Taming Continuous Posteriors for Latent Variational Dialogue Policies

Marin Vlastelica¹, Patrick Ernst², Gyuri Szarvas²

¹Autonomous Learning Group, Max Planck Institute for Intelligent Systems, Tübingen, Germany
²Amazon Development Center Germany GmbH, Berlin, Germany
marin.vlastelica@tuebingen.mpg.de, {peernst, szarvasg}@amazon.de

Abstract

Utilizing amortized variational inference for latent-action reinforcement learning (RL) has been shown to be an effective approach in Task-oriented Dialogue (ToD) systems for optimizing dialogue success. Until now, categorical posteriors have been argued to be one of the main drivers of performance. In this work we revisit Gaussian variational posteriors for latent-action RL and show that they can yield even better performance than categorical. We achieve this by introducing an improved variational inference objective for learning continuous representations without auxiliary learning objectives, which streamlines the training procedure. Moreover, we propose ways to regularize the latent dialogue policy, which helps to retain good response coherence. Using continuous latent representations our model achieves state of the art dialogue success rate on the MultiWOZ benchmark, and also compares well to categorical latent methods in response coherence.

1 Introduction

Task-oriented Dialogue (ToD) systems have reached a degree of maturity, which enables them to engage with human users and assist them in various tasks. They are able to steer natural-language conversations in order to complete users’ goals, such as booking restaurants, querying weather forecasts and resolving customer service issues. At their core, the behavior of these systems is controlled by a dialogue policy, which receives user inputs in the form of utterances and additional features or states. Template-based methods (Walker et al. 2007; Inaba and Takahashi 2016) leverage ranking or classification approaches to select the most fitting response from a pre-defined set of responses, i.e. templates. While template-based methods offer better control over the dialogue policy behavior, they are less versatile due to their dependency on template sets. Moreover, constructing comprehensive template sets is a challenge in itself (Gao, Galley, and Li 2019). In retrieval-based approaches (Yan, Song, and Wu 2016; Henderson et al. 2019; Tao et al. 2019) candidate responses are not pre-defined, but are retrieved from massive dialogue corpora, e.g. by executing ad-hoc search queries a priori. Generative models do not require such additional inputs as prior knowledge. They enable end-to-end (E2E) learning of dialogue policies and have the potential to generate diverse responses by leveraging a large vocabulary (Serban et al. 2017; Zhao, Zhao, and Eskenazi 2017; Gu et al. 2018). Even though such fully data-driven approaches offer great versatility and a faster adoption, they may exhibit degenerate behavior by generating incomprehensible utterances. This is apparent in multi-turn dialogues, which span hundreds of words, while the success signal is only observed at the end of dialogues. Reinforcement learning (RL) based approaches are able to optimize for such long-term, sparse rewards and have been applied in this setup. Previous approaches prominently applied word-level RL (Lewis et al. 2017; Kottur et al. 2017), where the action space is defined over the entire vocabulary. The response utterances are then generated auto-regressively by consecutive next-word predictions. Unfortunately, the use of large action spaces often impedes the convergence of policy learning algorithms, which makes it hard to ensure coherent responses. Prior work makes use of latent-action RL to address the dimensionality problem by utilizing variational inference approaches (Zhao, Xie, and Eskenazi 2019; Lubis et al. 2020). These methods rely on a supervised learning stage, where latent action representations are learned over response utterances, followed by fine-tuning via reinforcement learning in the latent space.

In this paper we follow this paradigm, but extend prior work substantially by introducing the TCUP approach, which aims to Tame Continuous Posteriors for latent variational dialogue policies. TCUP makes the following contributions: (i) A new formulation of the variational inference objective for learning continuous latent response representations without auxiliary learning objectives. (ii) A more robust approach for learning from ToD data in an offline RL setup, which utilizes the fact that we are dealing with expert dialogue trajectories. Using the MultiWOZ benchmark (Budzianowski et al. 2018), we show that TCUP is able to improve the state-of-the-art performance across different metrics. In addition to MultiWOZ’s context-to-text metrics and following Lubis et al. (2020), we demonstrate the benefits of the learned continuous latent representations quantitatively, using a clustering analysis.

2 Preliminaries

Latent-action reinforcement learning methods are trained in two stages: In the first stage, which we denote as SL Stage, an encoder-decoder architecture is trained using a supervised
We denote a dialogue context as $c$, which contains the dialogue history and state. Response utterances are denoted as $x$. Provided a dataset of context and optimal response pairs $(c, x)$, we want to extract a latent representation $z$, representing the dialogue responses given context. Approaches based on variational inference have shown to be beneficial for learning such latent representations. This is done by optimizing the evidence lower bound (ELBO)

$$
\mathcal{L}(\phi, \theta) = \mathbb{E}_{q_{\phi}(z|x,c)}[\log p_{\theta}(x|z)] + D_{\text{KL}}[q_{\phi}(z|x,c)||p(z|c)],
$$

where $q_{\phi}$ denotes the variational posterior parameterized by $\theta$ and $p_{\theta}$ the decoder parameterized by $\phi$. Prior work (Zhao, Xie, and Eskenazi 2019) argues that this full ELBO formulation suffers from “explaining away”, i.e. without additional incentive, the encoder only relies on $x$ for computing $z$. To mitigate this overexposure bias to responses, i.e. the insensitivity to the context, the authors introduce the “lite” ELBO:

$$
\mathcal{L}(\phi, \theta) = \mathbb{E}_{q_{\phi}(z|x,c)}[\log p_{\theta}(x|z)] + D_{\text{KL}}[q_{\phi}(z|x,c)||p(z)].
$$

The goal is to be closer to the information available during testing time, where the decoder only sees $z$ conditionally sampled on the context $q_{\phi}(z|x)$.}

### 2.2 RL Stage – Learning Latent Dialogue Policies

After learning to extract a compressed representation $z$ in the supervised learning stage, the encoder $q_{\phi}(z|x,c)$ is fine-tuned via reinforcement learning to optimize the dialogue reward, which is typically based on a dialogue success / completion signal. A Markov Decision Process (MDP) is defined as a tuple $(\mathcal{S}, \mathcal{A}, r, p)$ of state space $\mathcal{S}$, action space $\mathcal{A}$, reward function $r$, and transition density $p$. The general goal of reinforcement learning is to optimize the expected return of policy $\pi$, denoted as $\mathbb{E}[J(\pi)]$. Many works utilize a Monte-Carlo estimate of the policy gradient

$$
\nabla_{\phi} \mathbb{E}[J(\pi_{\phi})] = \mathbb{E} \left[ \nabla_{\phi} \sum_{(s,a) \in \tau} \log \pi_{\phi}(a|s)r(s,a) \right].
$$

The expectation is taken over the trajectory distribution $\eta^\tau(\tau)$, i.e. distribution of sequences of $(s, a)$. In the context of ToD, we may cast the state $s$ as being the underlying dialogue state which may be unobserved, in which case the problem becomes partially observable. In ToD systems based on word-level RL an action $a$ corresponds to predicting the next word out of a large vocabulary, where in latent-action RL $a$ is cast to predicting one element of a latent-space representation. In our setting, the latent dialogue policy is warm-started by the parameters of the variational posterior from the SL stage.

### 3 Method

Our method, called TCUP, is in line with the general latent action RL paradigm outlined in Sec. 2, i.e. TCUP is also based on a two-stage approach. However, we propose contributions to both stages, which significantly extend prior work. In Sec. 3.1 we describe a reformulation of the ELBO for the SL Stage, which is vital to reach state-of-the-art performance with continuous variational posteriors. In Sec. 3.2 we discuss our reinforcement learning setup and introduce methods to address the deterioration of response coherence in latent action-based dialogue policies through regularization.

#### 3.1 SL Stage – Revisiting the Full ELBO

As aforementioned, prior work by Zhao, Xie, and Eskenazi (2019) introduced the “lite” ELBO (described in equation 2), to alleviate the overexposure bias that emerges when response information is incorporated in the optimization objective. Indeed, if we only optimize for samples of $q(z|x,c)$, the most information about the responses $x$ is found within the responses themselves, which could encourage the model to ignore the context $c$. However, not conditioning on $x$ will reduce the expressiveness of the variational posterior for capturing generative factors of responses and for predicting the best next response given contexts. In this work we rely on a full ELBO and introduce a conditional prior $q_{\phi}(z|x)$ indicated by the superscript $p$ which is constrained by the free prior $p(z)$ and also parametrized.

**Proposition 1.** For the minimization problem $\min_{q_{\phi}} D_{\text{KL}}[q_{\phi}||p(z|x,c)]$, given sufficiently similar $p(z|x,c)$ and $p(z|x)$, it suffices to minimize

$$
D_{\text{KL}}[q_{\phi}||q_{\phi}^p] + D_{\text{KL}}[q_{\phi}^p||p(z)] - \mathbb{E}_{q_{\phi}}[\log p_{\theta}(x|z)] - \mathbb{E}_{q_{\phi}^p}[\log p_{\theta}(x|z)] + C,
$$

with respect to $q_{\phi}$ and $q_{\phi}^p$. 

![SL Stage Schema](image-url)
We achieve this by leveraging samples from the prior and we rely on the BLEU metric as specified in the MultiWOZ benchmark; an ablation study to assess the importance of our technical contributions from Sec. 3; and a detailed evaluation of TCUP’s dialogue policy in Sec. 4.1. This includes comparing its performance on the MultiWOZ benchmark; an ablation study to assess the importance of our technical contributions from Sec. 3; and a qualitative analysis of response coherence. Finally, we analyze the quality of the latent representations in Sec. 4.2 and provide evidence that TCUP learns representations, which yield a good clustering of MultiWOZ domains and actions.
### Table 1: Competitor comparison on MultiWOZ. RL methods are marked with $\dagger$, transformer architectures with $\ast$.

| Model       | BLEU | Inform | Success | Av. len. | CBE | # unigrams | # trigrams |
|-------------|------|--------|---------|----------|-----|------------|------------|
| MarCo$^\ast$ | 17.3 | 94.5   | 87.2    | 16.01    | 1.94| 319        | 3002       |
| HDSA        | **20.7** | 87.9   | 79.4    | 14.42    | 1.64| 259        | 2019       |
| HDNO$^\dagger$ | 17.8 | 93.3   | 83.4    | 14.96    | 0.84| 103        | 315        |
| SFN$^\dagger$ | 14.1 | 93.4   | 82.3    | 14.93    | 1.63| 188        | 1218       |
| UniConv     | 18.1 | 66.7   | 58.7    | 14.17    | 1.79| 338        | 2932       |
| LAVA$^\dagger$ | 16.2 | 89.7   | 77.6    | 14.41    | 1.96| 272        | 2365       |
| LAVA$^\dagger$-Cat | 10.8 | 95.9   | 93.5    | 13.28    | 1.27| 176        | 708        |
| TCUP$^\dagger$ | 10.3 | **96.3** | **95.9** | 15.14    | 1.44| 210        | 1838       |
| TCUP$^\dagger$-Cat | 14.3 | 96.1   | 92.1    | 14.33    | 1.61| 230        | 1490       |

### Table 2: Example of successful responses, which are generated by a naive system policy “gaming” the metric and producing incoherent utterances with low BLEU scores together with the corresponding TCUP responses.

**USER** we are staying [value_count] people for [value_count] nights starting from [value_day] i need the reference number?

**NAIVE SYSTEM** [hotel_name] [hotel_name] [hotel_name] [hotel_phone] . . .

**TCUP** [hotel_name] is located at [hotel_address]. their phone number is [hotel_phone]. is there anything else i can help with?

**USER** just the address please

**NAIVE SYSTEM** [attraction_name] is located at [attraction_address]. postcode [attraction_postcode]; the phone number is [attraction_phone]. [attraction_name] is located . . .

**TCUP** [attraction_name] s phone number is [attraction_phone]. it s located on [attraction_address], and their phone number is [attraction_phone].

We use a recurrent encoder-decoder architecture with dot-product attention, details on architecture and training are in the Appendix (Vlastelica, Ernst, and Szarvas 2022).

**SL Stage** Here, we learn the mapping from context to response using equation 1. In practice, we optimize the $p_c(x|z)$ part with a weighted cross-entropy loss that puts higher weights on slot placeholders in the response. The best model is selected based on its BLEU score.

**RL Stage** In this stage we fix the parameters of the decoder and train only the encoder parameters via policy gradient with two important modifications: (i) we use a batched version of the policy gradient which contains samples from multiple dialogues and hence reduces the variance of the gradient, and (ii) in each batch we sample a mix of newly generated and old experience (with which we have obtained a success signal earlier). This implicitly keeps us close to the starting policy which results in more stable training. It is beneficial to replace the standard $D_{KL}(q||p)$ term of the variational objective with a symmetric version $\frac{1}{2}(D_{KL}(q||p) + D_{KL}(p||q))$. This ensures that regions where the densities of $p$ and $q$ behave differently are treated equally irrespective of the ordering.

### 4.1 Context-to-Response Generation

We evaluate TCUP using MultiWOZ 2.1 (Wang et al. 2020) on the policy learning task for context to response generation.

MultiWOZ contains 10438 dialogues across six different domains, pre-split into 8438 training, 1000 validation, and 1000 testing records. The task’s objective is to generate the next response at every system turn given the ground-truth dialog belief state and the previous dialogue. For a fair comparison we use the same delexicalization approach as in prior work (Lubis et al. 2020). Additionally to our proposed approach based on isotropic Gaussians, we introduce a variant which is based on categorical latent distributions (coined TCUP-Categorical). Table 1 shows that TCUP improves the state-of-the-art inform- and success rate metric across all competitors. Moreover, it is also competitive in terms of language diversity metrics. Compared to the currently best performing latent action reinforcement learning approach (LAVA), we increase all metrics except for minor decreases in the BLEU metric. The response coherence is further discussed in Sec. 4.1.

**Latent Representations** While prior latent-action RL approaches (Lubis et al. 2020; Zhao, Xie, and Eskenazi 2019) favor categorical latent distributions with modified attention mechanisms in the decoder, our results demonstrate that relying on isotropic Gaussian latents is advantageous, even if we only use simple dot-product attention in the decoder. Focusing on TCUP-Categorical we observe competitive results in terms of inform and success rate compared to prior work, but inferior performance compared to TCUP with continuous latents. Using categoricals limits us in the sense that we don’t make full use of the nonlinearity in the decoder to decode diverse responses, which manifests itself with poor results in

---

*Results taken for best runs, mean performance is lower, discussion in the Appendix (Vlastelica, Ernst, and Szarvas 2022).*
Ablation Study

Fig. 2 presents an ablation study over different variants of TCUP. Variants only using the SL Stage (sl) have the highest BLEU metrics, but show low success rates compared to approaches with a RL Stage (sl + rl). Our experiments demonstrate that the proposed variational inference objective Sec. 3.1, i.e. the full ELBO with prior and posterior shuffling (sl shuff) and an identity covariance Gaussian prior on the contextual prior (sl pp shuff) leads to improvements across all metrics compared to the objective used in prior work (sl larl) LaRL (Zhao, Xie, and Eskenazi 2019) and LAVA (Lubis et al. 2020). Among the two, sl shuff performs better in the success and inform metrics, where sl pp shuff delivers minor improvements in BLEU. Through adding the RL Stage (sl + rl), significantly higher success rates are achieved while sacrificing BLEU. The BLEU decrease is alleviated by applying optimal replay (all variants with rep in their name) and the constrained rewards with the KL penalty (kl), with replay being the superior choice between the two. For instance, compare the scores of sl + rl ctxp rep and sl + rl ctxp kl. Further gains can be achieved by utilizing both sl + rl pp shuff kl rep. Not using optimal replay introduces higher variance across runs. Considering the variants with highest bi- and trigrams (sl + rl kl, sl + rl ctxp kl) we observe that high scores in these metrics are not indicative of high coherence, if we compare with their corresponding BLEU scores. In summary, the results (shuff kl rep) validate the benefits of our variational inference objective and regularization.

Response Coherence

We observed that it is possible to maximize success rates at the cost of lower coherence in terms of BLEU scores. Fig. 3a shows that we achieve state-of-the-art success rates, whilst achieving a BLEU score of effectively 0. This is an artifact of the success rate metric, i.e. it disregards response coherence and only checks if the correct slot values have been addressed by the dialogue policy. In Fig. 3b we see different runs of our method with different strength of regularization in terms of KL penalty and optimal replay fraction. Depending on the strength of regularization they form a Pareto front. This further shows the multi-objective trade-off between success rates and BLEU scores. The BLEU collapse is to be expected when using latent-action reinforcement learning, since longer and more diverse answers have positive impact on the dialogue success, which leads to the policy selecting degenerate responses. Furthermore, if no additional regularization is used, the latent policy learns to sample outliers in terms of the prior over z. We argue that it is not realistic to expect the policy resulting from the reinforcement learning stage to outperform the response coherence of the supervised learning stage in terms of BLEU (as long as the primary purpose of the RL stage is to improve dialogue success metrics). Instead, we show that through regularization and optimal replay, the BLEU score can be kept from deterioration through the RL stage while we optimize dialogue level metrics. An alternative approach to alleviating degenerate policies would be to simply make the BLEU score or other coherence metrics part of the reward function. For example, the final comparison of models in (Budzianowski et al. 2018) is done by comparing

the diversity metrics (Tab. 1). By using a continuous distribution, we improve the diversity of the generated responses compared to those two competitors.

Figure 2: Ablation study. TCUP configurations are averaged over three random seeds. sl denotes supervised training, sl+rl supervised stage followed by reinforcement learning sl+larl work in (Zhao, Xie, and Eskenazi 2019), ctxp contextual prior without prior on it (q_p, equation 4), pp identity covariance Gaussian prior on contextual prior (D_{KL}[q_p∥p(z)], equation 4), shuff prior and posterior shuffling (Sec. 3.1), rep replay buffer (Sec. 3.2), kl penalty term to identity covariance Gaussian (Sec. 3.2)
Appendix (Vlastelica, Ernst, and Szarvas 2022). (Fig. 4b) without applying regularization in the RL stage. In comparison to the results reported by Lubis et al. (2020), our model is able to obtain high scores in the supervised stage of training already. The scores drop slightly after RL fine-tuning, but remain at a higher level than those for categorical latents. It’s important to note that the scores for LaRL* and LAVA* were computed with such categoricals, whereas our scores are based on Gaussian latents, which are continuous and unbounded. Tab. ?? demonstrates the value in using continuous latent representations, rather than to make direct numeric comparisons.

5 Related Work

Supervised Learning The majority of prior work applies some form of Supervised Learning. For an extensive overview we refer to Gao, Galley, and Li (2018) and focus on the state-of-the-art competitors HDSA (Chen et al. 2019) as well as UniConv (Le et al. 2020). Both approaches demonstrate the benefits of jointly training multiple dialogue tasks at once, such as predicting dialogue acts and states.

Variational Inference Inspired by variational autoencoders (VAEs) (Kingma and Welling 2013), several works employ variational inference for learning conditional response distributions. VHRED (Serban et al. 2017) is a variational hierarchical RNN for modeling dependencies between words and utterances. Stochastic latent variables are introduced to generate the next utterance. To avoid collapsing posteriors Zhao, Zhao, and Eskenazi (2017) learn response representations using an auxiliary task introducing a bag of words loss. Shen et al. (2018) improves the stability of VHRED by splitting the training process into two parts. First, text is auto-encoded into continuous embeddings, which are the starting point for learning latent representations by reconstructing the embeddings. To address the degeneration problem, Variational Hierarchical Conversation RNNs (VHCR) (Park, Cho,
Figure 4: UMAP embeddings of latent representations. Figures a-c: domain labeled embeddings for the SL, SL+RL and SL+RL with replay sampling, bottom row: same embeddings labeled by response type. Executing the RL stage of the training results in representations that are difficult to separate (b and e). By applying replay we obtain higher specialization in the clusters (c and f).

and Kim 2018) exploits an utterance drop regularization. DialogWAE (Gu et al. 2018) represents the prior distribution as Gaussian mixture and adapts WGAN (Arjovsky, Chintala, and Bottou 2017) for training.

Combined Supervised- & Reinforcement Learning Henderson, Lemon, and Georgila (2008) was the first work to combine Reinforcement and Supervised Learning introducing a value function, which relies on SL for predicting the expected future reward of states not covered by the data directly. Williams, Asadi, and Zweig (2017) prevent high variance by performing a SL gradient update, if the RL policy output deviates from the training data. Fatemi et al. (2016); Su et al. (2017) apply two-stage approaches, where they pre-train a dialogue policy supervised, which is then further optimized by RL. HappyBot (Shin et al. 2019) relies on a weighted combination of a maximum likelihood and a REINFORCE objective. Saleh et al. (2020) uses REINFORCE-based policy gradients to update the prior probability distribution of the latent variational model trained using supervised learning. Structured Fusion Networks (SFNs) (Mehri, Srinivasan, and Eskenazi 2019) apply RL to fuse dialogue modules, where each module serves a different purpose (NLU, NLG, etc.) and is pre-trained in individual supervised stages. Apart from task-oriented dialogues a combination of SL and RL has also been applied for generating responses for open-domain dialogues (Xu, Wu, and Wu 2018). The aforementioned approaches are based on word-level RL, suffering from huge action-spaces covering the entire input vocabulary. Due to this, ensuring coherent responses is challenging (Lewis et al. 2017; Kottur et al. 2017), especially in multi-turn dialogues spanning hundreds of words. LaRL (Zhao, Xie, and Eskenazi 2019) and LAVA (Lubis et al. 2020) address this problem by learning a low-dimensional latent representation with amortized variational inference followed by RL fine-tuning.

6 Conclusion

We showed that with appropriate modifications to the reinforcement learning procedure and the latent space, it is possible to obtain state-of-the-art results with simple Gaussian latent distributions. The problem of coherence deterioration is alleviated with optimal replay and KL regularization, but points to the fact that better metrics are needed for developing efficient dialogue policies. TCUP shows promising results in the policy learning setting with access to ground-truth dialogue state, which would be interesting to transfer to the E2E learning setting.
References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In International conference on machine learning, 214–223. PMLR.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ—A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. arXiv preprint arXiv:1810.00278.

Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1): 1–27.

Chen, W.; Chen, J.; Qin, P.; Yan, X.; and Wang, W. Y. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. arXiv preprint arXiv:1905.12866.

Fatemi, M.; Asri, L. E.; Schulz, H.; He, J.; and Suleman, K. 2016. Policy networks with two-stage training for dialogue systems. arXiv preprint arXiv:1606.03152.

Gao, J.; Galley, M.; and Li, L. 2018. Neural approaches to conversational ai. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 1371–1374.

Gao, J.; Galley, M.; and Li, L. 2019. Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots. Now Foundations and Trends.

Gu, X.; Cho, K.; Ha, J.-W.; and Kim, S. 2018. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. arXiv preprint arXiv:1805.12352.

Henderson, J.; Lemon, O.; and Georgila, K. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. Computational Linguistics, 34(4): 487–511.

Henderson, M.; Vulić, I.; Gerz, D.; Casanueva, I.; Budzianowski, P.; Coope, S.; Spithourakis, G.; Wen, T.-H.; Mrkšić, N.; and Su, P.-H. 2019. Training neural response selection for task-oriented dialogue systems. arXiv preprint arXiv:1906.01543.

Inaba, M.; and Takahashi, K. 2016. Neural utterance ranking model for conversational dialogue systems. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 393–403.

Jiang, H.; Dai, B.; Yang, M.; Zhao, T.; and Wei, W. 2021. Towards Automatic Evaluation of Dialog Systems: A Model-Free Off-Policy Evaluation Approach. arXiv preprint arXiv:2102.10242.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.1172.

Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Natural language does not emerge naturally in multi-agent dialog. arXiv preprint arXiv:1706.08502.

Le, H.; Sahoo, D.; Liu, C.; Chen, N. F.; and Hoi, S. C. 2020. Uniconv: A unified conversational neural architecture for multi-domain task-oriented dialogues. arXiv preprint arXiv:2004.14307.

Lewis, M.; Yarats, D.; Dauphin, Y. N.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125.

Lubis, N.; Geishauser, C.; Heck, M.; Lin, H.; Moresi, M.; van Niekerk, C.; and Gašić, M. 2020. LAVA: Latent Action Spaces via Variational Auto-encoding for Dialogue Policy Optimization. In Scott, D.; Bel, N.; and Zong, C., eds., Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, 465–479. International Committee on Computational Linguistics.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mehri, S.; and Eskenazi, M. 2020. Unsupervised evaluation of interactive dialog with dialogpt. arXiv preprint arXiv:2006.12719.

Mehri, S.; Srinivasan, T.; and Eskenazi, M. 2019. Structured fusion networks for dialog. arXiv preprint arXiv:1907.10016.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Park, Y.; Cho, J.; and Kim, G. 2018. A hierarchical latent structure for variational conversation modeling. arXiv preprint arXiv:1804.03424.

Saleh, A.; Jaques, N.; Ghandeharioun, A.; Shen, J.; and Picard, R. 2020. Hierarchical reinforcement learning for open-domain dialog. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 8741–8748.

Serban, I.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31.

Shen, X.; Su, H.; Niu, S.; and Dembreg, V. 2018. Improving variational encoder-decoders in dialogue generation. In Proceedings of the AAAI conference on artificial intelligence, volume 32.

Shin, J.; Xu, P.; Madotto, A.; and Fung, P. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. arXiv preprint arXiv:1906.08487.

Su, P.-H.; Budzianowski, P.; Ultes, S.; Gašić, M.; and Young, S. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. arXiv preprint arXiv:1707.00130.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In Proceedings of the twelfth ACM international conference on web search and data mining, 267–275.

Vlastelica, M.; Ernst, P.; and Szarvas, G. 2022. Taming Continuous Posteriors for Latent Variational Dialogue Policies. arXiv preprint arXiv:2205.07633.

Walker, M. A.; Stent, A.; Mairesse, F.; and Prasad, R. 2007. Individual and domain adaptation in sentence planning for
dialogue. *Journal of Artificial Intelligence Research*, 30: 413–456.

Wang, K.; Tian, J.; Wang, R.; Quan, X.; and Yu, J. 2020. Multi-domain dialogue acts and response co-generation. *arXiv preprint arXiv:2004.12363*.

Williams, J. D.; Asadi, K.; and Zweig, G. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*.

Xu, C.; Wu, W.; and Wu, Y. 2018. Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*.

Yan, R.; Song, Y.; and Wu, H. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 55–64.

Zhang, Y.; Ou, Z.; and Yu, Z. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9604–9611.

Zhao, T.; Xie, K.; and Eskenazi, M. 2019. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1208–1218. Minneapolis, Minnesota: Association for Computational Linguistics.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*. 

