[CASPI] Causal-aware Safe Policy Improvement for Task-oriented Dialogue

Govardana Sachithanandam Ramachandran, Kazuma Hashimoto*, Caiming Xiong
Salesforce Research
gramachandran@salesforce.com
hassy@logos.t.u-tokyo.ac.jp
cxiong@salesforce.com

Abstract

The recent success of reinforcement learning (RL) in solving complex tasks is often attributed to its capacity to explore and exploit an environment. Sample efficiency is usually not an issue for tasks with cheap simulators to sample data online. On the other hand, Task-oriented Dialogues (ToD) are usually learnt from offline data collected using human demonstrations. Collecting diverse demonstrations and annotating them is expensive. Unfortunately, RL policy trained on off-policy data are prone to issues of bias and generalization, which are further exacerbated by stochasticity in human response and non-markovian nature of annotated belief state of a dialogue management system. To this end, we propose a batch-RL framework for ToD policy learning: Causal-aware Safe Policy Improvement (CASPI). CASPI includes a mechanism to learn fine-grained reward that captures intention behind human response and also offers guarantee on dialogue policy's performance against a baseline. We demonstrate the effectiveness of this framework on end-to-end dialogue task of the Multiwoz2.0 dataset. The proposed method outperforms the current state of the art. Further more we demonstrate sample efficiency, where our method trained only on 20% of the data, are comparable to current state of the art method trained on 100% data on two out of there evaluation metrics.

1 Introduction

Offline task-oriented dialogue (ToD) systems involves solving disparate tasks of belief states tracking, dialogue policy management, and response generation. Of these tasks, in this work we focus on dialogue policy management to improve the end-to-end performance of ToD. The need for sample efficiency is key for learning offline task-oriented dialogue system, as access to data are finite and expensive. Recent advancements in off-policy reinforcement learning methods that uses offline data as against a simulator has proven to be sample efficient (Thomas and Brunskill, 2016). The effective use of these techniques are hindered by the nature of ToD. For instance, bias correction in off-policy based methods usually requires estimation of behaviour policy for a given state of Markov Decision Process (MDP). In ToD, per-turn annotated belief-state does not capture the true state of the MDP. Example of such annotated belief-state are shown in Fig:1. Latent state information such as prosody, richness of natural language and among others induces stochasticity in the agents response. In addition to these short comings, the direct use of automatic evaluation metric as reward for policy learning is not desirable, since these automatic evaluation metrics are often for the entire dialogue and not per turn. Hence such rewards are sparse and under-specified (Wang et al., 2020). Use of under-specified reward will often lead to policy that suffers from high variance (Agarwal et al., 2019). Alternatively use of imitation learning based methods falls short of reasoning on the outcome. This is demonstrated in Fig:1. Turns#3 and #2 are rich in semantic information and Turn#3 is key to success of the booking process. While Turn#4 contributes least to successful outcome. Though the turns have varying levels of importance, each of the turns are treated equally in imitation learning. In worst case, turns like Turn#4 will appear more often than turns Turn#2 and #3 in a ToD dataset, there by taking greater share of the gradient budget.

We address aforementioned shortcomings with following key contributions:

1.We introduce pairwise causal reward learning to learn fine grained per turn reward that reason the intention of human utterance.

2.We propose a safe policy improvement method

*Contributed to this work during his time at Salesforce Research

Code: https://github.com/salesforce/CASPI

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 92 - 102

Figure 1: A typical Task oriented dialogue conversation in MultiWoz2.0 dataset

for task oriented dialogue setting that guarantees performance against a baseline.

By use of these two methods, we demonstrate performance and sample efficiency.

2 Related Works

With the release of multi-domain, multi-turn MultiWoz2.0 dataset (Budzianowski et al., 2018a), there has been flurry of recent works, of which Zhang et al. (2019) uses data augmentation. Rastogi et al. (2019) and Hosseini-Asl et al. (2020) frame dialogue policy learning as language modeling task. Among the works that uses reinforcement learning. Mehri et al. (2019) uses supervised learning to bootstrap followed by RL fine tuning, whereas Zhao et al. (2019) uses policy gradient on latent action space as against handcrafted ones. Jaques et al. (2019) and Wang et al. (2020) uses Batch-RL for dialogue policy learning. (Wang et al., 2020) is first to argue the use of automated evaluation metrics directly as reward is under-specified for ToD policy learning. Recently there's has been proliferation in use of large pretrained language model based systems like Hosseini-Asl et al. (2020), Lin et al. (2020), Chen et al. (2019) etc. More details on contrasting the merits and limitations of these methods can be found in Sec:A.1

The line of inverse RL used in this work can be traced back to Ziebart et al. (2008), which proposes roll-outs from expert demonstration should have rewards exponentially higher than any other arbitrary roll-outs. This method requires a normalizing constant that integrates across rollouts, which is challenging. Christiano et al. (2017) and Thananjeyan et al. (2020) propose to do relative comparison of two roll-outs there by eliminating the need for normalization constant and they demonstrate in online setting.

3 Method

3.1 Preliminaries

We model task-oriented dialogue as a Markov decision process (MDP) (Sutton and Barto, 2018) with set of states S and actions A. The agent at time step t with state s_t performs a composite action a_t as per a target policy $\pi_e(a_t|s_t)$ on the environment. The environment is defined by transition probabilities $P(s_{t+1}|s_t, a_t)$, a latent reward function, $R(s_t, a_t, g)$, discount factor $\gamma \in [0, 1]$ and goal of dialogue g. Then the objective of the target policy π_e, is to maximizes the discounted sum of future reward on the MDP, given by the state-action value function $Q^{\pi_e}(a_t, s_t) = \mathbb{E}_{a_t \sim \pi_e, s_t \sim P}[\sum_{t'=t}^{T} \gamma^{t-t'} R(s_{t'}, a_{t'}, g)]$.

In offline Batch-RL. The agent does not get to interact with the environment, instead we are provided with offline data D logged by human agents performing actions based on a latent stochastic behaviour policy π_b. Rollout of a dialogue $\tau^i \in D$ is composed of $\tau^i = ((o_0^i, a_0^i), ..., (o_{T-1}^i, a_{T-1}^i))$. Here o_t is the observation at turn t, composing of $o_t = (b_t, u_t^u, u_{t-1}^a)$, where b_t is the belief state of the agent at turn t, u_t^u and u_{t-1}^a are the user and agent utterance at time t and $t-1$ respectively.

3.2 Safe policy improvement

Batch-RL entails training target policy π_e on rollout generated by a latent behaviour policy π_b. Directly optimizing on the rollouts generated by policy other than the target policy, will lead to large bias in the value function estimation, poor generalization characteristic, and sample inefficiency (Thomas and Brunskill, 2016). Safe policy improvement ensures the new policy performance is bounded by performance against a baseline policy. This is expressed as:

$$Pr(V^{\pi_e} \geq V^{\pi_b} - \zeta) \geq 1 - \delta,$$

where V^{π_e} and V^{π_b} are value functions of the target and behaviour policy respectively. Here $1-\delta$ and ζ are the high probability and approximation meta-parameters respectively. Schulman et al. (2015)

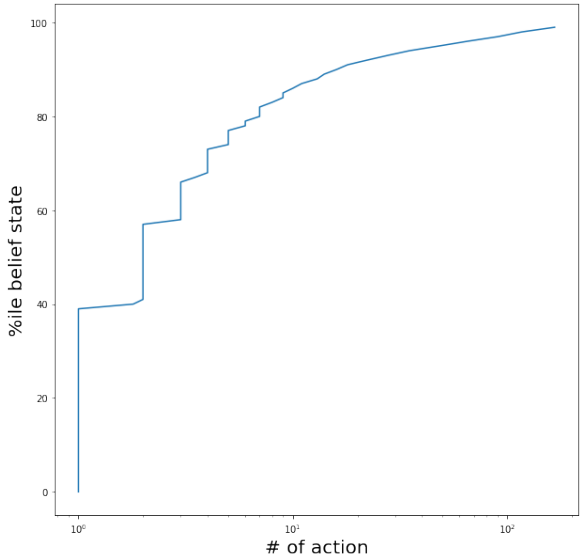

Figure 2: Shows stochacity i.e number of different dialogue act against each delexicalized belief state in MultiWoz2.0 dataset

provide such update mechanism, (1), whose errors are bounded as long as the constraints of (1) are met, where $D_{KL}(.||.)$ is the KL divergence and η is a hyper-parameter.

$$L_{sto}(\theta) = \min_{\substack{s_t \sim P^{\pi_{bs}} \\ a_t \sim \pi_{bs}}} -\mathbb{E}\left[\frac{\pi_e(a_t|s_t;\theta)}{\pi_{bs}(a_t|s_t)} Q^{\pi_{bs}}(s_t, a_t)\right]$$
$$s.t. \mathop{\mathbb{E}}_{s_t \sim P^{\pi_{bs}}}[D_{KL}(\pi_{bs}(.|s_t)||\pi_e(.|s_t))] \leq \eta \quad (1)$$

(Schulman et al., 2015) originally formulated (1) for online learning as trust region for policy updates and uses policy before gradient update as the baseline policy, $\pi_{bs}(a_t|b_t;\theta_{old})$. In this work we adapt it to offline setting and use behaviour policy π_b as the baseline policy. Use of this update rule requires access to the behavior policy $\pi_b(a_t|s_t)$ which is intractable to estimate and the learnt ones might have bias. Use of such behavior policy to perform bias correction by Important Sampling (Precup, 2000) might lead to worse policy. Instead we estimate the behaviour policy conditioned only the annotated belief-state b_t as against true state s_t in (1), which result in a stochastic behavior policy. This stochasticity of dialogue act vis-à-vis annotated belief state can observed in Fig:2. We also estimate the Q-function of the behavior policy, $Q^{\pi_b}(b_t, a_t)$ using learnt reward $R(s_t, a_t, g)$. More on learnt reward in Sec: 3.3.

The belief state b_t is part of the observation o_t, hence we purport that, on availability of more evidence of the observation o_t, (beside b_t) the mode of the policy collapse to a near deterministic action. To factor this into the policy learning, we have an additional loss:

$$L_{det}(\theta) = \min - \mathop{\mathbb{E}}_{(o_t,a_t)\sim D}[G(\tau, t) \log \pi_e(a_t|o_t;\theta)] \quad (2)$$

where return $G(\tau, t) = \sum_{t'=t}^{T} \gamma^{t'-t} R(s_{t'}, a_{t'}, g)$ is the discounted sum of future reward for rollout τ with goal g. Hence policy optimization loss function is given by:

$$L(\theta) = \alpha L_{sto}(\theta) + (1-\alpha) L_{det}(\theta) \quad (3)$$

We achieve this by doing two forward passes of the policy network $\pi_e(a_t|o_t;\theta)$, first with only the belief state, b_t as the input and second pass with entire observation i,e $o_t := (b_t, u_t^u, u_t^a)$ as input to the policy network. We then use the corresponding action distribution $\pi_e(a_t|b_t;\theta)$ and $\pi_e(a_t|o_t;\theta)$ in loss functions (1) and (2) respectively.

3.3 Pairwise causal reward learning

Algorithm 1 CASPI

Input: Dialogue dataset D and evaluation metric $M(.)$
Sub-sample K-folds of train and val set $\{(D_T, D_V)_1, ..., (D_T, D_V)_k | (D_T, D_V) \sim D\}$

for $\forall (D_T, D_V)$ **do**
  Learn ToD in supervised setting by optimizing for objective:
  $-\min \mathbb{E}_{a_t, s_t \sim D_T} \log(\pi_m(a_t|s_t))$
  for $\forall$ epoch **do**
    Using $\pi_m(a_t|s_t)$ predict actions on the valset D_V and add it to the dataset, D_P along with corresponding metric score $M(\tau)$ for pairwise causal reward learning $D_P = D_P \cup (\tau, M(\tau)) | \tau \sim \pi_m$
  end for
end for
repeat
  Sample pair of rollouts $(\tau^1, \tau^2) \sim D_P$
  Learn for $R(.)$ by optimizing for objective (4)
until Convergence using data D_P
repeat
  Optimize for policy π_e using objective (3)
until Convergence using data D

The policy optimization objective introduced in the previous section requires access to per time-step reward $R(s_t, a_t, g)$. To this end, we provide a

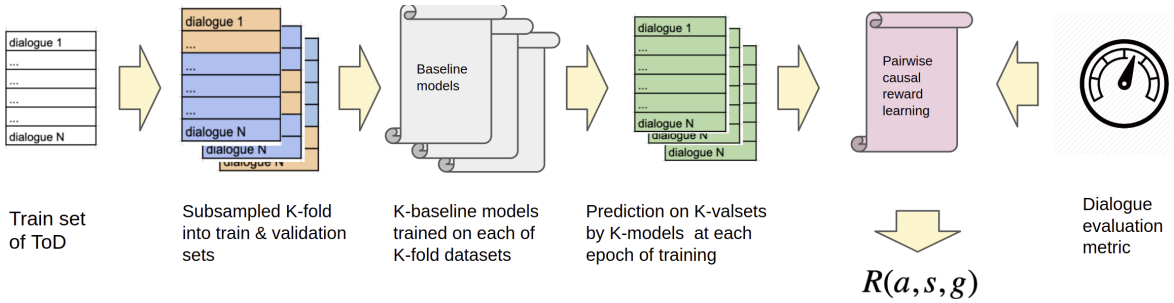

Figure 3: Process flow of pairwise causal reward learning

mechanism to learn a reward that is causally reasoned on the intention of the human demonstrator. Usually ToD are evaluated using dialogue level automatic evaluation metrics $M(.)$. Given the large state-action space of the dialogue management system, these dialogue level feedback are under-specified for for effective policy learning (Wang et al., 2020). Details about the the choice of evaluation metric $M(.)$ are covered in Sec:4.4.2.

To address this under-specified feedback, we adapt preference learning introduced by (Christiano et al., 2017) from an online to an offline setting, to learn fine grained per dialogue turn (ie. per timestep t) reward, $R(s_t, a_t, g)$. Given a pair of rollouts $\tau^1, \tau^2 \in D$ with actions for each state in the rollout is sampled from a pair of different policies π_m^1 and π_m^2 respectively. Let $\tau^1 \succ \tau^2$ represent preference of rollout τ^1 over rollout τ^2. This preference is true when sum of rewards of each dialogue turn of the two rollouts satisfies: $\sum_{t=0}^{T} R(s_t, a_t, g|(s_t, a_t) \in \tau^1) > \sum_{t=0}^{T} R(s_t, a_t, g|(s_t, a_t) \in \tau^2)$. For brevity, henceforth we refer $\sum_{t=0}^{T} R(s_t, a_t, g|(s_T, a_t) \in \tau)$ as $R(\tau)$. Then preferential probability of one rollout over an another, can be represented by:

$$P[\tau^1 \succ \tau^2] = \frac{\phi(R(\tau^1))}{\phi(R(\tau^1)) + \phi(R(\tau^2))}$$

Here $\phi(.)$ could either be $exp(.)$ or identity $\mathbb{1}(.)$. In our experiments, the later works best. We optimize for reward, $R(s_t, a_t, g)$ by minimizing binary cross-entropy loss between the preference probability and the normalized metrics score, $\mu(\tau)$ between a pair of rollout.

$$L(\theta) = \min - \mathop{\mathbb{E}}_{\tau^1 \sim \pi_m^1, \tau^2 \sim \pi_m^2} [\mu(\tau^1) \log P[\tau^1 \succ \tau^2] + \mu(\tau^2) \log P[\tau^2 \succ \tau^1]] \quad (4)$$

where,

$$\mu(\tau^1) = \frac{M(\tau^1)}{M(\tau^1) + M(\tau^2)} \quad (5)$$

We observe that the dialogue roll-outs are generated by expert latent policy. The data (dialogue rollouts) are distributed as per the optimal latent policy and transition probability. We propose that predictions made by a policy while in the process of learning to maximize the likelihood of the data is a good curriculum for exploring the state-action space for pairwise reward learning. This is a key insight of this work.

We formalize this insight into a method depicted in Fig:3 and Algo:1. The (train) dataset is subsampled into K-fold train & val sets. K-baseline policies are trained to fit the data distribution generated by experts using cross entropy loss, i.e supervised learning. During the process of fitting the data distribution, the still learning K-policies are used to predict on their corresponding K-fold valset at every epoch of the training. Each of these predictions are the scored by a chosen dialogue level metric, $M(.)$. On convergence of this supervised learning process, pairs of dialogue predictions generated by the above process, along with their corresponding metric score are used to train for fine grained reward $R(a_t, s_t, g)$ using objective (4).

The use of K-fold subsampling, K-baseline policies, π_m and actions sampled from these K-policies that are still in the process of learning help generate counter factual examples in the action space. These counter factual actions close to optimal policy, along with the goal of the dialogue helps us to learn subtle nuance of fine grained reward function $R(a_t, s_t, g)$ in the region of action space that matters the most.

4 Experimental Settings

4.1 Model

4.1.1 CASPI(.)

The learnt reward using CASPI $R(s_t, a_t, g)$ is akin to sample weights for each dialogue turn, that helps to redistribute the gradient budget among dialogue turns based of their contribution to the overall success of the ToD.

$$\theta := \theta - R(s_t, a_t, g)\nabla\pi_{blackbox}(a_t|s_t;\theta) \quad (6)$$

Hence we believe our pairwise casual reward learning and associated improvement in sample efficiency are independent of model architecture. To this end we choose two ToD methods that are at the extremes of model architecture spectrum 1) One uses a light weight custom model and 2) Other uses a large standard pre-trained out-of-the box universal language model.

4.1.2 CASPI(DAMD)

In this setting , we use the neural model proposed by Zhang et al. (2019). DAMD is composed of three $seq2seq$ generative model using GRUs. The three $seq2seq$ models are one each for belief state, dialogue act and response generation modules. An attention layers is used to attend the outputs of the $seq2seq$ models with the context vector of previous turn for copy over mechanism. The outputs of these attention layer are used as representation for predicting series of tokens for their respective modules. For more details on the model architecture and parameter setting refer Zhang et al. (2019). In this setting we use both stochastic, L_{sto} and deterministic, L_{det} loss functions on dialogue act. For DST and response generation, we retain the cross entropy loss as is from DAMD (Zhang et al., 2019).

4.1.3 CASPI(MinTL)

On the other extreme of model complexity, we use the Task oriented Dialogue model, MinTL(Lin et al., 2020). MinTL uses a large pretrained language model BART (Lewis et al., 2019). BART use as a standard encoder decoder transformer architecture with a bidirectional encoder and an autoregressive decoder. It is pre-trained on the task of denoising corrupt documents. BART is trained using cross-entropy loss between the decoder output and the original document. For more details of the model architecture and parameter setting, we suggest referring to (Lin et al., 2020) (Lewis et al., 2019).

MinTL doesn't explicitly predict dialogue act. Hence we only use the deterministic loss, L_{det} directly on the generated response and for DST we retain the loss as is from MintTL (Lin et al., 2020).

4.1.4 Pairwise Causal Learning Network

For k-model training of pairwise casual reward learning illustrated in Fig:3, we chose DAMD (Zhang et al., 2019) model for it's light weight model architecture. In all our experiments, we use $K = 10$.

For the pairwise casual reward learning network, we use three single bi-LSTM layers, one each to encode goal, belief state and either dialogue act or response sequences at each dialogue turn on each of the sampled roll-outs pairs, τ^1 and τ^2. The three encoded representations are concatenate and are fed through a couple of feed-forward layers before making a bounded reward prediction $R(s_t, a_t, g) \in [0, 1]$ for each turn using a sigmoid function. The per turn rewards are summed to form a global reward $R(\tau)$ for the roll-out τ. Using a pair of dialogue rewards $R(\tau^1)$ and $R(\tau^2)$, we compute the probabilistic preference between the roll-outs $P[\tau^1 \succ \tau^2]$ either by standard normalization or a softmax function. The output of this optimized using binary crossentopy loss described in Eqn:4. The above described architecture is illustrated in Fig:10 .

4.2 Dataset

To evaluate our proposed method on Multi-domain Wizard-of-Oz (MultiWoz) (Budzianowski et al., 2018a) dataset. It is a large scale multidomain, task oriented dataset generated by human-to-human conversation , where one participant plays the role of a user while the other plays the agent.The conversations are between a tourist and a clerk at an information center. The conversations span across 7 domains including attraction, hospital, hotel, police, restaurant, taxi and train. Each dialogue is generated by users with a defined goal which may cover 1-5 domains with a maximum of 13 turns in a conversation. The dataset has 10438 dialogues split into 8438 dialogues for training set and 1000 dialogues each for validation and test set.

4.3 Prepossessing

We represent DB results as one-hot vectors as proposed by Budzianowski et al. (2018b). To reduce surface-level variability in the responses, we use domain-adaptive delexicalization preprocess-

ing proposed in Wen et al. (2016). As proposed in Zhang et al. (2019), We generate delexicalized responses with placeholders for specific values which can be filled with information in DST and database.

4.4 Metrics

4.4.1 Evaluation

We evaluate performance of our method on end-to-end dialogue modeling task of Multiwoz2.0 (Budzianowski et al., 2018a). We uses three evaluations metrics proposed by (Budzianowski et al., 2018a). These include: 1) inform rate - measures the fraction of dialogue, the system has provided the correct entity, 2) success rate - fraction of dialogues, the system has answered all the requested information and 3) BLEU (Papineni et al., 2002) - measures the fluency of the generated response. We also report the combined score $(Inform + Success) \times 0.5 + BLEU$ proposed by Mehri et al. (2019). All the numbers of CASPI reported in this work are median of 5 runs with different seeds.

4.4.2 Training

For the metric M used in pairwise causal reward learning , we use the following:

$$M := Inform + Success + \lambda \times BLEU \quad (7)$$

This is very similar to combined score used in evaluation and both are equivalent when $\lambda = 2$. We introduced hyperparamter λ to normalize the achievable scale of $BLEU$. We observe that success rate, if used as is, will result in non-markovian and stochastic per turn reward function. This is because the reward of current state will depend on the performance of future states. Hence, we also use a soft version of the metric M_{soft}, where the success rate measures a fraction of requested information provided in a dialogue. We refer the original metric that uses the discrete variant of success rate as M_{hard}. The choice of action in reward function $R(s_t, a_t, g)$ can either be dialogue act or generate response, we refer corresponding variants of metrics as $M(act)$ and $M(resp)$. To demonstrate the versatility of our method to adapt to different metrics, we use all the discussed variants of the metric.

5 Result

We compare both adaptation of our methods CASPI(DAMD) and CASPI(MinTL) on the end-to-end dialogue tasks defined by MultiWoz2.0 (Budzianowski et al., 2018a). The results are tabulated at Table:1. CASPI(DAMD) with its light weight model architecture and no pretraining on any external corpus, except for (Lubis et al., 2020), out perform all other previous methods, these includes methods that use large pretrained language models such as Hosseini-Asl et al. (2020), Peng et al. (2020) and Lin et al. (2020). This show using CASPI to shepard the gradient update process as sample weights for each dialogue turn leads to a model that's well aligned with true objective of the task. CASPI(MinTL) with its robust pretrained model out performs CASPI(DAMD) and LAVA (Lubis et al., 2020) by a large margin. This demonstrates the ease of adaptation of existing methods with CASPI.

5.1 Sample Efficiency

Inverse reinforcement learning, coupled with off-policy policy learning and evaluation are proven to be sample efficient (Thomas and Brunskill, 2016) . We argue CASPI is competitive with other sample efficiency techniques, such as data augmentation and transfer learning as performed by Zhang et al. (2019) and Lin et al. (2020) respectively. To demonstrate the hypothesis, we test our method against baseline in a low sample complexity regime. For experimental setup, we adopt the low resource testing strategy from Lin et al. (2020). We train our model on 5%, 10%, and 20% of the training data and compared with other baselines on end-to-end dialogue task, Table 2 list the results. CASPI(MinTL) trained only on 20% of data was able to out perform previous state of the art method, LAVA (Lubis et al., 2020) and MINTL (Lin et al., 2020) trained on 100% data on two of the three performance metrics. This goes to show that having the right reward function to guide the budget of the gradient update process to reach the true objective is important in extremely low resource setting.

5.2 Human Evaluation

Automatic evaluation metrics have their own biases. True objective of ToD is human experience while interacting with the dialogue systems, which automatic evaluation metrics might fall short to capture. To this end we conduct human evaluation on the quality of the generated response. We define quality by the following criterias:

1) Appropriateness: Are the generated responses appropriate for the given context in the dialogue turn?

Model	Pre-trained model	Inform %	Success %	BLEU	Combined Score
DAMD	No	72.79	60.45	16.93	83.55
DAMD + multi-action	No	76.33	64.35	17.96	88.30
SimpleTOD	Yes	84.4	70.10	15.01	92.26
SOLOIST	Yes	85.5	72.90	16.54	95.74
MinTL-BART	Yes	84.88	74.91	17.89	97.79
LAVA	Yes	91.80	81.80	12.03	98.47
CASPI(DAMD), $M_{soft}(act)$	No	89.1	76.1	**18.08**	100.68
CASPI(MinTL), $M_{soft}(act)$	Yes	**94.59**	**85.59**	17.96	**108.05**
CASPI(MinTL), $M_{hard}(act)$	Yes	93.79	84.88	17.47	106.81

Table 1: Comparison of results for end-to-end task of Multiwoz2.0.

Model	5%			10%			20%		
	Inform	Success	BLEU	Inform	Success	BLEU	Inform	Success	BLEU
MD-Sequicity	49.40	19.70	10.30	58.10	34.70	11.40	64.40	42.10	13.00
DAMD	56.60	24.50	10.60	62.00	39.40	14.50	68.30	42.90	11.80
MinTL	75.48	60.96	**13.98**	78.08	66.87	**15.46**	82.48	68.57	13.00
CASPI(MinTL), $M_{soft}(resp)$	87.69	**71.17**	13.51	82.08	72.27	14.10	89.39	78.58	**15.16**
CASPI(MinTL), $M_{hard}(resp)$	**89.69**	69.47	13.33	**92.59**	**78.58**	14.48	**94.19**	**83.28**	13.65

Table 2: Comparison of results for end-to-end of Multiwoz2.0. in low resource setting

2) Fluency: Are the generated responses coherent and comprehensible?

A dialogue turn in the test set is randomly picked. The human evaluators were shown context leading up to the turn. The predictions for the turn by different methods were anonymized and displayed to the evaluators. This is illustrated in Fig:4. The human evaluators were asked to give a score between 1 and 5 for appropriateness and fluency, with score of 5 being best and 1 being the worst. 100 randomly selected dialogue turns were presented to 10 participants .We report the mean and variance of the score. We compare our model performance against MinTL (Lin et al., 2020), SimpleTOD (Hosseini-Asl et al., 2020), LAVA (Lubis et al., 2020) and DAMD (Zhang et al., 2019). Fig:5 shows the results of the evaluation. CASPI(MinTL) outperforms all other models in appropriateness score. While fluency score of CASPI(MinTL), MinTL and SimpleTOD are comparable to each other. It is worth noting that though LAVA (Lubis et al., 2020) performs well on automatic evaluation metrics, it performs poorly on human evaluation. We suspect the policy learnt by (Lubis et al., 2020) exploits gaps in the reward function. In case of LAVA (Lubis et al., 2020), success rate is used as the reward function. In our analysis, low BLEU score is good indicator if the learnt policy indulges in reward hacking, which LAVA (Lubis et al., 2020) exhibits. More on reward hacking in Sec:5.4.2.

5.3 Human in the loop training

In the previous section we argued that automatic dialogue evaluation metrics are biased and doesn't truly reflect the human objective, but in our method we use these very same dialogue evaluation metrics to learn reward $R(s_t, a_t, g)$. To bridge this gap, we performed the following human-in-the-loop (HITL) experiment. We first trained a pair CASPI(MINTL) models with different seeds, on 5% of Multiwoz2.0 dataset. We then used these pair of models to predict on 0.5% of Multiwoz2.0 train data (40 dialogues) and had a human score these pairs of generated response relative to each other. We then trained for reward $R(s_t, a_t, g)$ using pairwise causal reward learning as described in Sec:3.3, where examples of the mini batch are randomly sampled either from human scored examples or the ones scored by the automatic evaluation metric as show in Fig:6. We then trained a fresh CASPI(MINTL) model on the original 5% of data and the learnt $R(s_t, a_t, g)$. We perform human evaluation on 24 dialogues using 3 participants. Fig:7

user	i need to find a [value_type] that has free parking .
agent	sure i can help with that . what area will you be staying in ?
user	in the [value_area] , i more specifically am looking for a [value_type] with a [value_stars] star rating in that area .
agent-gt	[value_name] is a [value_stars] star [value_type] located in the [value_area] has both free parking and wifi . what days are you interested in staying there ? can i book for you ?
MinTL	i have [value_choice] [value_type] in the [value_area] with a [value_stars] star rating . it is [value_name] . would you like me to book it for you ?
CASPI(MinTL)	i have [value_choice] [value_type] in the [value_area] . [value_name] is [value_price] and [value_stars] stars . would you like me to book it ?
DAMD	i have [value_choice] options for you . is there a certain area you are looking for ?
SimpleTOD	there are no guesthouses with a [value_count] star rating in the [value_area], but there are [value_count] star guesthouses with free parking in the [value_pricerange] to [value_pricerange] price range
LAVA	the [hotel_name] is located at [hotel_address] , postcode [hotel_postcode] . the phone number is [hotel_phone] .

Figure 4: Example of generated responses by different ToD models

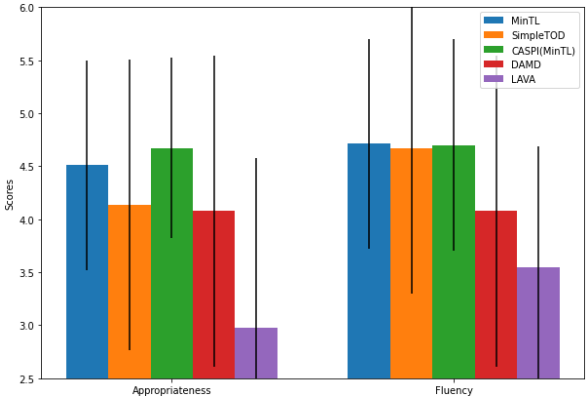

Figure 5: Human evaluation on criterias:Appropriateness and Fluency

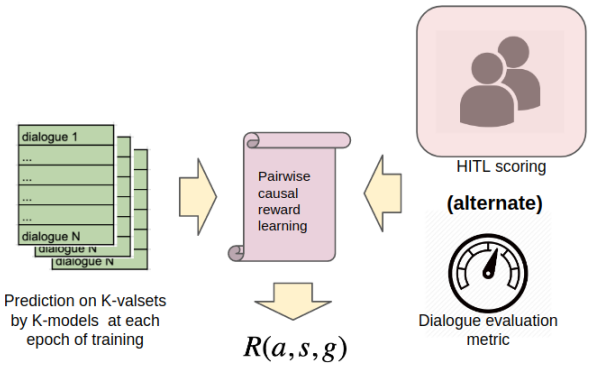

Figure 6: Mixed Human-in-the-loop and automatic evaluation metric scores for pairwise causal reward learning

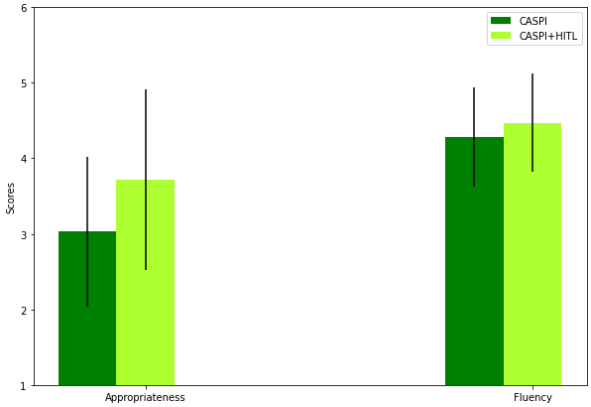

Figure 7: Human evaluation of Human in the loop training of CASPI(MinTL) on 5% of Multiwoz2.0 dataset

shows the performance.

Though CASPI(MINTL) using just 5% of the data outperforms DAMD trained on 100% of data in 2 out of the 3 automatic evaluation metrics shown in Table:1 and 2, performs poorly in human appropriateness score. With the HITL score in the reward learning, we see a boost in performance in both the human evaluation criteria: appropriateness and fluency. The 5% data CASPI(MINTL)'s human approriateness score is now comparable to 100% data DAMD. This goes to show the versatility of the pairwise causal reward learning. With enough expressiveness of the neural network used, the pairwise causal reward learning can generalize to unknown dialogue evaluation criteria.

5.4 Analysis

5.4.1 Rewards

In this section we qualitatively analyze the results of pairwise causal reward learning. Fig:8 is the same conversation between a tourist and information center agents that we introduced earlier, now we have learnt reward $R(s_t, a_t, g)$, against each turn. We observe that Turn#3 has received the highest reward, retrospectively we realize the transaction happens in this turn, which is crucial and has to be risk averse for the success of the dialogue. Turn#2 gets the next best reward which captures crucial information needed for transaction to happen in Turn#3. Turn#4 gets reward an order lower than Turn#3 & 2 because other than nicety, it doesn't contribute much to the success of the conversation. It should be noted that responses like Turn#4 will appear in almost all conversations and in supervised learning, these turns will be receiving the highest share of the gradient budget. The learnt reward redistributes the gradient budget based on the turns contribution to the success of the dialogue objective.

5.4.2 Type of agents

In this section we analyze the type of behaviour CASPI agents sometime exhibit, especially when trained in low sample regime.

Greedy agent: In certain domains, the agents has a tendency to book a service before it has gathered all the required information or before the user requested or agreed for booking a service. The first example in Fig:9 demonstrate this behaviour. Here the user has requested for a taxi, before enough information such as destination or time of departure

	Goal	{"train": {"info": { "leaveat": "09:00","destination": "broxbourne","day": "sunday","departure": "cambridge"},"fail_info": {},"book": {"invalid": true,"people": "8"},"fail_book": {}}}	R(a,s,g)
Turn-1	User	i am looking for a train . the train should leave after [value_leave] and should go to [value_destination] .	
	Agent	[train] [request] day	0.0614
		what day will this be on ?	
Turn-2	User	on [value_day] .	
	Agent	[train] [offerbook] leave id	0.1298
		[value_id] leaves at [value_leave] would you like me to book it ?	
Turn-3	User	yes please . i need that booked for [value_people] people .	
	Agent	[train] [offerbooked] reference people [general] [reqmore]	0.2295
		i was able to get you [value_people] tickets . your reference number is [value_reference] . can i help you with anything else ?	
Turn-4	User	no , thank you , that is all for now . have a great day .	
	Agent	[general] [welcome] [bye]	0.0249
		you are welcome . have a great trip !	

Figure 8: Example of learnt reward

	Greedy agent
	Cautious agent
user	help me reserve a taxi please
agent	okay , what is your departure site ?
user	i will be leaving from [value_departure]
agent	i have booked you a [value_car] . the contact number is [value_phone] .
user	i also need to book a taxi to get to the attraction by 03:30 and i will need the contact number and car type .
agent	i have booked a taxi for you . the contact number is [value_phone] . the car type is a [value_car] . the address is [value_address] . the phone number is [value_reference] . the postcode is [value_postcode] . the total fee is [value_price] . the reference number is : [value_reference] .

Figure 9: Example of agent behaviour in low sample regime.

are gathered, the agent books the taxi. This happens because there are gaps in automatic evaluation metrics. A low BLEU score and relatively high inform and success rate might indicate greedy agent behaviour. Other reasons for low BLEU score includes: lack of diversity in the responses or malformation of response.

Cautious agent: The agent tends to be cautious by providing long winded replies packed with more information than needed. Agent tend to do this to prevent the risk of loosing rewards by missing out any requested information. This behaviour is demonstrated in the second example in Fig:9

These subtle behaviour demonstrates gap in automatic evaluation metrics, which could be weeded out using Human in the loop learning described in Sec:5.3.

6 Conclusion

In this work we introduced a fine grained reward learning process using an under-specified metrics and expert demonstrations for efficiently learn task oriented dialogue. We demonstrated the efficacy of our method on MultiWoz2.0 dataset with results comparable to the existing state of the art method with only 20% of data. We believe the methods is generic and can be extend to other NLP tasks.

References

Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. 2019. Learning to generalize from sparse and underspecified rewards. *arXiv preprint arXiv:1902.07198*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018a. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018b. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *arXiv preprint arXiv:1905.12866*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Nurul Lubis, Christian Geishauser, Michael Heck, Hsien-chin Lin, Marco Moresi, Carel van Niekerk, and Milica Gašić. 2020. Lava: Latent action spaces via variational auto-encoding for dialogue policy optimization. *arXiv preprint arXiv:2011.09378*.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. *arXiv preprint arXiv:1907.10016*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Doina Precup. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Brijen Thananjeyan, Ashwin Balakrishna, Ugo Rosolia, Felix Li, Rowan McAllister, Joseph E Gonzalez, Sergey Levine, Francesco Borrelli, and Ken Goldberg. 2020. Safety augmented value estimation from demonstrations (saved): Safe deep model-based rl for sparse cost robotic tasks. *IEEE Robotics and Automation Letters*, 5(2):3612–3619.

Philip Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *arXiv preprint arXiv:2006.06814*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2019. Task-oriented dialog systems that consider multiple appropriate responses under the same context. *arXiv preprint arXiv:1911.10484*.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. *arXiv preprint arXiv:1902.08858*.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning.

A Appendix

A.1 Baselines

DAMD: Introduced by (Zhang et al., 2019)is a domain-aware multi-decoder network. The method also exploits stochastic nature of the dialogue act by using a data-augmentation technique called the multi-action data augmentation. DAMD with data augmentation is denoted here as DAMD + multiaction.

HDSA by (Chen et al., 2019) proposes to use hierarchical graph representation for dialogue act. It uses a pre-trained 12-layer BERT model (Devlin et al., 2019) to represent dialogue act. The predicted dialogue act is transformed to the hierarchical graph structure using disentangled self-attention model, a 3-layer self-attention model (Vaswani et al., 2017)

SOLOIST (Peng et al., 2020) and SimpleTOD (Hosseini-Asl et al., 2020) uses pretrained GPT-2-based methods. These method are trained on turn-level data without generated belief state and system act in dialog history.

MinTL-BART (Lin et al., 2020), introduced Levenshtein belief spans framework that predicts only the incremental change in dialogue state per turn. It leverages the pretrained T5 and BART (Lewis et al., 2019) as backbone for model architecture.

LAVA (Lubis et al., 2020), reduces the action space of policy in end-to-end ToD, by using the latent space of a variational model with an informed prior. The work use variable distribution: via pretraining, to obtain an informed prior, and uses auto-encoding as the auxiliary task, to capture generative factors of dialogue responses.

HDNO proposed by (Wang et al., 2020) is a dialogue policy learning method to solve context-to-response generation task of Multiwoz2.0 (Budzianowski et al., 2018a). It exploits the hierarchical nature of dialogue act and response generation task by proposing an option based framework of Hierarchical RL and variational model to learn a latent dialogue act that corresponds to natural language response. Unlike our

method, HDNO though highlights the risk of sparsity of metric function such as success rate as reward function, resorts to shaping a proxy reward function. It uses markov language model as a proxy reward function. The language model is learnt independent of the metric function. Our method refrains from reward shaping and is independent of the nature of any underspecified metric function. Since we learn fine grained turn specific credit assignment, our solution can adapt to other metric function as long as the pairwise reward network is rich enough to factorize them.

A.2 Pairwise causal reward learning network architecture

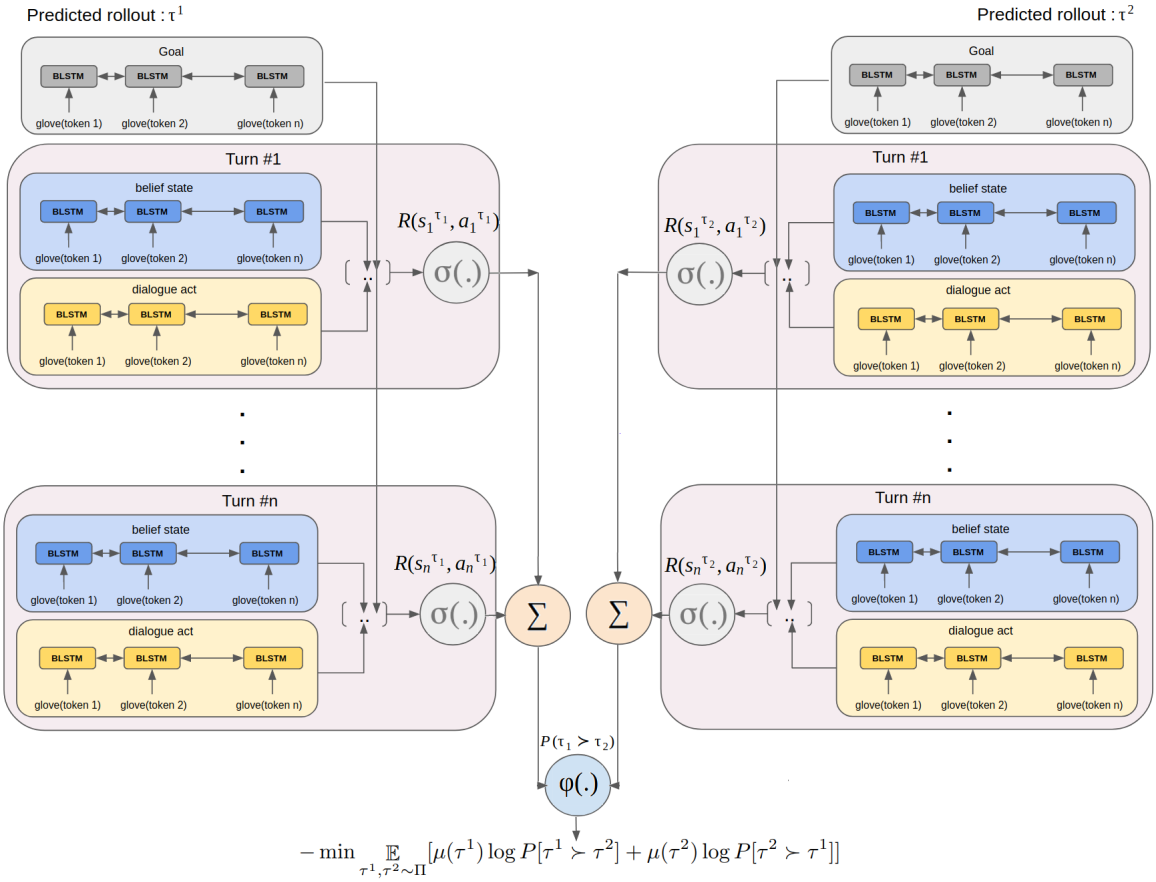

Figure 10: Pairwise causal reward learning network architecture