Improving End-to-End Task-Oriented Dialogue System with A Simple Auxiliary Task

Yohan Lee
Electronics and Telecommunications Research Institute, Daejeon, South Korea
carep@etri.re.kr

Abstract

The paradigm of leveraging large pre-trained language models has made significant progress on benchmarks on task-oriented dialogue (TOD) systems. In this paper, we combine this paradigm with multi-task learning framework for end-to-end TOD modeling by adopting span prediction as an auxiliary task. In end-to-end setting, our model achieves new state-of-the-art results with combined scores of 108.3 and 107.5 on MultiWOZ 2.0 and MultiWOZ 2.1, respectively. Furthermore, we demonstrate that multi-task learning improves not only the performance of model but its generalization capability through domain adaptation experiments in the few-shot setting. The code is available at github.com/bepoetree/MTTOD.

1 Introduction

Traditional task-oriented dialogue (TOD) systems are built on a modular pipeline architecture and their workflow is as follows: the natural language understanding (NLU) module identifies user intents and extracts task-specific slot values, and the dialogue state tracking (DST) module tracks the belief state (i.e., user goal) with considering dialogue history. By using the belief state as a database (DB) query, the system can obtain DB state, such as the number of matching entities and whether the booking is available. Based on the information, the dialogue policy (POL) module determines the next system action and then the natural language generation (NLG) module generates an appropriate natural language response according to the system action.

With the advances in neural approach, recent works on TOD system handle individual modules in a unified way. In particular, the approach to leveraging the large pre-trained language models for end-to-end dialogue modeling has shown very promising results (Ham et al., 2020; Lin et al., 2020; Lee et al., 2020; Yang et al., 2021). Such models are typically developed by fine-tuning the large pre-trained model, which learned task-agnostic language representations, with only the end-to-end dialogue modeling objective. Another approach to leveraging knowledge transfer is multi-task learning, which aims to learn universal representations (knowledge) between related tasks (Ruder, 2017). It has been shown that multi-task learning not only improves the performance of model, but also mitigates overfitting problem (Liu et al., 2015). Furthermore, Liu et al., (2019) demonstrate that the both approaches are complementary and combing them improves the performance of NLU.

In this sense, we introduce multi-task learning into fine-tuning an end-to-end TOD model, initialized with pre-trained language model. We use T5 (Raffel et al., 2020) as a backbone and adopt span prediction as an auxiliary task to boost the performance of NLU. Our model achieves new state-of-the-art results on both MultiWOZ 2.0 and MultiWOZ 2.1 in end-to-end setting. We also investigate the advantages of multi-task learning in end-to-end TOD modeling by conducting domain adaptation experiments with the few-shot setting.

2 Method

2.1 Task-Oriented Dialogue Model

The proposed model is built on a sequence-to-sequence architecture as illustrated in Figure 1. At each dialogue turn t, the encoder takes the user utterance U_t and dialogue history H_t. Based on the encoded dialogue, the belief decoder generates a belief state B_t, which consists of *(domain, slot,*

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1296–1303
November 7–11, 2021.

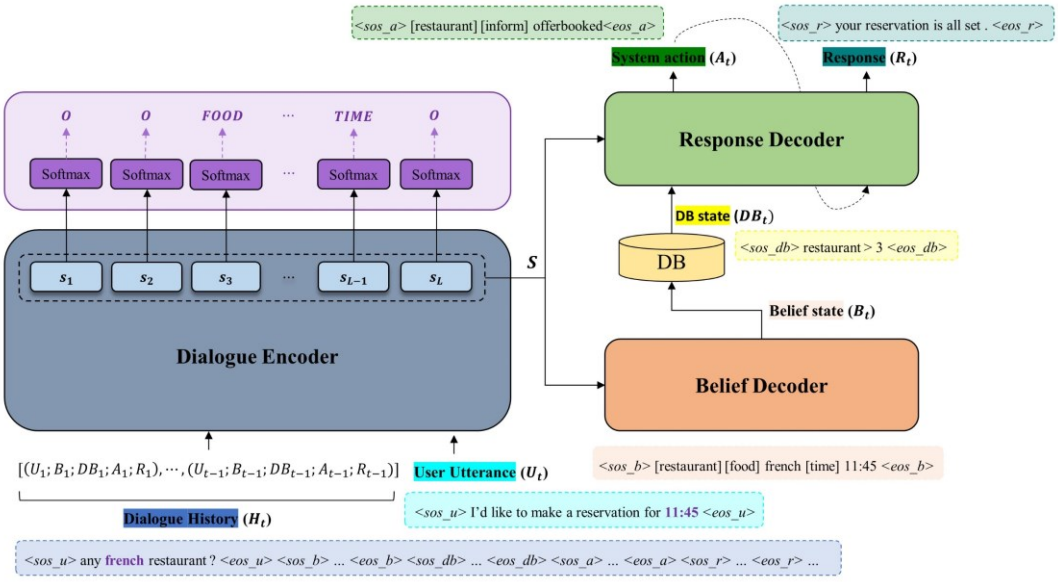

Figure 1: The overview of our end-to-end task-oriented dialogue model. The highlighted texts describe input and output example of each component. All sequences are surrounded by special tokens, such as <*sos_**> and <*eos_**> to indicate the sequence boundaries.

value) triples. The generated belief state is used to query a domain-specific database and the DB state DB_t is determined by the number of matching entities. Finally, conditioned on the encoded dialogue and the DB state, the response decoder first generates system action A_t, which consists of *(domain, action-type, slot)* triples, and natural language response R_t. Note that the natural language responses are also conditioned on the generated system action because the decoder generates tokens in an auto-regressive manner. Inspired by the state-of-the-art work (Yang et al., 2021), we treat the overall workflow of TOD as dialogue history. In other words, the current workflow sequence $(U_t; B_t; DB_t; A_t; R_t)$ is appended to the next dialogue history H_{t+1}. This procedure is repeated until the dialogue ends. The loss functions are defined as,

$$\mathcal{L}_{belief} = -\log p(B_t|H_t, U_t), \quad (1)$$

$$\mathcal{L}_{resp} = -\log p(A_t, R_t|H_t, U_t, DB_t), \quad (2)$$

for the belief and system action/response generation, respectively.

2.2 Auxiliary Task

Some recent approaches to DST use span-based method to address the out-of-vocabulary problem (Gao et al., 2020; Zhang et al., 2020; Heck et al., 2020). For each *(domain, slot)* pair, span-based DST extracts its value through span matching with start and end positions in utterances.

We adopt span prediction as an auxiliary task on the grounds of that the labels can be easily obtained by matching the ontology with dialogue context, and the task can improve the performance of NLU (Joshi et al., 2020). Different from the positional span-based DST works, we formulate span prediction as the slot tagging task, as shown in the purple box in Figure 1. Note that the domain information is excluded here because the meaning of slots is shared across all domains. The probability distribution over all possible slots for i_{th} input token is computed as,

$$p_i = softmax(W \cdot s_i + b), \quad (3)$$

where s_i is the i_{th} encoder hidden state, and W and b are trainable weights and bias, respectively. We consider only the extractive informable slots[1] defined in Gao et al., (2020) that categorize slots based on exact match rate of slot values in conversation. We use the cross-entropy loss function for the auxiliary task, $\mathcal{L}_{aux}$. Our model is trained to jointly minimize the weighted sum of the loss functions,

$$\mathcal{L} = \mathcal{L}_{belief} + \alpha\mathcal{L}_{resp} + \beta\mathcal{L}_{aux}. \quad (4)$$

In experiments, we set α and β to 1.0 and 0.5, respectively.

[1] name, leave, arrive, destination, departure, food, and type.

Model	MultiWOZ 2.0				MultiWOZ 2.1			
	Inform	Success	BLEU	Combined	Inform	Success	BLEU	Combined
SimpleTOD	84.4	70.1	15.0	92.3	-	-	-	-
SOLOIST	85.5	72.9	16.5	95.7	-	-	-	-
MinTL-BART	84.9	74.9	17.9	97.8	-	-	-	-
UBAR	**95.4**	80.7	17.0	105.1	**95.7**	81.8	16.5	105.7
MTTOD (ours)	91.0	**82.6**	**21.6**	**108.3**	91.0	**82.1**	**21.0**	**107.5**
−MTL (ablation)	90.4	81.9	21.3	107.4	89.1	80.7	20.9	105.8

Table 1: End-to-end evaluation on MultiWOZ 2.0 and MultiWOZ 2.1. In this evaluation, the generated belief state and system action are used. The additional results in different settings where the ground-truth belief state and system action are used are reported in Appendix C.

3 Experiments

3.1 MultiWOZ Dataset

Dataset. MultiWOZ (Budzianowski et al., 2018) is a large-scale TOD dataset collected via Wizard-of-Oz setup. The statistics of the dataset are presented in Appendix A. We evaluate our proposed model on both MultiWOZ2.0 and MultiWOZ 2.1 (Eric et al., 2020) which is cleaned version of MultiWOZ 2.0.

Pre-processing. The system response includes slot values that depend on the particular conversation, such as address. To reduce diversity of the surface form, the specific slot values are replaced with placeholders (Zhang et al., 2020). For example, the addresses are expressed by <*value_address*> in system response.

Evaluation Metrics. We follow the automatic evaluation metrics of Budzianowski et al., (2018): *Inform* measures whether an entity provided by system is correct, *Success* measures whether information has been provided for all user requests, and *BLEU* (Papineni et al., 2002) measures the fluency of the responses. We also report the combined score, which is computed as $Combined = (Inform + Success) \times 0.5 + BLEU$ (Mehri et al., 2019). To evaluate the DST, we use the joint goal accuracy measuring whether predicted belief state exactly matches ground-truth belief state.

3.2 Experimental Results

We developed our model using *T5-base* (220M) that consists of 12 layers of transformer blocks for the encoder and decoder, implemented in huggingface Transformers (Wolf et al., 2019). To generate the belief state and system response, we use the simple greedy decoding algorithm. The training details are described in Appendix B.

End-to-End Modeling. Table 1 compares our model (MTTOD) to the state-of-the art models leveraging large pre-trained language models in end-to-end setting. For end-to-end TOD modeling, SimpleTOD (Hosseini-Asl et al., 2020), SOLOIST (Peng et al., 2020), and UBAR (Yang et al., 2021) use GPT2 (Radford et al., 2018) and MinTL-BART (Lin et al., 2020) uses BART (Lewis et al., 2020). Our model achieves the best combined score with significantly outperforming other models in terms of *Success* and *BLEU* on both MultiWOZ 2.0 and MultiWOZ 2.1. For an ablation study, we also report the performance of the model trained without multi-task learning. MTTOD shows better performance on all the metrics, which indicates the usefulness of our auxiliary task and multi-task learning.

Dialogue State Tacking. Table 2 compares the model trained with and without multi-task learning in DST. The results with slight gains have consistency with end-to-end evaluation results, where the *Success* gains are greater than *Inform*.

Model	MultiWOZ Joint Acc.	
	2.0	2.1
MTTOD	**53.56**	**53.44**
MTTOD w/o MTL	53.17	53.25

Table 2: Dialog state tracking evaluation on MultiWOZ 2.0 and MultiWOZ 2.1.

Few-shot Domain Adaptation. In practice, it is hard to collect the massive dialogue data for each domain. Therefore, a dialogue system is required that have domain scalability with a few training examples. Following setup in Yang et al., (2021), we conduct domain adaptation experiments in the few-shot learning setting to test whether the multi-task learning improves the generalization capability of the model. We exclude attraction domain that only has 12 test dialogue sessions resulting the large variation of results in this setup.

Model	Train		Taxi		Restaurant		Hotel	
	Train	ΔOthers	Taxi	ΔOthers	Rest.	ΔOthers	Hotel	ΔOthers
MTTOD	**100.8**	-4.5	**74.3**	-3.4	**88.1**	**-12.0**	**82.1**	**-7.7**
MTTOD w/o MTL	96.8	**-2.9**	73.8	**-0.9**	79.6	-25.6	61.8	-11.2

Table 3: The combined scores of domain adaptation experiments in the few-shot learning setting. ΔOthers measures the difference of combined scores for the rest of domains (except the target domain) between the model fine-tuned on target domain and the model trained on the others. It indicates how much the model lost previous knowledge.

After the model is trained on 3 domains excluding a target domain, the trained model is fine-tuned with 100 dialogue examples which are randomly sampled from the target domain.

As shown in Table 3, the model trained with the multi-task learning achieves better performance on all target domains. This indicates that multi-task learning has major positive effects on the knowledge transfer in the low resource environment. It is also worth noting that the models have the different gaps of performance degradation between train/taxi and restaurant/hotel domains. The model trained with multi-task learning has smaller performance degradation in restaurant/hotel domains. The train and taxi domains share the same informable slots and many slot values such as arrival time and departure time. On the other hand, the restaurant and hotel domains have domain-specific slots such as food and stars. This property makes more difficult to transfer knowledge between domains and causes the catastrophic forgetting problem. Our empirical results show that the multi-task learning is helpful to alleviate this problem and improves the generalization capability of the model.

4 Related Work

Lei et al., (2018) first propose a sequence-to-sequence architecture for end-to-end TOD modeling with a belief sequence, named belief spans. Then, Zhang et al., (2020) extend the model in multi-domain scenarios with considering appropriate multiple responses. Recent approach employs transfer learning framework based on the large pre-trained language models such as GPT-2 (Radford et al., 2018), and T5 (Raffel et al., 2020) to generate the belief spans and responses. This approach has made significant progress on benchmarks for TOD system (Ham et al., 2020; Hosseini-Asl et al., 2020; Peng et al., 2020; Lin et al., 2020; Yang et al., 2021). Another approach to leveraging knowledge transfer is the multi-task learning. It has been shown that combining multi-task learning and transfer learning from pre-trained language model improves NLU tasks (Liu et al., 2019). In TOD systems, multi-task learning has been leveraged for DST (Rastogi et al., 2018; Quan and Xiong, 2020). Similar to our work, they adopt the language understanding as auxiliary task, but there is large difference in that we design the auxiliary task for end-to-end dialogue modeling in multi-domain scenarios.

5 Conclusion

In this work, we explored the approach to fine-tuning pre-trained model with multi-task learning for end-to-end TOD modeling. Our model establishes new state-of-the-art results on both MultiWOZ 2.0 and MultiWOZ 2.1 in end-to-end setting. We also demonstrate the effectiveness of multi-task learning in domain adaptation experiments with a few training examples. In future works, we plan to investigate various auxiliary tasks to enhance end-to-end TOD modeling.

Acknowledgement

We would like to thank Jonghun Shin, Ohwoog Kwon, and YoungKil Kim for helpful discussions and comments. We also thank the editors and anonymous reviewers for useful feedback.

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

References

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for Joint Language Understanding and Dialogue State Tracking. In

Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pages 376-384, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. 2018. *https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper.pdf*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. SOLOIST: Few-shot Task-Oriented Dialogue with A Single Pre-trained Auto-regressive Model. *arXiv preprint arXiv:2005.05298*.

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. *Journal of Machine Learning Research*, 21: 1-67.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583-592.

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*.

Hwaran Lee, Seokhawn Jo, Hyungjun Kim, Sangkeun Kim, and TaeYoon Kim. 2020. SUMBT+LARL: End-to-end Neural Task-oriented Dialog System with Reinforcement Learning. *arXiv preprint arXiv:2009.10447*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialogue State Tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2021. Modelling Hierarchical Structure between Dialogue Policy and Natural Language Generator with Option Framework for Task-oriented Dialogue System. In *Proceedings of Ninth International Conference on Learning Representations*.

Jun Quan and Deyi Xiong. 2020. Modeling Long Context for Task-Oriented Dialogue State Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7119-7124.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8: 64-77.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialogue State Tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 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*, pages 422-428.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871-7880.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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*, pages 5016-5026.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098*.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured Fusion Networks for Dialogue. In

Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 165-177.

Shuyang Gao, Sanchit Agarwal, Tagyoung Chung, Di Jin, and Dilek Hakkani-Tur. 2020. From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI,* pages 79-89.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan and William Yang Wang. 2019. Semantically Conditioned Dialogue Response Generation via Hierarchical Disentangled Self-Attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Wenqiang Lei, Xisen Jin, Zhaochun Ren, Xiangnan He, Min-Yen Kan, and Dawei Yin. 2018. Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv preprint arXiv:1901.11504*.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-Oriented Dialogue Systems that Consider Multiple Appropriate Responses under the Same Context. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialogue System with GPT-2. *arXiv preprint arXiv:2012.03539*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3391–3405.

A Data Statistics

Domain	# of dialogues		
	Train	Dev	Test
Police	245	0	0
Hospital	287	0	0
Attraction	127	11	12
Taxi	326	57	52
Train	282	30	33
Hotel	513	56	67
Restaurant	1,199	50	62
Train + Attraction	883	148	163
Hotel + Attraction	437	55	50
Restaurant + Attraction	396	78	70
Restaurant + Train	875	157	155
Restaurant + Hotel	462	59	49
Hotel + Train	1,077	149	144
Restaurant + Hotel + Taxi	454	41	42
Restaurant + Attraction + Taxi	431	53	59
Hotel +Attraction + Taxi	444	56	42
Total	8,438	1,000	1,000

Table 4: Statistics of train/dev/testset in MultiWOZ.

B Training Details

We train our model for 10 epochs (it takes about 10 hours on a single NVIDIA Quadro RTX 8000). The initial learning rates for end-to-end modeling and dialogue state tracking are 5e-4 and 1e-4, respectively. For all experiments, the batch size is set to 8 and the proportion of warmup steps is set to 0.1. We adopt an optimizer as AdamW (Loshchilov and hutter, 2019) with the linear learning rate decaying scheme. After the training is done, we select best checkpoint model based on performance on the development set.

C Additional Results

Table 5 compares our model (MTTOD) to action/response generation models including HDSA (Chen et al., 2019), and HDNO (Wang et al., 2021) as well as end-to-end models including MinTL (Lin et al., 2020), and UBAR (Yang et al., 2021) on MultiWOZ 2.0. Table 6 compares our model to UBAR on MultiWOZ 2.1. We also report the performance of the model trained without multi-task learning for ablation study.

Model	Belief State	System Action	Inform	Success	BLEU	Combined
HDSA	oracle	oracle	87.9	78.0	30.4	113.4
UBAR	oracle	oracle	**96.9**	**92.2**	28.6	123.2
MTTOD (ours)	oracle	oracle	93.6	89.9	**32.7**	**124.5**
−MTL (ablation)	oracle	oracle	93.3	89.6	32.6	124.0
HDSA	oracle	generated	82.9	68.9	**23.6**	99.5
HDNO	oracle	generated	**96.4**	**84.7**	18.9	**109.4**
UBAR	oracle	generated	94.0	83.6	17.2	106.0
MTTOD (ours)	oracle	generated	90.6	82.4	21.7	108.2
−MTL (ablation)	oracle	generated	91.6	82.6	21.4	108.5
MinTL-BART	generated	generated	84.9	74.9	17.9	97.8
UBAR	generated	generated	**95.4**	80.7	17.0	105.1
MTTOD (ours)	generated	generated	91.0	**82.6**	**21.6**	**108.3**
−MTL (ablation)	generated	generated	90.4	81.9	21.3	107.4

Table 5: Results of response generation on MultiWOZ 2.0

Model	Belief State	System Action	Inform	Success	BLEU	Combined
UBAR	oracle	oracle	**95.4**	**91.4**	28.8	122.2
MTTOD (ours)	oracle	oracle	93.8	90.0	**32.3**	124.1
−MTL (ablation)	oracle	oracle	93.9	90.3	32.1	**124.2**
UBAR	oracle	generated	**92.7**	81.0	16.7	103.6
MTTOD (ours)	oracle	generated	91.4	**82.7**	**21.2**	**108.2**
−MTL (ablation)	oracle	generated	91.1	82.5	21.0	107.8
UBAR	generated	generated	**95.7**	81.8	16.5	105.7
MTTOD (ours)	generated	generated	91.0	**82.1**	**21.0**	**107.5**
−MTL (ablation)	generated	generated	89.1	80.7	20.9	105.8

Table 6: Results of response generation on MultiWOZ 2.1.