SynTOD: Augmented Response Synthesis for Robust End-to-End Task-Oriented Dialogue System

Quang-Chieu Nguyen[†], Quang-Minh Tran[†], Khac-Hoai Nam Bui[*]
Viettel AI, Viettel Group, Vietnam
{chieunq, minhtq21, nambkh}@viettel.com.vn

Abstract

Task-oriented dialogue (TOD) systems are introduced to solve specific tasks, which focus on training multiple tasks such as language understanding, tracking states, and generating appropriate responses to help users achieve their specific goals. Currently, one of the remaining challenges in this emergent research field is the capability to produce more robust architectures fine-tuned for end-to-end TOD systems. In this study, we consider this issue by exploiting the ability of pre-trained models to provide synthesis responses, which are then used as the input for the fine-tuned process. The main idea is to overcome the gap between the training process and inference process during fine-tuning end-to-end TOD systems. The experiment on Multiwoz datasets shows the effectiveness of our model compared with strong baselines in this research field. The source code is available for further exploitation.

Keywords: task oriented dialogue, end-to-end system, synthesis data augmentation, pre-trained language models

1. Introduction

Different from open domain dialogue (ODD), which aims to provide smooth conversations with humans on various topics, task-oriented dialogue (TOD) systems are developed to assist users in achieving some specific goals such as hotel booking or restaurant recommendation (Fu et al., 2022). The traditional approach follows a modular pipeline architecture, which is divided into three separate components: natural language understanding (NLU), dialogue state tracking (DST), and natural language generation (NLG). Nonetheless, this approach has two main limitations (Chen et al., 2017): i) it is difficult to propagate when processing the end user's feedback; and ii) it requires significant human efforts to adapt to the new environments (e.g., training with new data). In this regard, many studies attempt to construct an end-to-end trainable framework for TOD systems based on large pre-trained language models (Hosseini-Asl et al., 2020; Yang et al., 2021, 2022). Accordingly, those models are typically developed by fine-tuning the pre-trained model, which utilizes the strength of pre-trained networks to learn task-agnostic language representations on specific data. However, the limitation of this approach is the over-fitting of the final task and forgetting the useful capabilities from the pre-training phase (Kulhánek et al., 2021). Therefore, recent studies try integrating multi-task learning into fine-tuning end-to-end TOD models by adding related auxiliary tasks. For instance, Lee (2021) defines span selection as an auxiliary task for enhancing the encoder of the T5 backbone model. Kulhánek et al. (2021) demonstrate that the response selection tasks are helpful on top of GPT-2 (Radford et al., 2019). Sequentially, Cholakov and Kolev (2022) combine the two above models by exploiting response selection tasks based on the T5 (Raffel et al., 2020) as the backbone model. Furthermore, PPTOD (Su et al., 2022) is introduced as the first pre-trained multiple tasks with task-specific prompts and alleviates the error accumulation in a plug-and-play fashion.

In this paper, inspired by the recent studies in this research field, we try to exploit deeply the capability of pre-trained TOD models for robust fine-tuning of end-to-end models with multi-task learning. The basic idea is to overcome the gap between the training and inference process for fine-tuning end-to-end TOD models. Accordingly, based on our observation, for the fine-tuning end-to-end model with the

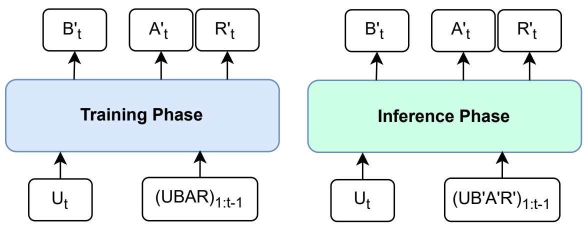

Figure 1: The gap between the training and inference phase for fine-tuning the end-to-end TOD model. BAR and B'A'R' denote the belief state, action, and response of the ground truth and generate results, respectively.

[†] Equal contribution
[*] Corresponding author

LREC-COLING 2024, pages 15493–15499
20-25 May, 2024.

sequence of tasks such as DST and NLG tasks, the training process uses ground truth to calculate the cross-entropy loss, however, the inference phase uses generated responses, which is illustrated in the Figure 1. Specifically, we utilize the well-known strong baseline models in this research field such as T5 (Raffel et al., 2020), MT-TOD (Lee, 2021), and PPTOD (Su et al., 2022) to generate the synthetic data. The results are then utilized to improve the auxiliary tasks for a robust fine-tuning process. According to the experiment reports, our proposed method outperforms the strong baseline model on two benchmark TOD datasets, which are MultiWOZ 2.1 (Eric et al., 2020) and MultiWOZ 2.2 (Zang et al., 2020), and achieves new state-of-the-art results on MultiWOZ 2.1 in the end-to-end paradigm. We release the source code for further exploitation[1].

2. Methodology

2.1. End-to-end TOD model

In contrast with pipeline-based TOD systems, end-to-end TOD directly optimizes and trains with a single model by mapping the input to the output. A general approach for building end-to-end systems is to fine-tune pre-trained language models, for instance, GPT-2 (Hosseini-Asl et al., 2020; Yang et al., 2021) or T5 (Lee, 2021; Bang et al., 2023) to task-specific data. In this study, we use T5, the encoder-decoder architecture as the backbone for our proposed model, which is illustrated in Figure 2. Specifically, for the current dialogue turn t, the

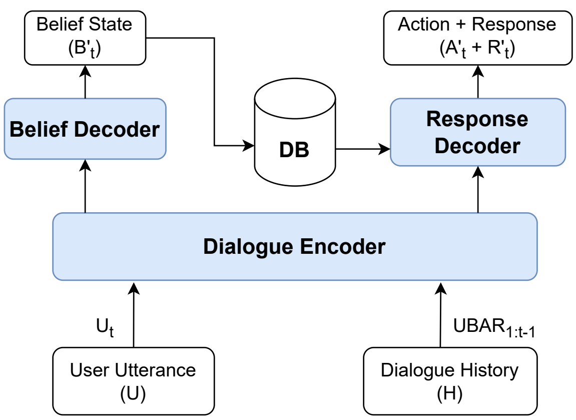

Figure 2: The overview architecture of end-to-end TOD model.

encoder layer encodes the utterance U_t and the dialogue history H_t. Technically, H_t obtain the previous workflow sequence of utterance U, belief state B, database DB, action A, and response R, which is formulated as follows:

$$H_t = [(U_1, B_1, DB_1, A_1, R_1), ..., (U_{t-1}, B_{t-1}, DB_{t-1}, A_{t-1}, R_{t-1})] \quad (1)$$

The belief decoder generates a belief state B_t, which includes a sequence of a domain name, slot names, and slot values. These values are then used to query a domain-specific database to determine a list of matching entities DB_t. The list of entities is then used as the input for the response decoder to generate system action A_t , which consists of (domain, action-type, slot) triples, and natural language response R_t, which generates tokens in an auto-regressive manner. This procedure is repeated until the dialogue is complete. In this regard, the joint loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{Belef} + \mathcal{L}_{Resp} \quad (2)$$

where $\mathcal{L}_{Belef}$ and $\mathcal{L}_{Resp}$ are negative log-likelihood language modeling losses for the two decoders layers, which are sequentially calculated as follows:

$$\mathcal{L}_{Belef} = -log(p(B_t|H_t, U_t)) \quad (3)$$

$$\mathcal{L}_{Resp} = -log(p(A_t, R_t|H_t, U_t, DB_t) \quad (4)$$

2.2. Augmented responses synthesis for end-to-end TOD models

Recent studies attempt to add training auxiliary tasks to improve end-to-end TOD systems' performance. For instance, MTTOD (Lee, 2021) utilized span selection as the auxiliary task for improving the encoder layer. RSTOD (Cholakov and Kolev, 2022) improves the MTTOD model by exploiting response selection. PPTOD (Su et al., 2022) is a new dialogue multi-task pre-training stage with four TOD-related tasks. TOATOD (Bang et al., 2023) enhances the performance of the DST and NLG tasks using the reinforcement learning method.

In this study, we try to exploit the auxiliary task to develop a robust fine-tuned end-to-end model. The main idea is to overcome the gap between the training and inference processes. Observationally, for the inference process, the multi-task in the end-to-end model takes the generated results of the previous turn $(B'A'R')_{1:t-1}$ to predict the current turn, which is different from the training process using ground truth $(BAR)_{1:t-1}$. In this regard, we utilize the generated responses of the pre-trained TOD model as the additional samples for a robust end-to-end TOD system. This strategy enables the fine-tuning of end-to-end TOD models to mitigate propagation errors during the inference phase, particularly when some of the generated states are incorrect. Our proposed method is illustrated in the Figure 3. Accordingly, with each utterance at turn t, two dialogue histories H and H' denote ground

[1]https://github.com/nqchieutb01/SynTOD

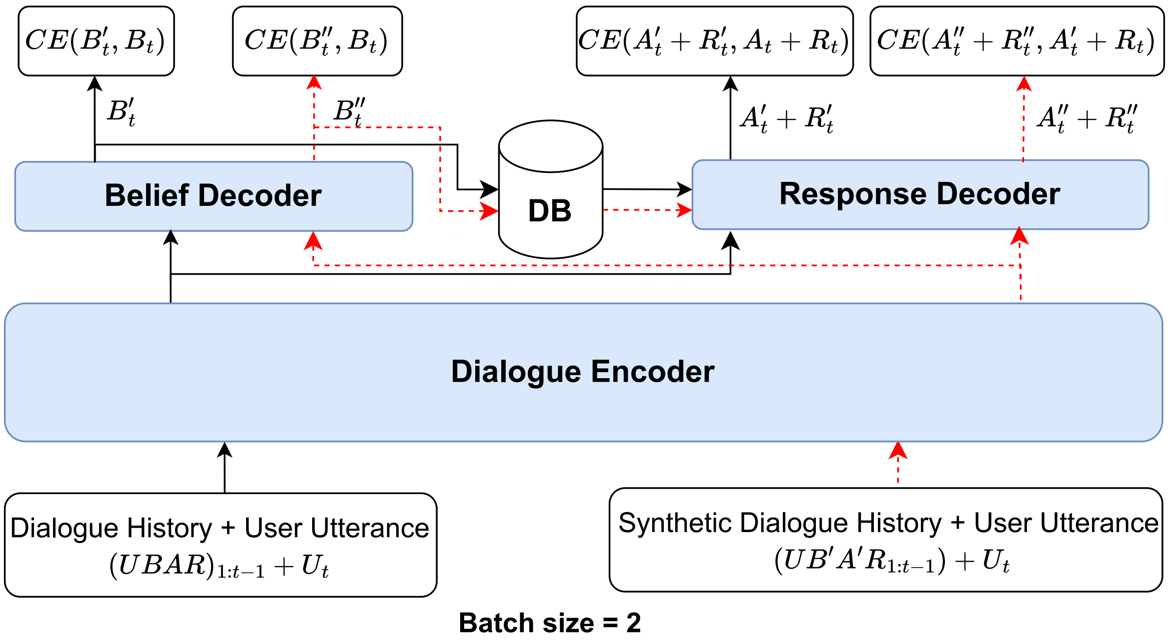

Figure 3: The overview of the proposed model (SynTOD) with batch size = 2. The black and dashed red lines refer to our model flow when the inputs are original and synthesis dialogues, respectively. CE denotes cross-entropy loss function. B' denotes the belief state output of the original dialogue, and B'' is the belief state output of the synthesis dialogue. Similar meaning for action (A) and response (R).

truth and synthesis history, respectively. Notably, the synthesis responses are generated from pre-trained TOD models. The loss function in Equation 2 can be reformulated as follows:

$$\mathcal{L} = \alpha(\mathcal{L}_{Belef} + \mathcal{L}_{Resp}) + \beta(\mathcal{L'}_{Belef} + \mathcal{L'}_{Resp}) \quad (5)$$

where α and β are hyperparameters, which are both set to 1 for the experiments. $\mathcal{L'}_{Belef}$ and $\mathcal{L'}_{Resp}$ denotes the log-likelihood language modeling loss of synthesis samples, which are sequentially calculated as follows:

$$\mathcal{L'}_{Belef} = -log(p(B_t|H'_t, U_t)) \quad (6)$$

$$\mathcal{L'}_{Resp} = -log(p(A'_t, R_t|H'_t, U_t, DB_t) \quad (7)$$

where H' represents the synthesis dialogue history:

$$H'_t = [(U_1, B'_1, DB_1, A'_1, R_1), ..., (U_{t-1}, B'_{t-1}, DB_{t-1}, A'_{t-1}, R_{t-1})] \quad (8)$$

3. Experiment

3.1. Experiment Setup

Dataset: We used MultiWOZ 2.1 and MultiWOZ 2.2 as the two benchmark datasets for our evaluation. Specifically, the first version of the MultiWOZ dataset was released in Budzianowski et al. (2018) which consists of 8438, 1000, and 1000 for training, dev, and test sets, respectively. MultiWOZ 2.1 (Eric et al., 2020) is released to correct dialogue states and add explicit system action annotation in the original version. MultiWOZ 2.2 (Zang et al., 2020) has fixes for state annotation of turns, a redefined ontology, canonical forms for slot values, and slot span annotations.

Baselines and Setup: We re-implement three strong baseline models of end-to-end TOD models (in the Figure 2) for the evaluation:

- **T5** (Raffel et al., 2020): is a well-known pre-trained language model with encoder-decoder architecture.
- **MTTOD** (Lee, 2021): uses the T5 backbone and integrates multi-task learning by adopting span prediction as an auxiliary task.
- **PPTOD** (Su et al., 2022): we replace the T5 backbone with the pre-trained weight of the PPTOD model, which is trained on large-scale dialogue datasets.

All evaluated models use a small version of the backbone T5 model, which includes 512 dimensions, 8-headed attention, and 6 layers for both the encoder and decoder. Regarding the hyperparameter, the models are trained for 10 epochs with the batch size 8, the learning rate is 5e-4, and the proportion of warm-up steps is set to 0.2. We utilize the AdamW (Loshchilov and Hutter, 2019) optimizer with the linear learning rate decaying scheme for

Model	MultiWOZ 2.1				MultiWOZ 2.2			
	Inform	Success	BLEU	Combined	Inform	Success	BLEU	Combined
T5	92,1	81,9	19,05	106,05	78,3	70,5	**18,57**	92,97
MTTOD	92,4	83.0	18,47	106,17	79.0	72,2	18,43	94,03
PPTOD	91.4	82.5	18.88	105.83	80,3	**73,8**	18,38	95,43
Ours:								
+ T5	**93.3**	83.4	**19.82**	**108.17**	84.1	73.0	17.82	96.37
+ MT-TOD	92.1	83.6	18.58	107.43	82.8	73.0	18.48	96.38
+ PPTOD	92.7	**84.3**	19.30	107.80	**85.7**	73.0	17.73	**97.08**

Table 1: Reported results on MultiWOZ 2.1 and MultiWOZ 2.2 with end-to-end evaluation. Bold texts indicate the best results.

optimization.

Evaluation Metrics In our experiments on the MultiWOZ 2.1 dataset, we employ an evaluation script as in (Lee, 2021). For the MultiWOZ 2.2 dataset, we incorporate a newly standard evaluation script, which is published in Nekvinda and Dušek (2021). More specially, both evaluation scripts contain automatic metrics as follows: **Inform** measures whether a system has provided a correct entity. **Success** measures whether it has answered all the requested information. Both are calculated on the level of dialogues. **BLEU** (Papineni et al., 2002) is calculated by comparing n-grams in human-written references and machine-generated hypotheses. It measures the fluency of output responses, where human utterances are used as the reference. Finally, the **Combined score** is computed as:

$$Combined = (Inform + Success)0.5 + BLEU \quad (9)$$

3.2. Experiment Results

Table 1 reports the results of the proposed method, which executes on three strong baselines. By incorporating synthetic dialogue responses, the input size is doubled. As reported results, the performances are consistently improving in the final scores (combined metrics) across all models, showing that augmented data is effective for boosting model performance. More specifically, in the MultiWOZ 2.1 dataset, four metrics of the proposed method outperform baseline models. In the MultiWOZ 2.2 dataset, the *Inform* score and *Success* score are better than the baseline, while the BLEU score slightly decreases. The results indicate the impact of our method for adjusting trade-off problems between BLEU scores and others (i.e., Inform) since the response from our model is robust by learning the auxiliary task (i.e., cross-entropy loss) between generate response ($A'' + R''$) and the synthesis value of the action with ground truth response ($A' + R$), as shown in Equation 7. In this regard, the generated response can be learned robustly with the noise values in previous processes (e.g., R”). Generally, our method achieves the best results on the MultiWOZ 2.1 dataset using T5 as the backbone and MultiWOZ 2.2 using PPTOD, respectively. We assume that with the cleaned version of the MultiWOZ 2.1 dataset, MultiWOZ 2.2 can achieve better performance on pre-trained TOD models.

Furthermore, Table 2 reports the results of our model compared with other recent studies on MultiWOZ 2.1 datasets, which is widely used as the benchmark dataset in this research field. Specifically, we compare our models with other strong baselines including SimpleTOD (Hosseini-Asl et al., 2020), UBAR (Yang et al., 2021), MTTOD (Lee, 2021), RSTOD (Cholakov and Kolev, 2022), PPTOD (Su et al., 2022), GALAXY (He et al., 2022), and TOATOD (Bang et al., 2023). Accordingly, our

Model	Backbone(size)	Combined
SimpleTOD	DistilGPT2(82 M)	92.98
UBAR	DistilGPT2(82 M)	105.70
MTTOD*	$T5_{small}$(102.2 M)	103.99
	$T5_{base}$(360.9M)	107.50
RSTOD$*$	$T5_{small}$(105.5 M)	108.34
PPTOD†	$T5_{small}$(60 M)	101.52
	$T5_{base}$(220 M)	102.26
GALAXY	UniLM(340 M)	105.92
GALAXY$\dagger$	UniLM(340 M)	110.76
TOATOD	$T5_{small}$(60 M)	104.54
	$T5_{base}$(220 M)	109.32
SynTOD	$T5_{small}$(60 M)	**108.17**
	$T5_{base}$(220 M)	**108.75**

Table 2: Reported results on MultiWOZ 2.1 Compared with state-of-the-art end-to-end models. The values with $*$ are from (Cholakov and Kolev, 2022). Other results are from the respective papers. $\dagger$ denotes pre-trained TOD models.

model outperforms similar trainable params models and is competitive with large models in this research field. It demonstrates that our model, by augmenting the synthesis responses as the auxiliary task can improve the performance of the end-to-end TOD model. Furthermore, different from other models in this research field, the performance between the large version (e.g., the base model) and the

Our Model	Size	Inform	Success	BLEU	Combined
+ T5	small - 60 M	**93,3**	83,4	**19,82**	**108.17**
	base - 220 M	**93,7**	**84,8**	19,5	**108.75**
+ MT-TOD	small - 102 M	92,1	**83,6**	18,58	107,43
	base - 360.9 M	92,4	83,1	19,65	107,40
+ PPTOD	small - 60 M	92,7	84,3	19,30	107,80
	base - 220 M	**93,7**	84,1	**19,54**	108,44

Table 3: Reported results on MultiWOZ 2.1 of the proposed approach with different backbone versions (i.e., small and base). Bold texts indicate the best results of each version, respectively.

small version of the proposed model is not significant. More details of this issue are reported in the Table 3. In this regard, we make a hypothesis that using generated results for the training process of TOD tasks can reduce the gap between the small and larger versions of the backbone models. Specifically, this issue is taken into account as the future work of this study.

3.3. Case Study

Figure 4 shows a case in which the proposed synthetic approach can overcome the complex query. Specifically, by using generated results for the training process, SynTOD can provide both request booking and give the reference number that other methods might fail to mention the important entity of the complex query, which leads the model to be more stable.

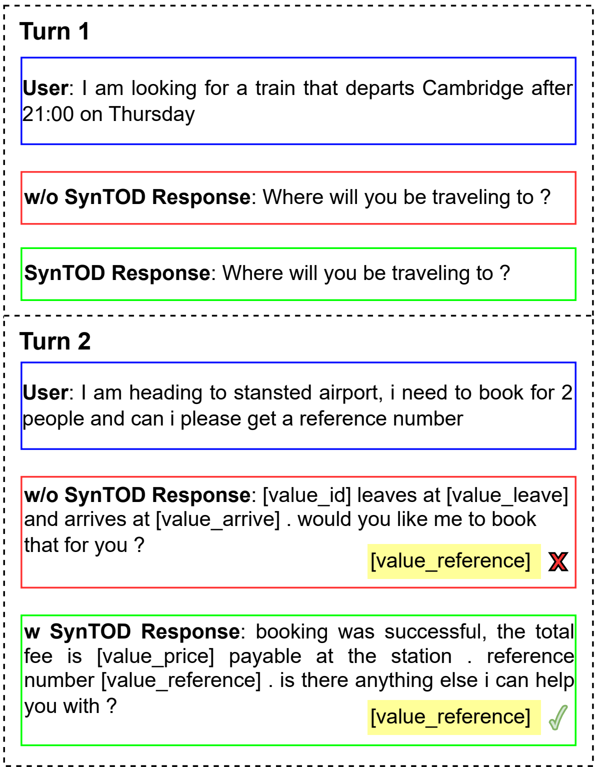

Figure 4: The example of two consecutive dialog turns in dialog session SNG0446 from the MultiWOZ2.1 dataset. The green boxes and red boxes indicate the response of with and without generated results, respectively. The task-related entities are highlighted in yellow.

4. Related Work

In the era of pre-trained language models, the end-to-end trainable framework has been an emergent research for TOD systems. A general approach for building end-to-end systems is to fine-tune pre-trained language models. SimpleTOD (Hosseini-Asl et al., 2020) and UBAR (Yang et al., 2021) are two well-known end-to-end models of this approach. Furthermore, pre-training language models for dialogue tasks, such as GALAXY (He et al., 2022) and PPTOD (Su et al., 2022) have provided promising results. MTTOD (Lee, 2021) and RSTOD (Cholakov and Kolev, 2022) apply auxiliary tasks to improve the performance of TOD systems. Recently, with the rapid growth of large pre-trained language models (billions of parameters), synthetic data generation become a promising approach for developing TOD systems (Lin et al., 2022; Bao et al., 2023). However, most previous studies used the synthetic approach to augment grounded dialogue data. To the best of our knowledge, this paper is the first study to generate synthetic results for improving the training process of TOD systems.

5. Conclusion

In this paper, we propose a new strategy to incorporate synthesis data for the training process to enable the TOD models to be more robust. Specifically, we utilize the synthesis responses from backbone models as the augmented data and training together with the fine-tuned end-to-end TOD model to make the model more robust. Experiments show the effectiveness of the proposed method in various strong baseline models. Regarding the future work of this study, We plan to extend our new version with large backbones (e.g., a large language model with billions of parameters) to generate synthetic responses for improving the performance.

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7355–7369. Association for Computational Linguistics.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10866–10882. Association for Computational Linguistics.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor.*, 19(2):25–35.

Radostin Cholakov and Todor Kolev. 2022. Efficient task-oriented dialogue systems with response selection as an auxiliary task. In *5th International Conference on Natural Language and Speech Processing, ICNLSP 2022, Trento, Italy, December 16-17, 2022*, pages 12–18. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 422–428. European Language Resources Association.

Tingchen Fu, Sheng Gao, Xueliang Zhao, Ji-Rong Wen, and Rui Yan. 2022. Learning towards conversational AI: A survey. *AI Open*, 3:14–28.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. GALAXY: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10749–10757. AAAI Press.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303. Association for Computational Linguistics.

Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, and Dilek Hakkani-Tur. 2022. Knowledge-grounded conversational data augmentation with generative conversational networks. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*, pages 26–38. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Pro-*

ceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4661–4676. Association for Computational Linguistics.

Shiquan Yang, Xinting Huang, Jey Han Lau, and Sarah M. Erfani. 2022. Robust task-oriented dialogue generation with contrastive pre-training and adversarial filtering. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1220–1234. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14230–14238. AAAI Press.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. *CoRR*, abs/2007.12720.