of all [exchanges] at time steps t', and their similarity $\boldsymbol{\alpha}_{t,t'}$ to the hidden state representation $\mathbf{h}_t$ of the current [exchange]. Essentially, $\mathbf{l}_t$ dictates how much to attend to an [exchange] at any time step conditioned on their neighbourhood context."

To calculate the final estimate $\mathbf{y}_t$ of the current IQ value at time t, a softmax layer is introduced:

$$\mathbf{y}_t = \mathrm{softmax}(\mathbf{l}_t) \tag{9}$$

For estimating the interaction quality using a BiLSTM, the proposed architecture frames the task as a classification problem where each sequence is labelled with one IQ value. Thus, for each time step t, the IQ value needs to be estimated for the corresponding sub-dialogue consisting of all exchanges from the beginning up to t. Framing the problem like this is necessary to allow the application of a BiLSTM-approach while still being able to only use information that would be present at the current time step t in an ongoing dialogue interaction.

To analyse the influence of the BiLSTM, a model with a single forward-LSTM layer is also investigated where

$$\mathbf{h}_t = \vec{\mathbf{h}}_t \ . \tag{10}$$

Similarly, a model without attention is also analysed where

$$\mathbf{l}_t = \mathbf{h}_t \ . \tag{11}$$

A deep learning approach using only non-temporal features has been previously proposed and achieved an UAR on the full feature set of 0.55 (Rach et al., 2017).

4. Simulation Experiments and Results

The IQ estimators are both trained and evaluated on the LEGO corpus and applied within the IQ reward estimation framework (Figure 1) on several domains within a simulated environment.

4.1 Interaction Quality Estimation

To evaluate the BiLSTM model with attention (BiLSTM+att), it is compared with three of its own variants: a BiLSTM without attention (BiLSTM) as well as a single forward-LSTM layer with attention (LSTM+att) and without attention (LSTM). An additional baseline is defined by Rach et al. (2017) who already proposed an LSTM-based architecture that only uses non-temporal features. Furthermore, the conventional IQ estimator using a linear SVM is evaluated as originally used for reward estimation by Ultes et al. (2015).

Table 3: Performance of the proposed LSTM-based variants with the traditional cross-validation setup. Due to overlapping sub-dialogues in the train and test sets, the performance of the LSTM-based models achieve unrealistically high performance. Significant differences are observed between BiLSTM and all other variants/models ($p < 0.05$, Wilcoxon signed-rank test (Wilcoxon, 1945)).

	UAR	κ	ρ	*eA*	*Ep.*
LSTM	0.78	0.85	0.91	**0.99**	101
BiLSTM	**0.78**	**0.85**	**0.92**	**0.99**	100
LSTM+att	0.74	0.82	0.91	**0.99**	101
BiLSTM+att	0.75	0.83	0.91	**0.99**	93
Rach et al. (2017)	0.55	0.68	0.83	0.94	-
Ultes et al. (2015)	0.55	-	-	0.89	-

The deep neural net models have been implemented with Keras[4] using the self-attention implementation as provided by Zheng et al. (2018)[5]. The input vector consists of one-hot-encodings of the three nominal features ASRRecognitionStatus, ActivityType, and Confirmation? and the numerical features ASRConfidence and RePrompt? resulting in a vector of size 11. The target IQ labels are also one-hot-encoded. The LSTM and BiLSTM embeddings have a dimension of 64 and 128, respectively. The maximum dialogue length has been set to 60: dialogues with fewer turns are padded wiht zero vectors and dialogues with more turns are truncated. All models were trained against cross-entropy loss using RmsProp (Tieleman and Hinton, 2012) optimisation with a learning rate of 0.001 and a mini-batch size of 32.

Interaction Quality estimation is evaluated by using three commonly applied evaluation metrics: *Unweighted Average Recall (UAR)*, *Cohen's Kappa* (Cohen, 1960), and *Spearman's Rho* (Spearman, 1904) comparing the estimated IQ ratings x with the true IQ ratings y. As missing the correct estimated IQ value by only one has little impact for modelling the reward, a measure we call the *extended accuracy (eA)* is used where neighbouring values are taken into account as well.

Recall in general is defined as the rate of correctly classified samples belonging to one class. The **unweighted average recall** for multi-class classification problems with C classes is computed by the class-wise recalls $recall_c$ for each class c and then averaged over all class-wise recalls:

$$UAR = \frac{1}{C} \sum_{c=1}^{N} recall_c \ . \tag{12}$$

Cohen's Kappa measures the relative agreement between two corresponding sets of ratings, here the estimate x and ground truth y. In our case, Cohen's *weighted* kappa is applied as ordinal scores are compared (Cohen, 1968):

$$\kappa = 1 - \frac{\sum_{c=1}^{C} \sum_{k=1}^{C} w_{ck} \cdot m_{ck}}{\sum_{c=1}^{C} \sum_{k=1}^{C} w_{ck} \cdot \frac{m_{c.} \cdot m_{.k}}{N}} \ . \tag{13}$$

4. https://keras.io/
5. Code freely available at https://github.com/CyberZHG/keras-self-attention

$m_{c,k}$ is the number of samples where, for a corresponding (x_i, y_i) pair, the estimator predicted k for the true class c, and N is the total number of samples. $m_{c.}$ represents the sum of all estimated class samples for true class c and $m_{.k}$ the sum of all true class samples for estimated class k. A weighting factor w is introduced reducing the discount of disagreements the smaller the difference is between two ratings:

$$w_{ck} = \frac{|r_c - r_k|}{|r_{max} - r_{min}|} \,. \tag{14}$$

Here, r_c and r_k denote the rating pair and r_{max} and r_{min} the maximum and minimum ratings possible.

The correlation of two variables describes the degree by that one variable can be expressed by the other. **Spearman's Rank Correlation Coefficient** is a non-parametric method assuming a monotonic function between the two variables (Spearman, 1904). It is defined by

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \,, \tag{15}$$

where x_i and y_i are corresponding ranked ratings and $\bar{x}$ and $\bar{y}$ the mean ranks. Thus, two sets of ratings can have total correlation even if they never agree. This would happen if all ratings are shifted by the same value, for example.

The **extended accuracy** is computed similar to regular accuracy. However, instead of only using the main diagonal of the confusion matrix, the two secondary diagonals are considered additionally. For N classes, the extended accuracy is computed by

$$eA = \frac{\sum_{c=1}^{C} m_{c,c} + m_{c,c-1} + m_{c,c+1}}{\sum_{c=1}^{N} \sum_{k=1}^{N} m_{c,k}} \,, \tag{16}$$

where $m_{c,k}$ is the number of samples where the estimator predicted k for the true class c. Thus, neighbouring values are taken into account as well and being off by one is still considered as correct estimation[6].

All experiments were conducted with the LEGO corpus (Schmitt et al., 2012a) in a 10-fold cross-validation setup for a total of 100 epochs per fold. The results are presented in Table 3. Due to the way the task is framed (one label for each sub-dialogue), memorising effects may be observed with the traditional cross-validation setup that has been used in previous work. Hence, the results in Table 3 show very high performance, which is likely to further increase with ongoing training. However, the corresponding models are likely to generalise poorly.

To alleviate this, a dialogue-wise cross-validation setup has been employed also consisting of 10 folds of disjoint sets of dialogues. By that, it can be guaranteed that there are no overlapping sub-dialogues in the training and test sets. All results of these experiments are presented in Table 4 with the absolute improvement of the two main measures UAR and eA over the SVM-based conventional approach of Ultes et al. (2015) visualised in Figure 5.

The BiLSTM+att model outperforms the other models and baselines in all four performance measures by achieving an UAR of 0.54 and an eA of 0.94 after 40 epochs. Furthermore, both the BiLSTM and the attention mechanism by themselves improve the performance in terms of UAR.

Table 4: Performance of the proposed LSTM-based variants with the dialogue-wise cross-validation setup. The models by Rach et al. (2017) and Ultes et al. (2015) have been re-implemented. The BiLSTM with attention mechanism performs best in all evaluation metrics. All results are significantly different to each other ($p < 0.05$, Wilcoxon signed-rank test (Wilcoxon, 1945)).

	UAR	κ	ρ	*eA*	*Ep.*
LSTM	0.51	0.63	0.78	0.93	8
BiLSTM	0.53	0.63	0.78	0.93	8
LSTM+att	0.52	0.63	0.79	0.92	40
BiLSTM+att	**0.54**	**0.65**	**0.81**	**0.94**	40
Rach et al. (2017)	0.45	0.58	0.79	0.88	82
Ultes et al. (2015)	0.44	0.53	0.69	0.86	-

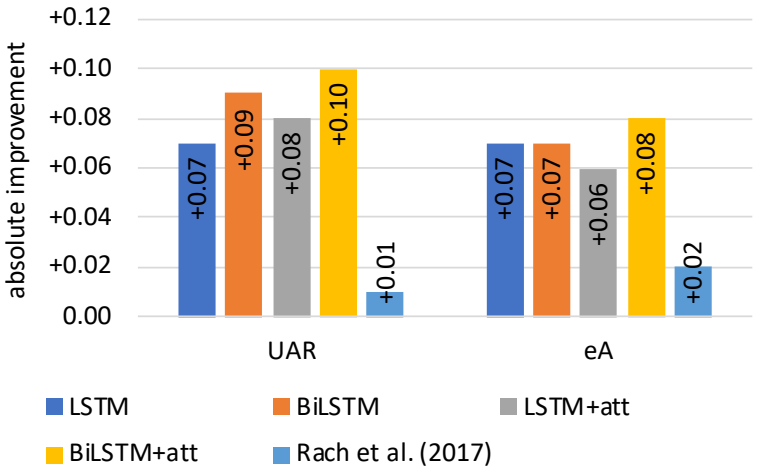

Figure 5: Absolute improvement of the IQ estimation models over the conventional model originally proposed by (Ultes et al., 2015) for IQ-based reward estimation with the dialogue-wise cross-validation setup. UAR and eA take values from 0.0 to 1.0.

4.2 Dialogue Policy Learning

To analyse the impact of the IQ reward estimator on the resulting dialogue policy, experiments are conducted comparing three different reward models. The baseline is in accordance to Ultes et al. (2017a): having the objective task success as principal reward component (R_{TS}). It will be compared to the conventional estimator having the interaction quality estimated by a support vector machine as principal reward component (R^{s}_{IQ}) and to the BiLSTM+att model to estimate the interaction quality used as principal reward component (R^{bi}_{IQ}). TS can be computed by comparing

6. For the bounds, the respective non-existing $m_{c,k}$ is omitted.

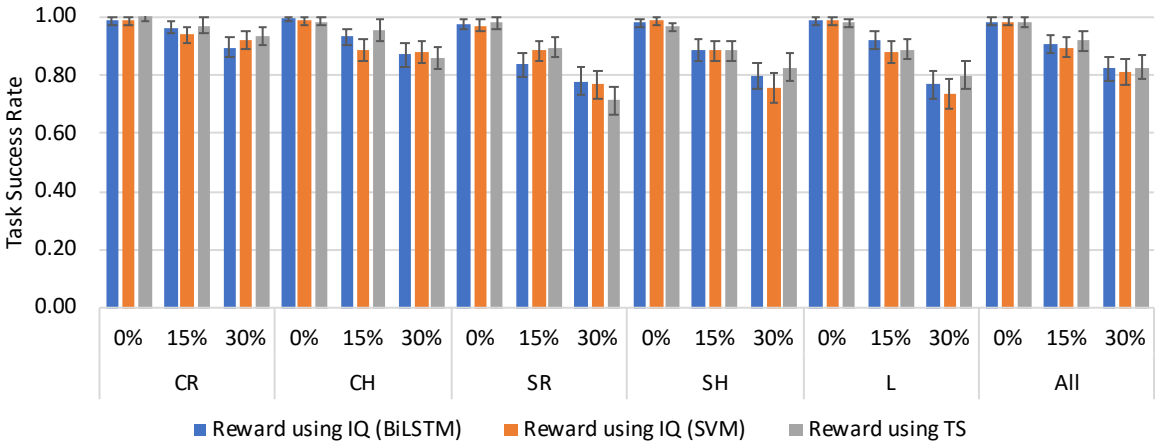

Figure 6: Results using GP-SARSA in task success rate (TSR) of the simulated experiments for all domains and semantic error rates 0%, 15%, and 30%. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials. Numerical results with significance indicators are shown in Table 6.

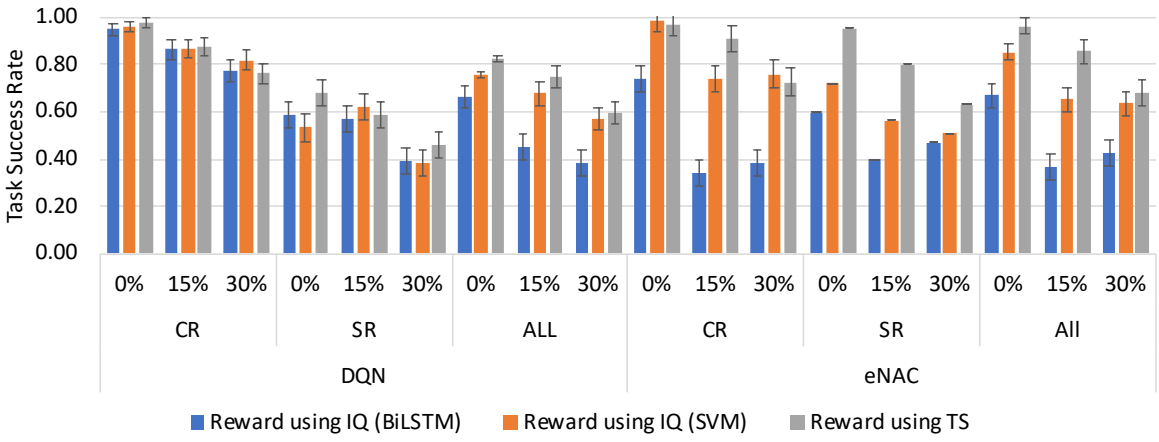

Figure 7: Results using DQN and eNAC in task success rate (TSR) of the simulated experiments for CamRestaurants and SFRestaurants and semantic error rates 0%, 15%, and 30%. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials. Numerical results with significance indicators are shown in Table 8 and Table 7.

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 estimators, nor is it required.

For learning the dialogue behaviour, three policy models are applied. The first is a policy model based on the GP-SARSA algorithm (Gašić and Young, 2014), which is a value-based method that

Table 5: Statistics of the domains the interaction quality estimators are trained on (LetsGo) and applied to (rest).

Domain	*Code*	*# constraints*	*# DB items*	*state size*
LetsGo		4	-	-
CamRestaurants	CR	3	110	268
CamHotels	CH	5	33	111
SFRestaurants	SR	6	271	636
SFHotels	SH	6	182	438
Laptops	L	6	126	204

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 (Gašić et al., 2013).

Additionally, two deep reinforcement learning algorithms are applied: Deep Q-Network (DQN) (Mnih et al., 2015) and episodic natural actor critic (eNAC) (Su et al., 2017). Similar to the GP-SARSA, the DQN also approximates the state-value function also known as Q-function. The eNAC directly learns the policy using a policy gradient approach in an actor-critic framework.

The decisions of the policy are based on a summary space representation of the dialogue state tracker. In this work, the focus tracker (Henderson et al., 2014)—an effective rule-based tracker—is used. The state space and summary state space follow previous work (e.g., Gašić et al., 2013) and comprise multiple dimensions: for each *informable* slot one probability distribution over all slot values plus the special values DONTCARE and NONE, for each *requestable* slot a Bernoulli distribution indicating whether the slot has been requested by the user, a probability distribution over the query method (e.g., search by constraints, search by name), and a status vector of the search results given the current state. *Informable* slots contain all information that the user has provided to the system as search constraints during the interaction. *Requestable* slots are usually a superset of the informable slots and contain additionally information that is part of the search result (e.g., the phone number). To map this dialogue state to a summary space, each of the informable slot probability distribution vectors are sorted (excluding NONE). This follows the idea that for making a decision, the actual slot value is not important, only how the probabilities are distributed over all values.

For each dialogue decision, the policy chooses exactly one summary action out of a set of summary actions, which are based on general dialogue acts like *request*, *confirm* or *inform*. The exact number of system actions varies for the domains and ranges from 16 to 25.

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 values of the respective model (AIQ^{bi}) for the BiLSTM and AIQ^{s} for the SVM) at the end of each dialogue. As there are two IQ estimators, a distinction is made between AIQ^{s} and AIQ^{bi}. Additionally, the average dialogue length (ADL) is reported.

Table 6: Results of the **simulated experiments for all domains using GP-SARSA** showing task success rate (TSR), average interaction quality estimated with the SVM (AIQ^s) and the BiLSTM (AIQ^{bi}), and average dialogue length (ADL) in number of turns. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials with different random seeds. ^{1,2,3} marks statistically significant difference compared to R_{TS}, to R^s_{IQ}, and to R^{bi}_{IQ}, respectively ($p < 0.05$, T-test for TSR and ADL, Mann-Whitney-U test for AIQ).

Domain	*SER*	*TSR*			AIQ^s		AIQ^{bi}		*ADL*		
		R_{TS}	R^s_{IQ}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R_{TS}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R^{bi}_{IQ}
CR	0%	**1.00**^{2,3}	0.99^{1}	0.99^{1}	3.64^{2}	**3.90**^{1}	3.68^{3}	**3.83**^{1}	4.68	4.88	**4.59**
	15%	**0.97**	0.94	0.96	3.35^{2}	**3.65**^{1}	3.45^{3}	**3.63**^{1}	5.85^{3}	5.33	**5.10**^{1}
	30%	**0.94**	0.92	0.90	3.15^{2}	**3.34**^{1}	3.22	**3.30**	6.34	6.30	**6.25**
CH	0%	0.98	0.99	**0.99**	3.26^{2}	**3.62**^{1}	3.33	**3.44**	5.71	5.61	**5.40**
	15%	**0.96**^{2}	0.89^{1,3}	0.93^{2}	**2.90**	2.88	**3.14**	3.14	**6.28**^{2}	7.26^{1,3}	6.31^{2}
	30%	0.86	**0.88**	0.87	2.38^{2}	**2.79**^{1}	2.79^{3}	**3.02**^{1}	7.94^{3}	7.31	**6.99**^{1}
SR	0%	**0.98**	0.97	0.98	3.04^{2}	**3.53**^{1}	3.13^{3}	**3.37**^{1}	6.26	6.03	**5.80**
	15%	**0.90**^{3}	0.88	0.84^{1}	2.40^{2}	**3.00**^{1}	2.85^{3}	**3.01**^{1}	7.99	7.55	**7.33**
	30%	0.71	0.77	**0.78**	2.03^{2}	**2.52**^{1}	2.46^{3}	**2.78**^{1}	9.77^{3}	9.41	**8.50**^{1}
SH	0%	0.97	**0.99**	0.98	3.15^{2}	**3.52**^{1}	3.17^{3}	**3.36**^{1}	5.99^{2}	**5.50**^{1}	5.76
	15%	0.88	0.88	**0.89**	2.63^{2}	**2.94**^{1}	2.77^{3}	**3.17**^{1}	7.98^{3}	7.59^{3}	**6.63**^{1,2}
	30%	**0.83**^{2}	0.76^{1}	0.80	2.50	**2.63**	2.70^{3}	**2.87**^{1}	8.38	9.21	**8.37**
L	0%	0.98	**0.99**	**0.99**	3.26^{2}	**3.61**^{1}	3.28	**3.41**	5.78	**5.44**	5.60
	15%	0.89	0.88	**0.92**	2.58^{2}	**2.97**^{1}	2.92^{3}	**3.17**^{1}	7.19	7.34	**6.73**
	30%	**0.80**	0.74	0.77	2.43	**2.57**	2.79	**2.92**	8.22^{2}	9.32^{1,3}	**7.97**^{2}
All	0%	0.98	0.98	**0.98**	3.23^{2}	**3.65**^{1}	3.31	**3.48**	5.76	5.50	**5.47**
	15%	**0.92**	0.89	0.91	2.76^{2}	**3.10**^{1}	3.02^{3}	**3.20**^{1}	7.13	7.06	**6.52**
	30%	**0.83**	0.81	0.82	2.49	**2.80**	2.78	**2.97**	8.20^{2}	8.23^{1,3}	**7.66**^{2}

For the simulation experiments with the GP-SARSA, the performance of the trained policies 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 5 and compared to the LetsGo domain (the domain the estimators have been trained on). The DQN and eNAC are evaluated only on the two domains Cambridge (most simple) and San Francisco Restaurants (most complex).

The dialogues were created using the publicly available spoken dialogue system toolkit PyDial (Ultes et al., 2017b)[7], which contains implementations of all the applied policy models[8] and an implementation of the agenda-based user simulator (Schatzmann and Young, 2009) with an ad-

7. Code freely available at `http://www.pydial.org`
8. Please refer to PyDial code for details about the exact model implementation.

Table 7: Results of the **simulated experiments for CR and SFR using DQN** showing task success rate (TSR), average interaction quality estimated with the SVM (AIQ^s) and the BiLSTM (AIQ^{bi}), and average dialogue length (ADL) in number of turns. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials with different random seeds. ^{1,2,3} marks statistically significant difference compared to R_{TS}, to R^s_{IQ}, and to R^{bi}_{IQ}, respectively ($p < 0.05$, T-test for TSR and ADL, Mann-Whitney-U test for AIQ).

Domain	*SER*	*TSR*			AIQ^s		AIQ^{bi}		*ADL*		
		R_{TS}	R^s_{IQ}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R_{TS}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R^{bi}_{IQ}
CR	0%	**0.98**^{3}	0.96	0.95^{1}	**3.69**^{2}	3.17^{1}	**3.52**	3.51	**4.32**^{2}	4.68^{1}	4.53
	15%	**0.87**	0.87	0.86	3.05	**3.11**	3.39	**3.42**	5.37	5.32	**5.14**
	30%	0.76	**0.82**	0.77	2.65^{2}	**2.94**^{1}	**3.34**	3.28	6.04	**5.59**	5.74
SR	0%	**0.68**^{2,3}	0.53^{1}	0.59^{1}	**2.57**^{2}	2.01^{1}	**3.10**	3.07	**6.29**	6.56	6.56
	15%	0.59	**0.62**	0.57	**2.20**	2.18	**3.10**	3.08	**6.58**^{2}	7.10^{1}	6.85
	30%	**0.46**^{2}	0.38^{1}	0.39	**1.85**	1.71	3.01^{3}	**3.08**^{1}	**7.55**^{2}	8.79^{1,3}	7.66^{2}
All	0%	**0.83**^{3}	0.75	0.77^{1}	**3.13**^{2}	2.59^{1}	**3.31**	3.29	**5.30**^{2}	5.62^{1}	5.54
	15%	0.73	**0.75**	0.72	2.63	**2.65**	3.25	**3.25**	**5.98**	6.21	5.99
	30%	**0.61**	0.60	0.58	2.25^{2}	**2.32**^{1}	3.17	**3.18**	6.79	7.19	**6.70**

Table 8: Results of the **simulated experiments for CR and SFR using eNAC** showing task success rate (TSR), average interaction quality estimated with the SVM (AIQ^s) and the BiLSTM (AIQ^{bi}), and average dialogue length (ADL) in number of turns. Each value is computed after 100 evaluation / 1,000 training dialogues averaged over three trials with different random seeds. ^{1,2,3} marks statistically significant difference compared to R_{TS}, to R^s_{IQ}, and to R^{bi}_{IQ}, respectively ($p < 0.05$, T-test for TSR and ADL, Mann-Whitney-U test for AIQ).

Domain	*SER*	*TSR*			AIQ^s		AIQ^{bi}		*ADL*		
		R_{TS}	R^s_{IQ}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R_{TS}	R^{bi}_{IQ}	R_{TS}	R^s_{IQ}	R^{bi}_{IQ}
CR	0%	0.97^{3}	**0.99**^{3}	0.74^{1,2}	3.31^{2}	**3.65**^{1}	**3.45**	3.41	4.78^{3}	**4.78**^{3}	5.82^{1,2}
	15%	**0.91**^{2,3}	0.74^{1,3}	0.34^{1,2}	**3.25**	3.16	**3.46**^{3}	3.18^{1}	**5.18**^{1,2}	7.72^{1,3}	8.71^{1,2}
	30%	0.73^{3}	**0.76**^{3}	0.39^{1,2}	2.78	**2.98**	**3.29**^{3}	3.14^{1}	**6.49**^{1,2}	7.23^{1,3}	8.77^{1,2}
SR	0%	**0.95**^{2,3}	0.72^{1,3}	0.60^{1,2}	2.52	**2.65**	**3.21**^{3}	3.11^{1}	**5.38**^{2,3}	7.40^{1,3}	8.47^{1,2}
	15%	**0.80**^{2,3}	0.57^{1,3}	0.40^{1,2}	2.23	**2.35**	**3.13**^{3}	3.07^{1}	**6.67**^{2,3}	9.39^{1,3}	7.90^{1,2}
	30%	**0.63**^{2,3}	0.51^{1}	0.47^{1}	1.71^{2}	**1.94**^{1}	**3.03**	**3.03**	**8.60**^{2,3}	10.01^{1}	9.38^{1}
All	0%	**0.96**^{3}	0.85^{3}	0.67^{1,2}	2.92^{2}	**3.15**^{1}	**3.33**	3.26	**5.08**^{3}	6.09^{3}	7.15^{1,2}
	15%	**0.86**^{2,3}	0.65^{1,3}	0.37^{1,2}	2.74	**2.76**	**3.29**^{3}	3.13^{1}	**5.93**^{2,3}	8.56^{1,3}	8.31^{1,2}
	30%	**0.68**^{3}	0.64^{3}	0.43^{1,2}	2.25	**2.46**	**3.16**^{3}	3.09^{1}	**7.54**^{2,3}	8.62^{1,3}	9.08^{1,2}

ditional error model. The error model simulates the required semantic error rate (SER) caused in a real system by the noisy speech channel. For each domain, all three reward models are compared on three SERs: 0%, 15%, and 30%. More specifically, the applied evaluation environments are based on Env. 1, Env. 3, and Env. 6, respectively, as defined by Casanueva et al. (2017). These environments also contain all parameters used for the training of GP-SARSA, DQN and eNAC implementations of PyDial.

For each domain and for each SER, policies have been trained using 1,000 dialogues followed by an evaluation step of 100 dialogues. The task success rates for GP-SARSA in Figure 6 and DQN and eNAC in Figure 7 with exact numbers shown in Table 6, Table 7, and Table 8, respectively, were computed based on the evaluation step averaged over three train/evaluation cycles with different random seeds.

The results of training the GP-SARSA with the SVM IQ reward estimator are similar in terms of TSR for R^s_{IQ} and R_{TS} in all domains for an SER of 0%. This finding is even stronger when comparing R^{bi}_{IQ} and R_{TS}. These high TSRs are achieved while having the dialogues of both IQ-based models result in higher AIQ values compared to R_{TS} throughout the experiments. Of course, only the IQ-based model is aware of the IQ concept and indeed is trained to optimise it.

For higher SERs, the TSRs lightly degrade for the IQ-based reward estimators. However, there seems to be a tendency that the TSR for R^{bi}_{IQ} is more robust against noise compared to R^s_{IQ} while still resulting in better AIQ values.

Finally, even though the differences are mostly not significant, there is also a tendency for R^{bi}_{IQ} to result in shorter dialogues compared to both R^s_{IQ} and R_{TS}.

The results of training a DQN or an eNAC policy model are different, though. While the DQN shows similar TSRs for all SERs and reward models in the CR domain, R_{TS} clearly shows best performance in the SFR domain. Furthermore, even when using an IQ-based reward model, the respective *AIQ* does not result in higher scores than using R_{TS} which shows the stability problems that come with the usage of deep reinforcement learning. The eNAC policy model is even more prone to these effects having R_{TS} always resulting in the highest TSR with R^s_{IQ} and R^{bi}_{IQ} performing rather poorly.

5. Analysis of Learned Behaviour

To further analyse the learning behaviour of the different reward estimators and policy models and thus to gain deeper insights, the similarity scoring framework (Ultes and Maier, 2020) is applied. It uses a standardised setup to feed a fixed set of dialogue states into each trained policy model and compares the resulting system actions and quantifies their similarities. Three different metrics are computed: the total match rate (TMR), the dialogue act match rate (DMR) and the concept match rate (CMR).

A similarity score is computed for the comparison of two behavioural models π and π'. Depending on the nature of the behavioural model, for each context $c_i \in C$, each may produce an abstract system response actions a_i, and an text response p_i. Each abstract system action $a_i = act_i(s^i_1 = v^i_1, \ldots, s^i_j = v^i_j)$ consists of a dialogue act act_i, representing the communicative function like *inform* or *request*, and a set S_i of j slot-value-pairs $S_i = \{(s^i_1, v^i_1), \ldots, (s^i_j, v^i_j)\}$

representing the concepts and their respective values[9]. To compute each similarity score, $|C|$ action/text response pairs are compared using the following similarity score measures.

According to Ultes and Maier (2020), the metrics are defined as follows: the **total match rate (TMR)** is based on a binary score that regards two actions a, a' as equal only if they completely match, i.e., $\delta_{a,a'} = 1$ iff $a = a'$, else 0. The TMR is then defined by

$$TMR = \frac{1}{|C|} \sum_{i=1}^{|C|} \delta_{a_i, a'_i} \ . \tag{17}$$

The **dialogue act match rate (DMR)** is based on a binary score comparing the actions a, a' where both match if the corresponding dialogue acts are the same: $\delta_{act,act'} = 1$ iff $act = act'$, else 0. The DMR is defined by

$$DMR = \frac{1}{|C|} \sum_{i=1}^{|C|} \delta_{act_i, act'_i} \ . \tag{18}$$

The symmetric **concept match rate (CMR)** counts concepts γ that are present in both dialogue actions where $\tilde{m}(a, a', \gamma)$ defines if a match occurred:

$$\tilde{m}(a, a', \gamma) = \begin{cases} 1, \text{if } \gamma \in S \text{ and } \gamma \in S' \\ 0, \text{otherwise} \ . \end{cases} \tag{19}$$

The concept match CM takes into account the dialogue acts and the unified set of concepts $\tilde{S} = S_1 \cup S_2$ of both dialogue actions treating slots s and values v in $\tilde{S}$ as individual γ:

$$\tilde{CM}(a, a') = \delta_{act,act'} + \sum_{\in \tilde{S}} \tilde{m}(a, a', \gamma) \tag{20}$$

A concept match of two dialogue actions a and a' is thus defined by

$$CM(a, a') = \frac{\tilde{CM}(a, a')}{1 + |\tilde{S}|} \tag{21}$$

and the concept match rate by

$$CMR = \frac{1}{|C|} \sum_{i=1}^{|C|} CM(a, a') \ . \tag{22}$$

In short, the TMR only evaluates to true in case of a complete match of both actions. The DMR evaluates to true already if only the dialogue acts of the system actions are equal. The CMR counts concepts that are present in both dialogue actions and normalises this count by the total number of distinct concepts in both actions. Using this analysis setup, the following questions are addressed within this section:

1. How dependent is the resulting policy on the chosen random seed?

9. For the abstract system action a = inform(name='Golden House', area=centre), act = inform and S = {(name,'Golden House'),(area,centre)}.

2. How similar are learned policies across different reward models?

3. How similar are learned policies across different policy models?

For analysing the learned behaviour and computing similarity scores within and between learning setups, the CamRestaurants domain is used.

5.1 Consistency of Policies

As each combination of reward model, policy model and semantic error rate has been trained with three random seeds, this section addresses the question about the consistency of the resulting learned behaviour. Thus, for each setup, the policy of each random seed has been evaluated with the dialogue states from 100 evaluation dialogues taken from one of the policies. A different set of states is used for each noise level. Each random seed's policy is then compared with the policy of the two other random seeds resulting in a total of three comparisons. For each policy model and reward model, the similarity scores of these three comparisons are averaged. The full scores are shown in the appendix in Table 12.

The similarity scores shown in Table 9 show the average scores for TMR, DMR, and CMR. In the left table, the scores are averaged over the different reward models to identify which policy model offers the most consistency during training, i.e., ends up in similar behaviour independent of the random seed. Overall, eNAC shows the most consistency in all three metrics. For 0% SER, eNAC seems to be less dependent on the random seed. The GP-SARSA shows average consistency with the worst CMR for 0% and 15% SER.

On the right side of Table 9, the scores are averaged over the policy models to identify which reward model offers the most consistency during training. Here, R_{TS} shows overall good performance but also R^{s}_{IQ} and R^{bi}_{IQ} are not fare off and in some cases even more consistent than R_{TS}, e.g., TMR with 30% SER.

5.2 Similarity between Reward Models

To gain deeper understanding of how similar the behaviour of the trained policies are that originate from the different reward models, this section compares policies from different reward models resulting in pair-wise comparisons of the policy of each random seed of one reward model with the policies of each random seed of another reward model within one policy model. With three random seeds each, this results in nine comparisons. As in the previous section, each policy is evaluated with 100 evaluation dialogues taken from one of the policies. For each policy model and reward model comparison, the similarity scores are averaged. The full scores are shown in the appendix in Table 13.

The similarity scores shown in Table 10 show the average scores for TMR, DMR, and CMR. In the left table, the scores are averaged over the different reward model pairs to identify which policy model results in the most similar behaviour during training, i.e., ends up in similar behaviour independent of the reward model. Overall, DQN shows the highest similarity in all three metrics. For 0% SER, GP-SARSA seems to result in more similar behaviour independent for the reward model.

On the right side of Table 10, the scores are averaged over the policy models to identify which reward model pair results in the highest similarity in behaviour independent of the chosen policy model. Overall, R_{TS} – R^{bi}_{IQ} shows the highest similarity. Interestingly, the low score of R^{s}_{IQ} – R^{bi}_{IQ}

Table 9: The similarity scores analysing the **consistency of policies**. On the left side, the mean match rates for different policy models is shown averaged over the respective reward models. The right side shows the mean match rates for different reward models averaged over the respective policy models.

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	GP	0.46	0.72	0.44
	DQN	0.42	0.73	0.50
	eNAC	**0.55**	**0.83**	**0.64**
15%	GP	**0.52**	0.73	0.48
	DQN	0.50	**0.76**	0.42
	eNAC	0.49	0.72	**0.62**
30%	GP	0.45	0.64	0.52
	DQN	**0.52**	**0.70**	0.45
	eNAC	0.42	0.69	**0.57**
All	GP	0.48	0.70	0.48
	DQN	0.48	0.73	0.46
	eNAC	**0.49**	**0.75**	**0.61**

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	R_{TS}	**0.51**	**0.79**	**0.58**
	R^{s}_{IQ}	0.47	0.75	0.46
	R^{bi}_{IQ}	0.46	0.73	0.55
15%	R_{TS}	0.49	**0.77**	**0.57**
	R^{s}_{IQ}	0.48	0.71	0.54
	R^{bi}_{IQ}	**0.54**	0.73	0.41
30%	R_{TS}	0.43	0.68	0.52
	R^{s}_{IQ}	**0.51**	**0.69**	**0.53**
	R^{bi}_{IQ}	0.44	0.66	0.48
All	R_{TS}	0.48	**0.74**	**0.56**
	R^{s}_{IQ}	**0.49**	0.72	0.51
	R^{bi}_{IQ}	0.48	0.71	0.48

shows that both reward models result in quite different behaviour even though both models estimate the same quantity, i.e., the interaction quality.

5.3 Similarity between Policy Models

To analyse the behaviour of the trained policies that originate from the different policy models, this section compares policies from different policy models resulting in pair-wise comparisons of the policy of each random seed of one policy model with the policies of each random seed of another policy model within one reward model. With three random seeds each, this results in nine comparisons. As in the previous sections, each policy is evaluated with 100 evaluation dialogues taken from one of the policies. For each policy model and reward model comparison, the similarity scores are averaged. The full scores are shown in the appendix in Table 14.

The similarity scores shown in Table 11 show the average scores for TMR, DMR, and CMR. In the left table, the scores are averaged over the different reward models to identify which policy model pair learns the most similar behaviour, i.e., ends up in similar behaviour independent of the reward model. Overall, GP – eNAC shows the highest similarity in all three metrics. For 15% and 30% SER, the two deep reinforcement learning models DQN – eNAC show rather low similarity in behaviour.

On the right side of Table 11, the scores are averaged over the policy model pairs to identify which reward model results in the highest similarity among the policy models. Overall, R^{s}_{IQ} seems to be the strongest learning signal resulting in the most similar behaviour between policy models. The similarity resulting from using R^{s}_{TS}, instead, is rather low.

Table 10: The similarity scores analysing the **similarity between reward models**. On the left side, the mean match rates for different policy models is shown averaged over the respective reward model pairs. The right side shows the mean match rates for different reward model pairs averaged over the respective policy models.

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	GP	**0.42**	**0.62**	**0.47**
	DQN	0.38	0.59	0.42
	eNAC	0.37	0.53	0.39
15%	GP	0.33	0.61	0.38
	DQN	**0.40**	**0.64**	**0.46**
	eNAC	0.37	0.58	0.42
30%	GP	0.36	0.52	0.42
	DQN	**0.41**	**0.60**	**0.47**
	eNAC	0.40	0.56	0.45
All	GP	0.37	0.59	0.42
	DQN	**0.40**	**0.61**	**0.45**
	eNAC	0.38	0.56	0.42

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	$R_{TS} - R^{s}_{IQ}$	0.34	0.54	0.41
	$R_{TS} - R^{bi}_{IQ}$	**0.53**	**0.71**	**0.59**
	$R^{s}_{IQ} - R^{bi}_{IQ}$	0.29	0.48	0.27
15%	$R_{TS} - R^{s}_{IQ}$	0.37	0.58	0.43
	$R_{TS} - R^{bi}_{IQ}$	**0.47**	**0.70**	**0.53**
	$R^{s}_{IQ} - R^{bi}_{IQ}$	0.27	0.55	0.31
30%	$R_{TS} - R^{s}_{IQ}$	0.34	0.48	0.41
	$R_{TS} - R^{bi}_{IQ}$	**0.52**	**0.71**	**0.59**
	$R^{s}_{IQ} - R^{bi}_{IQ}$	0.30	0.49	0.34
All	$R_{TS} - R^{s}_{IQ}$	0.35	0.53	0.42
	$R_{TS} - R^{bi}_{IQ}$	**0.51**	**0.71**	**0.57**
	$R^{s}_{IQ} - R^{bi}_{IQ}$	0.29	0.51	0.31

Table 11: The similarity scores analysing the **similarity between policy models**. On the left side, the mean match rates for different policy model pairs is shown averaged over the respective reward models. The right side shows the mean match rates for different reward models averaged over the respective policy model pairs.

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	GP – DQN	0.23	0.41	0.27
	GP – eNAC	**0.35**	0.52	0.39
	DQN – eNAC	0.33	**0.53**	**0.41**
15%	GP – DQN	0.27	0.50	0.35
	GP – eNAC	**0.30**	**0.54**	**0.38**
	DQN – eNAC	0.25	0.53	0.32
30%	GP – DQN	**0.38**	**0.54**	**0.45**
	GP – eNAC	0.33	0.51	0.42
	DQN – eNAC	0.34	0.51	0.39
All	GP – DQN	0.29	0.48	0.36
	GP – eNAC	**0.33**	0.52	**0.40**
	DQN – eNAC	0.31	**0.52**	0.37

SER	*Type*	*TMR*	*DMR*	*CMR*
0%	R_{TS}	0.29	0.48	**0.38**
	R^{s}_{IQ}	0.30	**0.50**	0.37
	R^{bi}_{IQ}	**0.31**	0.48	0.33
15%	R_{TS}	0.26	0.52	0.35
	R^{s}_{IQ}	**0.31**	0.51	**0.38**
	R^{bi}_{IQ}	0.25	**0.54**	0.32
30%	R_{TS}	0.26	0.43	0.34
	R^{s}_{IQ}	**0.51**	**0.68**	**0.62**
	R^{bi}_{IQ}	0.28	0.44	0.31
All	R_{TS}	0.27	0.48	0.36
	R^{s}_{IQ}	**0.38**	**0.56**	**0.45**
	R^{bi}_{IQ}	0.28	0.49	0.32

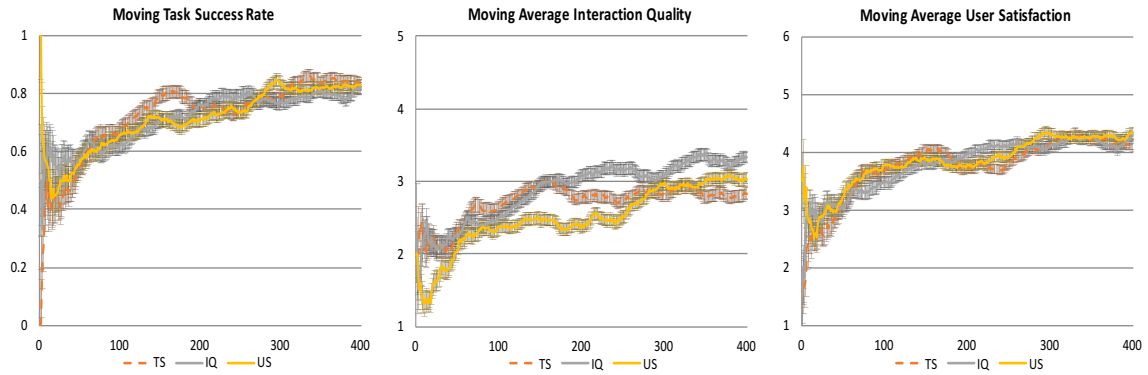

Figure 8: Moving TSR (left), moving AIQ^s (middle) and moving AUS (right) for using either R_{TS}, R^s_{IQ}, or R_{US} as reward averaged over two policies respectively, computed on windows consisting of 120 dialogues.

6. Learning from Real Humans

For learning a policy directly from the interaction with real humans, the IQ-based policy was only learned using R^s_{IQ} using GP-SARSA for dialogues in the CR domain[10]. Subjects were recruited through Amazon Mechanical Turk to talk to a 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?") and 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 (Gašić et al., 2013) ($obj = subj$).

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^s_{IQ}:

$$R_{US} = T \cdot (-1) + (US - 1) \cdot 5 \; . \tag{23}$$

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^s and moving *AUS* and are presented in Figure 8. Each value in the graphs is calculated using the respective previous 120 training dialogues averaged over the three trials.

The results show that R^s_{IQ} may successfully be used for learning through interactions with real users. It achieves the same TSR as with R_{TS} while resulting in better results for AIQ. The results for moving AUS also 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^s_{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.

10. These experiments and results have previously been published by Ultes et al. (2017a).

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 9: Example dialogue observed during training using R_{TS} after 387 training dialogues resulting in $iq = 1$.

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 10: Example dialogue observed during training using R^s_{IQ} after 365 training dialogues resulting in $iq = 4$.

Figures 9 and 10 show two successful example dialogues for the models trained with R_{TS} and R^s_{IQ}, respectively. One effect of training with R^s_{IQ} was a reduced number of system repetitions (which may be linked to the RePrompt? feature).

Obviously, the human experiments have only been conducted using R^s_{IQ} and not with R^{bi}_{IQ}. However, the general framework of applying an IQ reward estimator for learning a dialogue policy has shown to be applicable in such a setup and it seems rather unlikely that the changes we induce by changing the reward estimator lead to a fundamentally different result.

7. Discussion

One of the major questions of this work addresses the impact of an IQ reward estimator on the resulting dialogues as different IQ estimators achieves different levels of performance. Analysing the results of the dialogue policy learning experiment leads to the conclusion that, for GP-SARSA, the policy learned with R_{IQ}^{bi} performs similar or better than R_{IQ}^{s} throughout all experiments while still achieving better average user satisfaction compared to R_{TS}. Especially for noisy environments, the improvement is relevant.

This finding does not transfer to the other policy models, though. For both, DQN and eNAC, using R_{TS} results in the overall best peformance for both. This can be explained by the nature of the IQ reward estimates as they are not as noise-free as R_{TS}. Here, Gaussian processes are known to be able to deal with this type of noisy targets better while deep learning approaches are known to be quite sensitive.

The BiLSTM clearly performs better on the LEGO corpus while learning the temporal dependencies instead of using handcrafted ones. However, it entails the risk that these learned temporal dependencies are too specific to the original data so that the model does not generalise well anymore. This would mean that it would be less suitable to be applied to dialogue policy learning for different domains. The experiments clearly show that this is not the case.

One additional aspect for discussion is the definition of R_{IQ}, where the interaction quality is paired with a dialogue length penalty to guarantee a better comparison with R_{TS}. This length penalty is not strictly necessary as the annotated scores of the LEGO corpus already contain a notion of dialogue length implicitly as long dialogues usually tend to have lower quality ratings (in contrast to other work, e.g., (Foster et al., 2009)). Furthermore, the IQ estimation approach described in the following uses the dialogue length as an explicit input parameter. However, adding a dialogue length penalty to R_{IQ} is not regarded as harmfull to the overall approach as dialogue length plays a subordinate role in the learning process. Moreover, even though this limits the usage of the LEGO corpus for its applicability to types of interactions, it does not limit the overall approach presented in this work: for different tasks, new data needs to be collected and annotated.

8. Conclusion

This article has demonstrated 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 reward estimators 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. The reward estimator using BiLSTMs with attention mechanism achieved better estimation performance than a SVM-based estimator while learning all temporal dependencies implicitly. Moreover, one of the estimators 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, we aim at extending the user satisfaction estimator by incorporating domain-independent linguistic data to further improve the estimation performance. To tackle the problem of degrading performance if the noise level increases, 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. Furthermore, the effects of using a user

satisfaction-based reward estimator needs to be applied to more complex tasks, e.g., as defined by the Conversational Entity Dialogue Model (Ultes et al., 2018).

9. Acknowledgements

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

References

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. Multi-domain conversation quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1911.08567*, 2019a.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1908.07064*, 2019b.

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.347. URL https://www.aclweb.org/anthology/2020.findings-emnlp.347.

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

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

Jacob Cohen. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46, April 1960.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Heriberto Cuayáhuitl, Seonghan Ryu, Donghyeon Lee, and Jihie Kim. A study on dialogue reward prediction for open-ended conversational agents. In *2018 NeurIPS Workshop on Conversational AI: "Today's Practice and Tomorrow's Potential", Montréal, Canada.*, 2018.

Lucie Daubigney, Matthieu Geist, and Olivier Pietquin. 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), 2012. IEEE. doi: 10.1109/ICASSP.2012.6289040.

Layla El Asri, Romain Laroche, and Olivier Pietquin. Reward Function Learning for Dialogue Management. In *Proceedings of the 6th Starting AI Researchers' Symposium (STAIRS)*, pages 95–106. IOS Press, 2012. doi: 10.3233/978-1-61499-096-3-95.

Layla El Asri, Romain Laroche, and Olivier Pietquin. Reward shaping for statistical optimisation of dialogue management. In Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, and Bianca Truthe, editors, *Statistical Language and Speech Processing*, pages 93–101, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39593-2.

Layla El Asri, Hatim Khouzaimi, Romain Laroche, and Olivier Pietquin. Ordinal regression for interaction quality prediction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3245–3249. IEEE, May 2014a. doi: 10.1109/ICASSP.2014.6854195.

Layla El Asri, Romain Laroche, and Olivier Pietquin. Task completion transfer learning for reward inference. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014b.

Maxine Eskenazi, Alan W Black, Antoine Raux, and Brian Langner. 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*, 2008.

Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 879–887, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P09-1099`.

Milica Gašić and Steve J. Young. Gaussian processes for POMDP-based dialogue manager optimization. *IEEEACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40, 2014. doi: 10.1109/TASL.2013.2282190.

Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve J. Young. 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, 2013. doi: 10.1109/ICASSP.2013.6639297.

Milica Gašić, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve J. Young. Incremental on-line adaptation of POMDP-based dialogue managers to extended domains. In *Proceedings of the 15th International Conference on Spoken Language Processing (INTERSPEECH)*, pages 140–144. ISCA, August 2014.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013. doi: 10.1109/ASRU.2013.6707742.

James Henderson, Oliver Lemon, and Kallirroi Georgila. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511, 2008.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4337. URL https://www.aclweb.org/anthology/W14-4337.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977. ISSN 0006-341X.

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

Oliver Lemon and Olivier Pietquin. *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer New York, 2012. ISBN 978-1-4614-4802-0. doi: 10.1007/978-1-4614-4803-7.

Esther Levin and Roberto Pieraccini. A stochastic model of computer-human interaction for learning dialogue strategies. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 1883–1886, 1997.

Bing Liu and Ian Lane. Adversarial learning of task-oriented neural dialog models. In *19th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 350–359, 2018.

Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 84–93, Seoul, South Korea, July 2012. Association for Computational Linguistics. URL http://aclweb.org/anthology-new/W/W12/W12-1611.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. URL https://doi.org/10.1038/nature14236.

Niklas Rach, Wolfgang Minker, and Stefan Ultes. Interaction quality estimation using long short-term memories. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 164–169, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5520. URL https://www.aclweb.org/anthology/W17-5520.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 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)*, September 2006.

Verena Rieser and Oliver Lemon. Automatic learning and evaluation of user-centered objective functions for dialogue system optimisation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/592_paper.pdf.

Verena Rieser and Oliver Lemon. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proceedings of ACL-08: HLT*, pages 638–646, Columbus, Ohio, June 2008b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P08-1073.

Jost Schatzmann and Steve J. Young. The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):733–747, 2009. doi: 10.1109/TASL.2008.2012071.

Alexander Schmitt and Stefan Ultes. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36, November 2015. ISSN 0167-6393. doi: 10.1016/j.specom.2015.06.003. URL http://www.sciencedirect.com/science/article/pii/S0167639315000679.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 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, May 2012a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/333_Paper.pdf.

Alexander Schmitt, Stefan Ultes, and Wolfgang 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)*, pages 3369–337, May 2012b.

Weiyan Shi and Zhou Yu. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1140. URL https://www.aclweb.org/anthology/P18-1140.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.

Charles Edward Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

Pei-Hao Su, David Vandyke, Milica Gašić, Dongho Kim, Nikola Mrkšić, Tsung-Hsien Wen, and Steve J. Young. Learning from real users: Rating dialogue success with neural networks for

reinforcement learning in spoken dialogue systems. In *Interspeech*, pages 2007–2011. ISCA, September 2015.

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1230. URL https://www.aclweb.org/anthology/P16-1230.

Pei-Hao Su, Paweł Budzianowski, Stefan Ultes, Milica Gašić, and Steve Young. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 147–157, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5518. URL https://www.aclweb.org/anthology/W17-5518.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981. URL http://portal.acm.org/citation.cfm?id=551283.

T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Stefan Ultes. 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, September 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5902. URL https://www.aclweb.org/anthology/W19-5902.

Stefan Ultes and Wolfgang Maier. 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, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.sigdial-1.38.

Stefan Ultes and Wolfgang Minker. Improving interaction quality recognition using error correction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 122–126, Metz, France, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-4018.

Stefan Ultes and Wolfgang Minker. 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., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4328. URL https://www.aclweb.org/anthology/W14-4328.

Stefan Ultes, Tobias Heinroth, Alexander Schmitt, and Wolfgang Minker. A theoretical framework for a user-centered spoken dialog manager. In Ramón López-Cózar and Tetsunori Kobayashi, editors, *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue*

Systems Workshop, pages 241–246, New York, NY, September 2011. Springer New York. ISBN 978-1-4614-1334-9. doi: 10.1007/978-1-4614-1335-6_24.

Stefan Ultes, Alexander Schmitt, and Wolfgang 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)*, pages 49–52, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-1819.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 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, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1064.

Stefan Ultes, Matthias Kraus, Alexander Schmitt, and Wolfgang Minker. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4649. URL https://www.aclweb.org/anthology/W15-4649.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young. Domain-independent user satisfaction reward estimation for dialogue policy learning. In *Proc. Interspeech 2017*, pages 1721–1725. ISCA, August 2017a. doi: 10.21437/Interspeech.2017-1032. URL http://dx.doi.org/10.21437/Interspeech.2017-1032.

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. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada, July 2017b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4013.

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yen-Chen Wu, Steve Young, and Milica Gašić. Addressing objects and their relations: The conversational entity dialogue model. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5032. URL https://www.aclweb.org/anthology/W18-5032.

Stefan Ultes, Juliana Miehle, and Wolfgang Minker. *On the Applicability of a User Satisfaction-Based Reward for Dialogue Policy Learning*, pages 211–217. Springer International Publishing, Cham, 2019. ISBN 978-3-319-92108-2. doi: 10.1007/978-3-319-92108-2_22. URL https://doi.org/10.1007/978-3-319-92108-2_22.

David Vandyke, Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young. Multi-domain dialogue success classifiers for policy training. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 763–770. IEEE, 2015. doi: 10.1109/ASRU.2015.7404865.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8. doi: 10.1007/978-1-4757-3264-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Marilyn Walker. 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. doi: https://doi.org/10.1613/jair.713.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, July 1997. Association for Computational Linguistics. doi: 10.3115/976909.979652. URL https://www.aclweb.org/anthology/P97-1035.

Marilyn A. Walker, Jeanne C. Fromer, and Shrikanth Narayanan. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, 1998. URL https://www.aclweb.org/anthology/C98-2214.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Jason D. Williams and Steve J. Young. Characterizing task-oriented dialog using a simulated asr chanel. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004)*, pages 185–188, 2004.

Steve J. Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. doi: 10.1109/JPROC.2012.2225812.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1049–1058. ACM, 2018.

Appendix A. Tables

Table 12: Full results showing the **consistency of learned policies** in terms of total match rate (TMR), dialogue act match rate (DMR) and concept match rate (CMR) for three different semantic error rates (SERs). Similarity scores are computed by comparing the learned behaviour of the three policies originating from different random seeds for each policy model and each reward model.

SER	*Policy*	*TMR*			*DMR*			*CMR*		
		R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}	R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}	R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}
0%	GP	0.48	0.44	0.47	0.72	0.70	0.74	0.55	0.26	0.51
	DQN	0.42	0.46	0.39	0.76	0.71	0.70	0.51	0.53	0.48
	eNAC	0.61	0.52	0.52	0.90	0.85	0.76	0.67	0.59	0.67
15%	GP	0.51	0.39	0.66	0.74	0.66	0.78	0.58	0.40	0.45
	DQN	0.40	0.52	0.57	0.74	0.75	0.78	0.50	0.57	0.18
	eNAC	0.56	0.52	0.39	0.82	0.72	0.63	0.63	0.64	0.60
30%	GP	0.36	0.52	0.46	0.62	0.61	0.70	0.45	0.47	0.63
	DQN	0.51	0.54	0.50	0.69	0.73	0.66	0.59	0.49	0.27
	eNAC	0.44	0.46	0.37	0.71	0.74	0.61	0.53	0.64	0.55

Table 13: Full results showing the **comparison of learned policies using different reward models** in terms of total match rate (TMR), dialogue act match rate (DMR) and concept match rate (CMR) for three different semantic error rates (SERs). Similarity scores are computed by pair-wise comparing the learned behaviour of the three policies of each reward model originating from different random seeds with each trained policy of the other reward model within each policy model.

SER	*Type*	*TMR*			*DMR*			*CMR*		
		GP	*DQN*	*eNAC*	*GP*	*DQN*	*eNAC*	*GP*	*DQN*	*eNAC*
0%	R_{TS} – R^{s}_{IQ}	0.31	0.33	0.39	0.57	0.55	0.53	0.40	0.40	0.44
	R_{TS} – R^{bi}_{IQ}	0.58	0.58	0.45	0.72	0.75	0.65	0.64	0.64	0.50
	R^{s}_{IQ} – R^{bi}_{IQ}	0.38	0.45	0.27	0.55	0.65	0.42	0.36	0.50	0.23
15%	R_{TS} – R^{s}_{IQ}	0.24	0.41	0.45	0.48	0.64	0.62	0.31	0.48	0.51
	R_{TS} – R^{bi}_{IQ}	0.49	0.53	0.37	0.78	0.74	0.59	0.57	0.60	0.41
	R^{s}_{IQ} – R^{bi}_{IQ}	0.26	0.37	0.27	0.57	0.59	0.53	0.27	0.41	0.34
30%	R_{TS} – R^{s}_{IQ}	0.22	0.36	0.45	0.30	0.55	0.57	0.29	0.43	0.49
	R_{TS} – R^{bi}_{IQ}	0.48	0.57	0.51	0.68	0.76	0.70	0.55	0.63	0.59
	R^{s}_{IQ} – R^{bi}_{IQ}	0.37	0.51	0.24	0.58	0.70	0.41	0.40	0.59	0.26

Table 14: Full results showing the **comparison of learned policies using different policy models** in terms of total match rate (TMR), dialogue act match rate (DMR) and concept match rate (CMR) for three different semantic error rates (SERs). Similarity scores are computed by pair-wise comparing the learned behaviour of the three policies of each policy model originating from different random seeds with each trained policy of the other policy model within each reward model.

SER	*Type*	*TMR*			*DMR*			*CMR*		
		R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}	R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}	R_{TS}	R^{s}_{IQ}	R^{bi}_{IQ}
0%	GP – DQN	0.25	0.20	0.22	0.46	0.40	0.37	0.35	0.27	0.20
	GP – eNAC	0.35	0.37	0.33	0.50	0.56	0.49	0.43	0.43	0.32
	DQN – eNAC	0.28	0.33	0.38	0.48	0.53	0.59	0.36	0.40	0.46
15%	GP – DQN	0.24	0.24	0.33	0.45	0.45	0.61	0.34	0.31	0.39
	GP – eNAC	0.30	0.35	0.25	0.57	0.53	0.52	0.39	0.43	0.32
	DQN – eNAC	0.23	0.36	0.18	0.56	0.55	0.47	0.33	0.40	0.23
30%	GP – DQN	0.29	0.55	0.31	0.41	0.72	0.49	0.36	0.65	0.35
	GP – eNAC	0.23	0.50	0.26	0.44	0.67	0.41	0.31	0.62	0.33
	DQN – eNAC	0.25	0.49	0.29	0.45	0.66	0.42	0.35	0.58	0.24