

Resource Constrained Dialog Policy Learning via Differentiable Inductive Logic Programming

Zhenpeng Zhou¹, Ahmad Beirami¹, Paul Crook¹, Pararth Shah¹,
Rajen Subba¹, and Alborz Geramifard¹

¹Facebook

{zzp, beirami, pacrook, pararths, rasubba, alborzg}@fb.com

Abstract

Motivated by the needs of resource constrained dialog policy learning, we introduce dialog policy via differentiable inductive logic (DILOG). We explore the tasks of one-shot learning and zero-shot domain transfer with DILOG on SimDial and MultiWoZ. Using a *single* representative dialog from the restaurant domain, we train DILOG on the SimDial dataset and obtain 99+% in-domain test accuracy. We also show that the trained DILOG zero-shot transfers to all other domains with 99+% accuracy, proving the suitability of DILOG to slot-filling dialogs. We further extend our study to the MultiWoZ dataset achieving 90+% inform and success metrics. We also observe that these metrics are not capturing some of the shortcomings of DILOG in terms of false positives, prompting us to measure an auxiliary Action F1 score. We show that DILOG is 100x more data efficient than state-of-the-art neural approaches on MultiWoZ while achieving similar performance metrics. We conclude with a discussion on the strengths and weaknesses of DILOG.

1 Introduction

To foster research on dialog policy learning for virtual digital assistants, several task-oriented dialog corpora have been introduced in recent years, such as SimDial (Zhao and Eskenazi, 2018), MultiWoZ (Budzianowski et al., 2018), Taskmaster (Byrne et al., 2019), and Schema Guided Dialog (Rastogi et al., 2019), to name a few. Deep learning approaches, including mixture models (Pei et al., 2019) hierarchical encoder/decoder (Zhang et al., 2019; Chen et al., 2019), reinforcement learning (Zhao et al., 2019), and pre-trained language models (Wu et al., 2019; Peng et al., 2020; Hosseini-Asl et al., 2020), have significantly advanced dialog policy research in the past few years (Gao et al., 2019), setting new state-of-the-art performance limits.

However, collecting annotated data for supervised dialog policy learning is an expensive and time-consuming process. Hence, it is desirable to explore approaches to train dialog policy with limited data and transfer an existing policy with few or even no additional training data to new domains. This practical requirement has motivated the community to research resource-constrained dialog policy learning in the past few decades. Researchers have explored approaches including employing grammar constraints for dialog policy (Eshghi et al., 2017), transfer learning (Shalyminov et al., 2019), or pre-trained language models (Zhao et al., 2020). Few-shot domain adaptation has been researched since the 2000s (Litman and Pan, 2002) on both end-to-end dialog systems (Qian and Yu, 2019; Zhao and Eskenazi, 2018) as well as dialog policy learning (Vlasov et al., 2018).

In a traditional modular dialog system, the dialog policy aims to decide a dialog action given a dialog state, while assuming the tasks of language understanding and generation are handled by other components. Under such assumptions, a task-oriented dialog policy mostly follow the slot-filling scheme, which can be described by a set of probabilistic rules. Therefore, we hypothesize that dialog policy in this limited sense can be constructed by learning the underlying rules. To this end, we draw upon

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

the recent advances in developing differentiable inductive logical programs (DILP) (Evans and Grefenstette, 2018) that use neural architectures to learn almost rule-based policies. We present DILOG, an adaptation of DILP to dialog policy learning. Briefly, DILOG discerns a set of logical rules from the examples by using inductive reasoning.¹ We introduce DILOG in Section 2. We apply DILOG to the SimDial Dataset (Zhao and Eskenazi, 2018) (Section 3), and MultiWoZ Dataset (Budzianowski et al., 2018) (Section 4), showing that on the task of one-shot dialog policy learning and zero-shot domain transfer, DILOG outperforms several other neural baselines. Finally, Section 5 concludes this paper.

2 DILOG: Dialog Policy Via Differentiable Inductive Logic

Inductive Logic Programming (ILP) is a paradigm which derives a hypothesized first-order logic given the background knowledge, positive, and negative examples (Muggleton and De Raedt, 1994). The central component of ILP is known as clauses. A **clause** is a rule expressed as $\alpha \leftarrow \alpha_1, \dots, \alpha_n$, where α is the **head** atom and $\alpha_1, \dots, \alpha_n$ are **body** atoms. An **atom** $p(t_1, \dots, t_m)$ is composed of an m -ary **predicate** p and a tuple of **terms** t_1, \dots, t_m , which can be variables or constants. An atom is **ground** if it only contains constants. For example, The following clause defines when to perform the action of confirmation:

$$\text{confirm}(S) \leftarrow \text{user_request}(S, T), \text{not_confident}(S),$$

which means if the user requests a slot S in a task T , and the system is not confident about S , then the system should confirm S with user. Employing the clause above on the grounding atoms of $\text{user_request}(\text{contact}, \text{calling})$ and $\text{not_confident}(\text{contact})$, we will be able to deduce the action of $\text{confirm}(\text{contact})$.

DILP (Evans and Grefenstette, 2018) combines ILP with a differentiable neural network architecture, to make ILP robust to noisy or ambiguous data. In short, DILP generates a collection of clauses based on the rule templates,² and assigns trainable weights to those clauses. Then logical deduction is applied recursively using the weighted sum of the clauses on the valuation vector $\mathbf{a} \in [0, 1]^g$, where g is the number of grounding atoms and $\mathbf{a}[i]$ is the probability that the grounding atom i is true. With the trainable weights, DILP can be trained using the gradient descent method.

DILOG is based on DILP with several modifications: (1) We include an option of adding ℓ_1 or ℓ_2 regularizers to the weight matrix of the clauses to improve generalization. (2) We allow adding clauses as background knowledge, so as to enable continual learning using pre-learned rules. (3) We use element-wise max instead of probabilistic sum as the amalgamate function to update the valuation vector. As probabilistic sum is accumulating the valuation at each step, it is easier to get into a local optimum when the inference steps are long. (4) In DILP, a problem is defined as $(\mathcal{L}, \mathcal{B}, \mathcal{P}, \mathcal{N})$, where \mathcal{L} is the language frame to generate potential clauses, \mathcal{B} is the background knowledge, \mathcal{P} is the positive examples, and \mathcal{N} is the negative examples. In this definition, all positive and negative examples are grounded on the same set of constants \mathcal{C} . As the number of clauses scales quadratically with the number of constants, it can be computationally expensive when the sample size is large. Comparatively, we define a problem as $(\mathcal{L}, \mathcal{S})$, where \mathcal{S} is the set of samples, and each $x \in \mathcal{S}$ is a tuple of $(\mathcal{B}, \mathcal{P}, \mathcal{N}, \mathcal{C})$. This modification allows each sample to define its own set of constants, which makes it computationally tractable under the setting of dialog system where each dialog consists of multiple turns.

In this paper, we demonstrate the application of DILOG on two tasks in dialog policy learning: one-shot learning and zero-shot domain transfer. In a modular dialog system, which consists of automatic speech recognition, natural language understanding, dialog state tracking (DST), dialog manager, and

¹Inductive reasoning tries to summarize general principles from special cases. For example, the fact “cars A, B, and C drive on the right side of the road” **induces** that “all cars drive on the right side of the road.”

²A template describes the rule used to generate the clauses. A rule template is defined by the number of additional variables $v \in \mathbb{N}$ and whether to allow intensional predicates $i \in \{0, 1\}$. For example, we can specify the template for predicate $p(X)$ to be $\{v = 0, i = 1\}$, which means using no additional variables (only X), and allowing intensional predicates. When two other 1-ary predicates exist as q and r , we first generate all possible atoms, namely $q(X)$, and $r(X)$, then we enumerate all combination of those atoms to generate the body of the clause. Therefore the clauses generated will be $p(X) \leftarrow q(X)$, $q(X)$, $p(X) \leftarrow r(X)$, $r(X)$, and $p(X) \leftarrow q(X), r(X)$. We refer to (Evans and Grefenstette, 2018) for a detailed elaboration on the generation process, and Appendix A for the process of choosing hyperparameters in a rule template.

natural language generation (NLG), the role of dialog policy is to map every state $s \in \mathbb{S}$ (represented by the DST) to a dialog action $a \in \mathbb{A}$. Compared with end-to-end policy learning that generates a natural language response, we only focus on learning the policy function $\pi : \mathbb{S} \rightarrow \mathbb{A}$, and leave NLG as a separate problem. Under this setting, DILOG offers several advantages:

- **Sample efficiency:** DILOG generalizes well from a small number of samples by introducing a language bias from the template used to generate the clauses, with which a set of succinct rule is preferred over a set of complex rules. This makes it useful in one-shot policy learning.
- **Interpretability:** The dialog policy learned via DILOG consists of a set of (probabilistic) rules. Each of the rules can be manually inspected and understood. This feature is desirable in industrial settings where interpretability and debuggability are key considerations.
- **Domain generalizability:** The rules learned by DILOG in one domain can be zero-shot transferred to new domains with unseen slots, by assuming symmetry between slots. This helps quickly enable a new domain with no new data collection and annotation.

3 DILOG on the SimDial Dataset

SimDial (Zhao and Eskenazi, 2018) is a multi-domain dialog generator that can generate conversations in domains including restaurant, movie, bus, and weather. Each domain is defined by a collection of meta data including user slots and systems slots. Ignoring the actions of greeting and goodbye, possible user actions are inform or request, while system actions include inform, request, and database query. We use SimDial to generate clean dialogs from all four domains, among which a single *representative* dialog (that contains all user and system actions) from the restaurant domain is used for training and 500 dialogs from each other domains are used as the test set.

Table 1: An illustrative example of adapting the dialog state and actions to the logical forms of DILOG. Note that we use a linked list to enumerate the user slots in the logical form, where `succ(X, Y)` denotes the successor of X is Y , and `terminal(X)` means X is a terminal node. The first 4 predicates of the belief state describe a linked list of `[usr_slot] → [food_pref] → [loc] → [term]`.

		State Representation	Logical Form	
Dialog:	Belief State	user_slot:	terminal(term)	known(loc)
		* food_pref: None	succ(usr_slot, food_pref)	unknown(food_pref)
		* loc: San Jose	succ(food_pref, loc)	unknown(open)
		sys_slot:	succ(loc, term)	unknown(price)
Previous Turn(s):	User Action	* open: None	usr_slots(usr_slot)	unknown(default)
		* price: None	known(usr_slot)	unknown(parking)
		* default: None		
		* parking: None		
Current Turn:	System Action	act: inform	inform(loc)	
		param: loc		
USR: What's up? I need a restaurant.		act: request	sys_request(food_pref)	
		param: food_pref		
SYS: Which place?				
USR: I am at San Jose.				
SYS: What kind of food do you like?				

To adapt the SimDial problem to DILOG, the delexicalized³ state and actions are converted into the form of atoms, whose predicate is the action or state, and term is the slot. For example, `request(loc)` denotes the action of requesting the location, while `unknown(price)` indicates that the price is unknown to the system. Each turn is converted to a sample s , where the background knowledge \mathcal{B} is the combination of user actions and the belief state, and the positive examples \mathcal{P} are the system actions. DILOG learns a mapping (a set of clauses) from the background \mathcal{B} to the positive examples \mathcal{P} . See Table 1 for an illustrative example of the adaption steps. The detailed process is described in Appendix C.

Note that the four domains have different slots, and during training time, the model is only aware of the slots in the training set. During test time, the learned rules are directly applied to the converted samples in the test set. Additionally, to demonstrate continual learning, we add the pre-trained basic relationships of `all` and `member` to the background knowledge. `all` identifies that all items in a list satisfy some property, and `member` indicates that an item belongs to a list. See Appendix C for a more detailed description of the training process.

³Delexicalization (Henderson et al., 2014) has been used previously to improve generalization in order to reinforce that the policy chooses a reasonable target.

Figure 1: The evaluation results on the test set of SimDial comparing DILOG with MLP and ZSDG (Zhao and Eskenazi, 2018). The In-Domain stands for the performance on the restaurant domain (500 samples, **standard error: $\pm 2.2\%$**). Out-of-Domain one is the average performance on movie, bus, and weather (1500 samples, **standard error: $\pm 1.3\%$**). See Table A1 (appendix) for the per-domain results. The plot shows the zero-shot performance on movie and weather. -1000 denotes that the model is trained with 1000 dialogs from the restaurant domain.

	In-Domain		Out-of-Domain	
	Intent F1	Entity F1	Intent F1	Entity F1
ZSDG	91.41	13.48	83.08	0.59
MLP	91.81	87.52	61.22	5.50
DILOG	99.74	99.78	99.75	99.81
ZSDG-1000	99.98	99.25	83.16	65.63
MLP-1000	98.97	97.98	54.49	6.63

We compare DILOG with two baselines, the first one is Zero-Shot Dialog Generation (ZSDG) (Zhao and Eskenazi, 2018), which learns a cross-domain embedding of actions to enable the policy to zero-shot transfer to new domains. The second one is a naive multi-layer perceptron (MLP) model mapping from the encoding of the states to the actions. The models are trained and evaluated on the same datasets. We employ two metrics to quantify the performance of different models: **Intent F1** measures whether the predicted dialog intent matches the ground truth, while **Entity F1** measures whether the entity is predicted correctly.

The evaluation results are shown in Figure 1. The In-Domain column demonstrates the performance of one-shot learning (trained with one dialog), while the Out-of-Domain one shows the performance of zero-shot domain transfer. We also include the ZSDG and MLP models trained with 1000 samples (denoted with -1000) from the restaurant domain. On the SimDial dataset, DILOG consistently outperforms other models.

The better performance of DILOG on one-shot learning comes from the **language bias** induced by the template that used to generate all the clauses, which can be regarded as a form of regularization. The ability of zero-shot domain transfer can be attributed to the **symmetry** assumed by DILOG. For example the slots of *parking* and *price* are symmetrical in a sense that the rules applied to *parking* should be directly applicable to *price* as well. DILOG only breaks symmetry when necessary (for example, to differentiate between user slots and system slots), while maintaining the symmetry otherwise. However, in the vector-form encoding used by the neural networks, it is difficult, if not impossible, to express this symmetry.

The rules learned by DILOG can be extracted and interpreted by human beings. For example, the rule learned for when to request a slot is:

$$\text{sys_request}(V0) \leftarrow \text{member_usr}(V0), \text{unknown}(V0),$$

which reads: If a slot is one of the user slots, and that slot is unknown, the system should request that slot. The full set of learned rules are listed in Appendix C.4. We also analyze an example that results in an error in Appendix C.5, which is made possible by the interpretability of the DILOG framework.

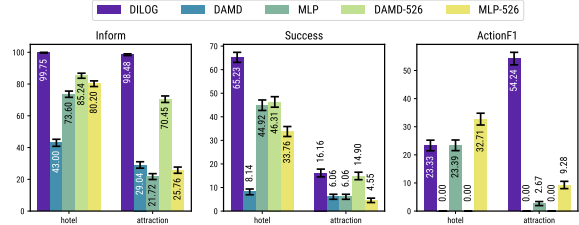
4 DILOG on the MultiWoZ Dataset

MultiWoZ 2.0 (Budzianowski et al., 2018) is a large-scale human to human English dialog dataset, which consists of dialogs on domains including restaurant, hotel, attraction, train, and taxi. It also includes dialogs that span multiple domains. The system response in MultiWoZ dataset is stochastic, as human clerks have the freedom to choose from the possible actions. The actions are annotated by human annotators, which are noisy as well. These features make MultiWoZ a more challenging task compared with SimDial. We used a variant of MultiWoZ 2.0 provided by ConvLab (Lee et al., 2019), which has annotated user actions.

The process of adapting the MultiWoZ dataset to DILOG is similar to that of SimDial (please refer to Appendix D for the details). For training DILOG, we selected a single representative dialog whose system action contains all possible intents (inform, request, offerbooked, nooffer) from the restaurant

Figure 2: The evaluation results on the test set of MultiWoZ: DILOG is compared with MLP and DAMD (Zhang et al., 2019). Note that DAMD applied its own heuristics in data preprocessing, so the Act. F1 of DAMD is not directly comparable with others. DAMD with the same preprocessing steps as others has a worse performance (Table A3). The Inform and Success metrics are not affected by preprocessing and therefore comparable. All stands for the performance on the all domains in the test set (1000 samples, **standard error: $\pm 1.6\%$**). Out-of-Domain is the average performance of hotel, attraction, train, and taxi (1480 samples, **standard error: $\pm 1.3\%$**). See Table A3 (appendix) for the per-domain results. The plot shows the performance on two example domains of hotel and attraction. -AM denotes adding the missing slots in NLG, and -526 denotes the model trained with all 526 dialogs from the restaurant domain.

	All			Out-of-Domain		
	Inform	Success	Act. F1	Inform	Success	Act. F1
DAMD	35.3	6.0	4.60	64.37	8.26	0.00
MLP	51.2	26.0	15.59	70.50	25.62	9.52
DILOG	91.5	43.4	23.67	95.92	65.49	23.54
DAMD-526	77.8	27.7	15.20	87.56	30.38	0.00
MLP-526	44.2	18.7	29.23	64.02	17.56	14.64
MLP-AM	51.4	43.3	15.59	70.44	42.15	9.52
DILOG-AM	91.4	90.2	23.67	95.60	91.24	23.54



domain. For testing, we use the original test set split in MultiWoZ dataset.

We compare DILOG with two baselines: DAMD (Zhang et al., 2019) is the state-of-the-art model with dialog action prediction, and MLP is a naive baseline. We calculate the following metrics to evaluate the quality of the models: **Inform** is a per diaog metric that measures whether the system finally provides a correct entity, **Success** measures whether all the requested information are provided, and **Action F1** is a per turn metric that checks if the predicted dialog action (both intent and entities) matches the ground truth. We use a template-based NLG to convert dialog actions predicted by DILOG and MLP to compute inform and success on generated utterances. There are three slots (postcode, phone, address) that only appear in the system actions but not dialog states. We manually add an option to the template-based NLG to inform those slots whenever an entity is informed, and denote that variant as -AM, which significantly improves the success rate. We do not compare the -AM variant against other methods.

The results are shown in Figure 2. As can be seen, DILOG outperforms other models on the overall test set, which shows DILOG is capable of doing one-shot learning and zero-shot domain transfer even under noisy data. Compared with the MLP policy trained with all the 526 samples in the restaurant domain (denoted as -526), DILOG trained with one sample has a higher inform/success rate but lower action F1 score. Noticeably, DILOG with the addition of missing slots in NLG achieves 91.40 and 90.20 inform and success rate overall, which is higher than the state-of-the-art DAMD.⁴ This shows inform and success are incomplete metrics which do not penalize false positives. Hence, we add action F1 as an additional metric to complement inform and success. Note that action F1 cannot solely capture the performance either, since there may be multiple possible true actions in a given state.

5 Conclusion

In this paper, we introduce DILOG for resource constrained dialog policy learning and zero-shot domain transfer. Empirically, we demonstrate that DILOG outperforms strong neural baselines on SimDial and MultiWoZ datasets, while offering interpretability. We also provide an intuitive explanation on why DILOG shows these features. On the other hand, the DILOG framework has certain weaknesses, one being that it is computationally expensive when the template space gets large, where distributed training will be desired. Further, the program template used to generate all possible clauses needs to be hand-crafted, which is not a straightforward process. Another disadvantage is that real-valued inputs, such as confidence scores cannot be taken into account automatically. The future work would naturally focus on solving these shortcomings. One way might be to jointly, in a multi-task setting, predict the templates or meta-learn them (Minervini et al., 2020).

⁴Note that DAMD is a more complex model with a higher capacity, which tends to overfit on extremely small datasets. Hence, MLP outperforms DAMD under the one-shot learning setting. However, when trained using all the dialogs in the restaurant domain, DAMD significantly outperformed the MLP when tested in-domain, as expected. Also, we are training DAMD-526 solely on the 526 dialogs in Restaurant domain. This results in lower inform and success rates compared to DAMD trained on the entire MultiWoZ dataset, which are 89.2 and 77.9, respectively (Zhang et al., 2019).

Acknowledgements

The authors are thankful to Edward Grefenstette, Seungwhan Moon, Chinnadhurai Sankar, and Zhiguang Wang for constructive discussions about this work and the future directions, as well as Cristine Cooper, Drea Modugno, Sheeva Slovan, Heidi Young, and Kevin Ahlstrom for their feedback.

References

- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278 Licensed as CC-BY 4.0*.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset.
- Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention.
- Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017. Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars. *arXiv preprint arXiv:1709.07858*.
- Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64.
- Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*. Now Foundations and Trends.
- Matthew Henderson, B. Thomson, and S. Young. 2014. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 360–365.
- 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*.
- Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Xiang Li, Yaoqin Zhang, Zheng Zhang, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. Convlab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Diane J Litman and Shimei Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2-3):111–137.
- Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. 2020. Learning reasoning strategies in end-to-end differentiable proving. *arXiv preprint arXiv:2007.06477*.
- Stephen Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679.
- Jiahuan Pei, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation.
- 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*.
- Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. *arXiv preprint arXiv:1906.03520*.
- 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*.
- Igor Shalyminov, Sungjin Lee, Arash Eshghi, and Oliver Lemon. 2019. Few-shot dialogue generation without annotated data: A transfer learning approach. *arXiv preprint arXiv:1908.05854*.
- Vladimir Vlasov, Akela Drissner-Schmid, and Alan Nichol. 2018. Few-shot generalization across dialogue tasks. *arXiv preprint arXiv:1811.11707*.

- Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019. Alternating roles dialog model with large-scale pre-trained language models.
- 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 and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. *arXiv preprint arXiv:1810.00278 SimDial is licensed under the Apache License 2.0*.
- 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*.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2002.10348*.

A Authoring the Program Template

A program template needs to be given to a ILP system, in order for it to generate the suitable rules. It is a general part of a system that does not need to be designed for each domain. Manual tuning is often required to get the desired program template. For example, a rule template is one of the important hyperparameters in the program template. A rule template is defined by the number of existentially quantified variables $v \in \mathbb{N}$ and whether to allow intensional predicates $i \in \{0, 1\}$. An **existentially** quantified variable is a variable which is interpreted as "there exists ...". An **intensional** predicate is a predicate defined by a set of clauses. Intuitively, the larger v , the more candidate rules will be generated, and thus easier for the model to overfit. Similarly, allowing intensional predicates will increase the number of rules generated. The hyperparameters can then be manually tuned to achieve the best performance.

Alternatively, we can define the complexity of the rule template to be the number of candidate clauses generated under this rule template. One can order all the possible rule templates by their complexities, and do iterative search starting from the least complex rule template. It is also possible to perform a blackbox optimization on the hyperparameters in order to save the computation.

B Learning the Relationship of `all`

The following example is used to learn the `all` relationship, which stands for all the items in a list satisfy some property.

The background knowledge \mathcal{B} is

```
terminal(t)
succ(a, b)
succ(b, c)
succ(c, d)
succ(d, e)
succ(e, t)
succ(f, g)
succ(g, h)
succ(h, t)
true(a)
true(c)
true(d)
true(e)
true(f)
true(g)
```

The list is presented in a format of linked nodes, where $\text{succ}(A, B)$ means the successor of A is B , $\text{terminal}(A)$ means A is a terminal state, and $\text{true}(A)$ specifies that the property under consideration holds for item A . In this example, the property does not hold for b and h . The positive examples are

```
all(c), all(d), all(e)
```

and the negative examples are

```
all(a), all(b), all(f), all(g), all(h)
```

The learned clauses for `all` are

```
pred1(V0, V1) <- succ(V0, V1), all(V1)
pred1(V0, V1) <- succ(V0, V1), terminal(V1)
all(V0) <- true(V0), pred1(V0, V1)
```

C Preprocessing Details for SimDial

C.1 Data Conversion

We illustrate the data conversion process using the following dialog:

SYS: Welcome to restaurant recommendation system. How can I help?
 USR: What's up? I need a restaurant.
 SYS: Which place?
 USR: I am at San Jose.
 SYS: What kind of food do you like?
 USR: Chinese.
 SYS: {"QUERY": {"#loc": "San Jose", "#food_pref": "Chinese"}, "GOALS": ["#default"]}
 USR: {"RET": {"#default": "6"}}
 SYS: Restaurant 6 is a good choice. Are you happy about my answer?
 USR: One more thing. Tell me if the restaurant is open.
 SYS: {"QUERY": {"#loc": "San Jose", "#food_pref": "Chinese"}, "GOALS": ["#open"]}
 USR: {"RET": {"#open": "open"}}
 SYS: The restaurant is open. Are you happy about my answer?
 USR: I have more requests. How expensive it is?
 SYS: {"QUERY": {"#loc": "San Jose", "#food_pref": "Chinese"}, "GOALS": ["#price"]}
 USR: {"RET": {"#price": "expensive"}}
 SYS: The restaurant serves expensive food. Anything else?
 USR: All good. That's all.
 SYS: Goodbye.

To adapt the DILP framework to the SimDial problem, the delexicalized states are converted to the form of atoms. For example, the state of the second turn is

```

'usr_slots': [{'name': '#loc', 'max_conf': 0.99, 'max_val': 'San Jose'},
              {'name': '#food_pref', 'max_conf': 0.0, 'max_val': None}],
'sys_goals': [{'name': '#default', 'delivered': False, 'value': None,
              'expected': None, 'conf': 1.2},
              {'name': '#open', 'delivered': False, 'value': None,
              'expected': None, 'conf': 0.0},
              {'name': '#price', 'delivered': False, 'value': None,
              'expected': None, 'conf': 0.0},
              {'name': '#parking', 'delivered': False, 'value': None,
              'expected': None, 'conf': 0.0}]
  
```

It was converted to the following grounding atoms:

```

known(loc)
unknown(food_pref)
terminal(term)
known(usr_slot)
usr_slots(usr_slot)
succ(usr_slot, food_pref)
succ(food_pref, loc)
succ(loc, term)
unknown(default)
unknown(open)
unknown(price)
unknown(parking)
  
```

Note that we use a linked list to enumerate the user slots, where $\text{succ}(A, B)$ means the successor of A is B , and $\text{terminal}(A)$ means A is a terminal state.

User and system actions are converted to the form of atoms in a similar way, for example, the user action of

```

[{'act': 'inform', 'parameters': [['#loc', 2]]}]
  
```

is converted to `inform(loc)`, and the system action of

```
[{'act': 'request', 'parameters': [['#food_pref', None]]}]
```

is converted to `sys_request(food_pref)`.

The constants \mathcal{C} are all the slot values appeared in the state and actions:

```
loc
food_pref
default
open
price
parking
term
usr_slot
```

At each turn, the user action and state are combined to be the background knowledge \mathcal{B} , the constants \mathcal{C} are all the possible slots, the positive examples \mathcal{P} are the system actions, and the negative examples \mathcal{N} are all other actions.

C.2 Domain Transfer

The model induced from the restaurant domain is tested on the other three domains of movie, weather and bus. The user slots and system goal slots can be different. For example, in the movie domain, the slots are

```
genre
years
default
country
rating
company
director
term
usr_slot
```

We can still follow the same conversion procedure for the restaurant domain. For example, the background \mathcal{B} of

```
known(genre)
known(years)
unknown(country)
terminal(term)
known(usr_slot)
usr_slots(usr_slot)
succ(usr_slot, genre)
succ(genre, years)
succ(years, country)
succ(country, term)
unknown(default)
unknown(rating)
unknown(company)
unknown(director)
inform(years)
request(default)
```

can be mapped to the positive example of `sys_request(country)`.

C.3 Results

The evaluation results on the SimDial dataset for different domains are shown in Table A1. The performance of different domains are similar.

	restaurant		movie		bus		weather	
	Intent F1	Entity F1	Intent F1	Entity F1	Intent F1	Entity F1	Intent F1	Entity F1
ZSDG	91.41	13.48	80.61	0.17	78.30	0.65	90.34	0.94
MLP	91.81	87.52	54.07	0.0	59.78	0.0	69.82	16.51
DILOG	99.74	99.78	99.83	99.89	99.63	99.72	99.78	99.82
ZSDG-1000	99.98	99.25	80.70	45.98	81.20	72.09	87.59	78.82
MLP-1000	98.97	97.98	54.07	0.00	59.78	0.00	49.63	19.88

Table A1: The evaluation results on the test set of SimDial, DILOG is compared with MLP and ZSDG (Zhao and Eskenazi, 2018). The models are trained on a dialog from the restaurant domain and evaluated on all other domains. We also included the ZSDG and MLP model trained with 1000 samples from the restaurant domain as an upper bound.

Interaction	True Action	Predicted Action
USR: Hi. I need a restaurant. SYS: Which city are you interested in?	sys_request(loc) sys_request(food_pref)	sys_request(loc) sys_request(food_pref)
USR: I am at Boston. SYS: What type of restaurant?	sys_request(food_pref)	sys_request(food_pref)
USR: French restaurant. SYS: {"QUERY": {"#loc": "Boston", "#food_pref": "French"}, "GOALS": [{"#default"}]}	sys_query(default)	sys_query(default)
USR: {"RET": {"#default": "56"}} SYS: Restaurant 56 is a good choice. Anything else?	sys_inform(default)	sys_inform(default)
USR: I like Chinese food. SYS: {"QUERY": {"#loc": "Boston", "#food_pref": "Chinese"}, "GOALS": [{"#default"}]}	sys_query(default)	None
USR: {"RET": {"#default": "94"}} SYS: Restaurant 94 is a good choice. What else can I do?	sys_inform(default)	sys_inform(default)
USR: One more thing. How easy is it to park? SYS: {"QUERY": {"#loc": "Boston", "#food_pref": "Chinese"}, "GOALS": [{"#parking"}]}	sys_query(parking)	sys_query(parking)
USR: {"RET": {"#parking": "valet parking"}} SYS: The restaurant has valet parking. What else can I do?	sys_inform(parking)	sys_inform(parking)
USR: One more thing. How expensive it is? SYS: {"QUERY": {"#loc": "Boston", "#food_pref": "Chinese"}, "GOALS": [{"#price"}]}	sys_query(price)	sys_query(price)
USR: {"RET": {"#price": "expensive"}} SYS: The restaurant serves expensive food. Anything else?	sys_inform(price)	sys_inform(price)
USR: No more questions. That's all		

Table A2: An example that DILOG predicted a wrong dialog action.

C.4 Learned rules

The rules learned by DILP are simple and can be extracted and interpreted by humans. In particular, DILP converges to these set of rules with probabilities close to one except for the predicate of `sys_inform`:

$$\begin{aligned}
 \text{sys_request}(V0) &\leftarrow \text{member_usr}(V0), \text{unknown}(V0), \\
 \text{sys_inform}(V0) &\leftarrow \text{kb_return}(V0), \\
 \text{sys_query}(V0) &\leftarrow \text{request}(V0), \text{pred3}(V0), \\
 \text{pred2}() &\leftarrow \text{all}(V0), \text{usr_slots}(V0), \\
 \text{pred3}(V0) &\leftarrow \text{pred2}(), \text{unknown}(V0),
 \end{aligned}$$

where `pred2` and `pred3` are *invented* predicates used as intermediate states. As an example, the first rule reads: If a slot is one of the user slots, and that slot is unknown, the system should request that slot.

C.5 A failure case

While DILOG achieves 99+% Intent F1 and Entity F1 scores, we wanted to understand the rare cases that lead to failures. A lemon-picked dialog in which DILOG failed to predict the correct system action is listed in Table A2.

The error happens in the 5th turn when user corrected the `food_pref` slot he previously provided. The true system action should be `sys_query(default)`, while the predicted one is `None`, which means no action should be taken. We can also pin-point the erroneous rule to be `pred3`. All these error analysis are made possible by the explainable rules from the DILP framework. By fixing these errors (retraining on scenarios where such slot correction is provided), we expect DILOG to fully learn SimDial with 100% Intent F1 and Entity F1 scores. However, that is not the point of this study.

D Preprocessing Details for MultiWoZ

We illustrated the data conversion process for MultiWoZ using the following dialog:

USR: I am looking to get some eritrean food in the west part of town.

SYS: Unfortunately there aren't any restaurants that fit your criteria.

Is there anything else that I can do for you?

USR: Can you see if there are any that serve indian food?

SYS: There are six restaurants, one in the moderate price range and the rest in the expensive price range.

Which one would you prefer?

USR: Any would be fine. But I need to book a table for 8 people for Sunday at 12:15.

SYS: Booking was unsuccessful at rice boat. Would you like to find another restaurant?

USR: Sure as long as it's in the same area and price range. I need the reference number too please.

SYS: Okay I booked you at the Tandoori Palace and your reference number is NBLD7E0V.

USR: Thank you, that's all I needed today!

SYS: Okay glad I could be of help!

In the first turn, the state of restaurant

```
'restaurant':
  {'book': {'booked': [], 'people': '', 'day': '', 'time': ''},
   'semi': {'food': 'eritrean',
            'pricerange': 'not mentioned',
            'name': 'not mentioned',
            'area': 'west'}}
```

is converted to the following grounding atoms:

```
"usr_inform(food) ",
"usr_inform(area) ",
"known(food) ",
"unknown(price) ",
"unknown(name) ",
"known(area) ",
"unknown(people) ",
"unknown(day) ",
"unknown(time) "
```

We also included the information of whether the book or match is successful from the database pointers. When there is no match or booking failed, `no_match()` or `book_fail()` is added to the state as well.

In the conversion of actions, we ignored all actions in the general domain including greeting, bye, thanks, etc. Besides, the system actions of `select`, `recommend`, and `offerbook` are treated to be the same as `inform`, as they are all providing information to the user. For example, the user actions of

```
[[ 'inform', 'restaurant', 'food' ],
 [ 'inform', 'restaurant', 'area' ]]
```

are converted to the atoms of `inform(food)` and `inform(area)`, while the system actions of

```
[[ 'nooffer', 'restaurant', 'none' ],
 [ 'reqmore', 'general', 'none' ]]
```

are converted to the atoms of `nooffer()`.

The construction of the sample $(\mathcal{B}, \mathcal{P}, \mathcal{N}, \mathcal{C})$ is similar to that of SimDial. Note that during training, we ignored all the domain information. During inference time, we separate the belief state by domains, and run inference on each domain using the same model. Finally, the predicted actions for each domain are combined to yield the final action prediction.

D.1 Results

The evaluation results for different domains on the MultiWoZ dataset are shown in Table A3. The performance of different domains are similar except train and taxi, this is because the goals in those two domains are less diverse. DAMD applied its own heuristics in data preprocessing, so the Act. F1 of DAMD is not directly comparable with others. We also included DAMD with the same data-preprocessing as the others (denoted as DAMD'), whose performance is worse.

	restaurant			hotel			attraction			train			taxi		
	inform	success	action F1	inform	success	action F1	inform	success	action F1	inform	success	action F1	inform	success	action F1
DAMD	50.11	9.15	2.81	43.0	8.14	0.00	29.04	6.06	0.00	85.43	18.83	0.00	100.0	0.00	0.00
MLP	83.30	27.92	23.74	73.60	44.92	23.39	21.72	6.06	2.67	86.67	51.52	12.01	100.0	0.00	0.00
DILOG	98.17	60.41	27.62	99.75	65.23	23.33	98.48	16.16	54.24	85.45	84.65	16.26	100.0	95.90	0.33
MLP-AM	84.44	40.05	23.74	73.35	52.28	23.39	21.72	17.42	2.67	86.67	51.72	12.01	100.0	47.18	0.00
DILOG-AM	99.08	91.30	27.62	98.98	93.65	23.33	97.98	90.15	54.24	85.45	84.24	16.26	100.0	96.92	0.33
DAMD-526	93.59	51.72	13.65	85.24	46.31	0.00	70.45	14.90	0.00	94.53	60.32	0.00	100.0	0.00	0.00
MLP-526	89.02	31.12	52.83	80.20	33.76	32.71	25.76	4.55	9.28	50.10	31.92	16.56	100.0	0.00	0.00
DAMD'	21.97	1.83	0.00	29.70	3.05	0.00	23.74	4.04	0.00	85.66	10.91	0.00	100.0	0.00	0.00
DAMD'-526	92.22	22.88	14.02	88.58	19.54	0.00	68.70	14.14	0.00	94.75	40.20	0.00	100.0	1.03	0.00

Table A3: The evaluation results on the test set of MultiWoZ, DILOG is compared with MLP and DAMD (Zhang et al., 2019). The models are trained on a dialog from the restaurant domain and evaluated on all other domains. We also included the DAMD and MLP model trained with all 526 samples from the restaurant domain as an upper bound.