User Satisfaction Modeling with Domain Adaptation in Task-oriented Dialogue Systems

Yan Pan
BMW Group, Germany
Technical University of Munich, Germany
Frank_panyan@outlook.com

Mingyang Ma
BMW Group, Germany
mingyang.ma@bmw.de

Bernhard Pflugfelder
BMW Group, Germany
bernhard.pflugfelder@bmw.de

Georg Groh
Technical University of Munich, Germany
grohg@in.tum.de

Abstract

User Satisfaction Estimation (USE) is crucial in helping measure the quality of a task-oriented dialogue system. However, the complex nature of implicit responses poses challenges in detecting user satisfaction, and most datasets are limited in size or not available to the public due to user privacy policies. Unlike task-oriented dialogue, large-scale annotated chitchat with emotion labels is publicly available. Therefore, we present a novel user satisfaction model with domain adaptation (USMDA) to utilize this chitchat. We adopt a dialogue Transformer encoder to capture contextual features from the dialogue. And we reduce domain discrepancy to learn dialogue-related invariant features. Moreover, USMDA jointly learns satisfaction signals in the chitchat context with user satisfaction estimation, and user actions in task-oriented dialogue with dialogue action recognition. Experimental results on two benchmarks show that our proposed framework for the USE task outperforms existing unsupervised domain adaptation methods. To the best of our knowledge, this is the first work to study user satisfaction estimation with unsupervised domain adaptation from chitchat to task-oriented dialogue.

1 Introduction

The developed task-oriented dialogue system has achieved great success for various business situations, such as virtual assistants and information-seeking systems with domain knowledge (Deriu et al., 2021). However, a dialogue chatbot with limited model capability sometimes fails to understand queries correctly and even annoys users with the wrong response. User Satisfaction Estimation (USE) is able to detect user satisfaction and enable adjustment of the strategy of the system. Liu et al. (2021) implemented a smooth handoff from the machine to a human agent when USE estimates a negative emotion from a user. When USE detects good user feedback in the deployment environment, chatbots can utilize this information to learn and improve continuously (Hancock et al., 2019).

In recent years, the USE in dialogue systems is always considered in the classification task. Previous works (Sun et al., 2021; Deng et al., 2022) show that data-driven pre-trained models can learn good exchange-level representations from task-oriented corpora and predict correct user satisfaction. Unfortunately, most user satisfaction datasets are very limited in size (Saha et al., 2020; Shi and Yu, 2018) or not publicly available due to user privacy policies (Wang et al., 2020). Moreover, it is time-consuming and expensive to conduct human evaluation experiments or crowd-sourcing for user satisfaction in a real-world task-oriented application.

Compared to the task-oriented dataset, the chitchat corpora from social media is easy-to-get but without explicit chatting targets. The underlying difference in linguistic patterns between the chitchat and task-oriented dialogue makes it difficult to utilize the chitchat dataset in the USE task directly. Therefore, unsupervised domain adaptation from chitchat to task-oriented dialogue is valuable and challenging in user satisfaction tasks.

As shown in Figure 1, we collect two dialogue sessions from human-human chitchat and human-machine task-oriented dialogue. In the chitchat, people talk around one topic and explicitly express their intents with emotions. In task-oriented dialogue, the user and system have explicit actions where the user wants to achieve his goal, and the system uses the background knowledge following the presetting actions. But users tend to implicitly show their emotions and are comfortable with the fulfillment of their goals.

To tackle the domain difference, we propose a novel USMDA framework and implement USE

Proceedings of the SIGdial 2022 Conference, pages 630–636
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

Chitchat	Emotion	Satisfaction
Listen to me! When my time comes, I wanna be buried at sea.	Neutral	Neutral
You what?	Scared	Dissatisfied
I wanna be buried at sea, it looks like fun.	Joyful	Satisfied

Task-oriented dialogue	Action	Satisfaction
Hello! Can you help me find a hotel room?	Inform Intent	Neutral
May I suggest 1 Hotel Brooklyn Bridge? It is a well reviewed, 4 star hotel.	Offer	-
What other options do I have?	Request Alternatives	Dissatisfied

Figure 1: Two example dialogue sessions in chitchat (Zahiri and Choi, 2018) and task-oriented dialogue (Rastogi et al., 2020).

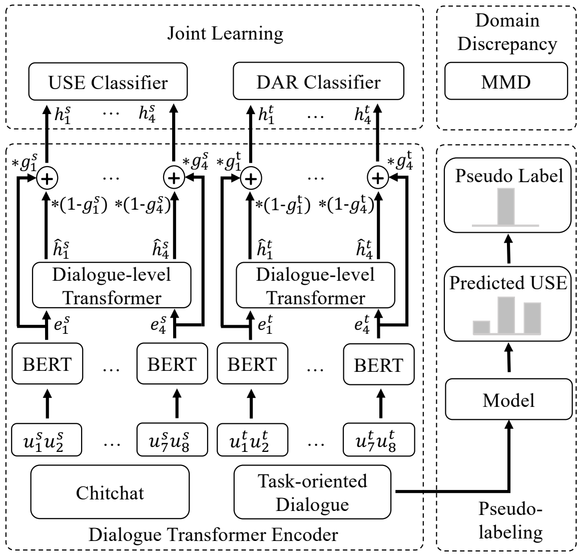

Figure 2: Overall framework architecture. The superscripts s and t denote the source chitchat data and target task-oriented dialogue data.

with unsupervised domain adaptation from chitchat to task-oriented dialogue. On the one hand, the model reduces the domain discrepancy of turn representations between chitchat and task-oriented dialogue datasets. On the other hand, the model learns satisfaction signals in context features from chitchat, and learns user actions in the task-oriented system with an additional Dialogue Action Recognition (DAR) task. Moreover, the framework utilizes the pseudo-labeling approach (Lee, 2013) to label the most confident predictions and build a stronger USE model.

To the best of our knowledge, our paper is the first attempt to explore the USE with domain adaptation from chitchat to task-oriented dialogue. In this work, we make the following contributions:

- We propose the USMDA framework to perform user satisfaction estimation with unsupervised domain adaptation from chitchat to task-oriented dialogue.
- The result shows that user actions and invariant dialogue-related features improve the performance of the USE model within an unsupervised domain adaptation setting.
- The results on two datasets demonstrate that the proposed framework in the USE task achieves better results than other domain adaptation approaches.

2 Problem Definition

We formulate the task of user satisfaction estimation with domain adaptation from chitchat to task-oriented dialogue. Given a set of chitchat and task-oriented dialogue sessions, each session contains N utterances $\{u_1, u_2, ..., u_N\}$. We split the N utterances into $\frac{N}{2}$ exchange turns $x_i = (u_{2i-1}, u_{2i})$. Each exchange turn is a communication either between multiple users or between user and system. Each exchange turn in chitchat is annotated with a satisfaction label y_i^s and each exchange turn in a task-oriented dialogue has a user action a_i^t. Our goal is to train a USE model using labeled chitchat data S and unlabeled task-oriented dialogue data T to predict the correct satisfaction label y_i^t on T.

3 Framework

This section introduces how to train a user satisfaction model with unsupervised domain adaptation. Figure 2 shows the overall architecture of our proposed framework USMDA with four different parts, including (1) dialogue Transformer encoder to capture a representation of each exchange-turn in the dialogue, (2) joint learning for USE with DAR, (3) reducing domain discrepancy between different distributed datasets, (4) predicting pseudo labels in the task-oriented dialogue, and retraining the model with the top-k pseudo labels.

3.1 Dialogue Transformer encoder

Chitchat and task-oriented dialogue samples are mixed in one batch X, which is fed into the shared backbone BERT (Devlin et al., 2019) to extract the exchange-level representation e_i of each exchange turn x_i. Each e_i represents the information from an exchange turn:

$$e_i = BERT([CLS]u_{2i-1}[SEP]u_{2i}[SEP]) \quad (1)$$

The shared dialogue-level transformer encoder is built upon the exchange-level representations $\{e_1, e_2, ..., e_M\}$ of M exchange turns within a dialogue window. We adopt a Transformer encoder with a gated attention mechanism to capture the context information in the conversation:

$$\hat{h}_i = \text{Dialogue-Transformer}(e_i) \quad (2)$$

$$g_i = \text{Sigmoid}(W[e_i; \hat{h}_i]) \quad (3)$$

$$h_i = g_i * e_i + (1 - g_i) * \hat{h}_i \quad (4)$$

where $\hat{h}_i$ is the dialogue-level representation, g_i is the learned gated attention weight to combine two different level representations, W is a trainable matrix and h_i is the final representation of x_i.

3.2 Joint learning

The model jointly trains with USE and DAR to learn the specific user actions in the task-oriented dialogue. The USE classifier calculates the loss between the labeled satisfaction classes and predictions in the chitchat dataset. The DAR classifier learns to predict correct user actions in the task-oriented dataset. The joint learning loss is the sum of losses from USE and DAR classifiers:

$$\mathcal{L}_{Joint} = \mathcal{L}_{USE} + \alpha \mathcal{L}_{DAR} \quad (5)$$

where α denotes the hyperparameter to balance USE and DAR tasks.

3.3 Domain discrepancy

The framework uses maximum mean discrepancy (MMD) (Gretton et al., 2012; Long et al., 2015) to measure the distance between chitchat and task-oriented dialogue dataset distributions. MMD computes the distance between two exchange-level representations with Gaussian kernel, i.e., $k(e_i^s, e_j^t) = \exp(-\|e_i^s - e_j^t\|^2)$. Finally, we combine the joint learning loss and MMD as the overall loss:

$$\mathcal{L} = \mathcal{L}_{Joint} + \beta\left(\frac{4}{|X|^2} \sum_{i=1}^{\frac{|X|}{2}} \sum_{j=1}^{\frac{|X|}{2}} k(e_i^s, e_j^t)\right) \quad (6)$$

where e_i^s and e_j^t are two exchange-level representations from chitchat and task-oriented dialogue, β denotes the hyperparameter balancing the joint-learning loss and MMD, and $|X|$ is the size of a mixed batch X.

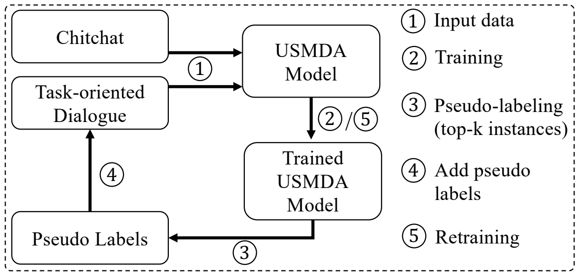

Figure 3: Retraining process with pseudo labels.

3.4 Pseudo-labeling

After joint learning and reducing domain discrepancy, the user satisfaction model makes the satisfaction prediction $\hat{y}_i^t$ on each exchange turn x_i^t from task-oriented dialogue. We measure the confidence of predictions by predicted scores. As shown in Figure 3, the top-k instances with the highest predicted scores are set as pseudo labels for retraining.

4 Experiments

4.1 Datasets and evaluation metrics

We conduct the proposed framework on the chitchat dataset EmoryNLP (Zahiri and Choi, 2018) and two task-oriented dialogue datasets: MultiWOZ 2.1 (MWOZ) (Eric et al., 2020) and Schema Guided Dialogue (SGD) (Rastogi et al., 2020). Moreover, we use the sampled 1000 dialogues from each of the MWOZ and SGD datasets, which are annotated with a five level satisfaction scale by Sun et al. (2021). The seven emotions in chitchat and five rating scores in task-oriented dialogue datasets are mapped into the coarse-grained labels "dissatisfied/neutral/satisfied" following existing work (Deng et al., 2022; Zahiri and Choi, 2018). For the DAR task, the MWOZ dataset is labeled with 21 actions by Eric et al. (2020), and the SGD dataset has 12 actions from Rastogi et al. (2020). We use the EmoryNLP dataset as a labeled source dataset and randomly choose 300 dialogues from each of the task-oriented dialogue datasets as unlabeled target datasets. The remaining 700 labeled dialogues from each task-oriented dialogue dataset are used for testing.

Following most existing work on emotion recognition in conversation, we report Macro-F1 and Micro-F1 scores for evaluating USE performance. Macro-F1 takes the average of all the per-class F1, and Micro-F1 computes the F1 of the aggregated contributions of classes.

	MWOZ		SGD	
Model	Macro	Micro	Macro	Micro
Performance without domain adaptation				
Bert (baseline)	37.98	45.51	40.66	49.15
ToD Bert	31.69	40.49	35.80	43.35
Performance with domain adaptation				
WDGRL	38.58	46.26	41.77	49.91
DANN	37.68	47.91	46.55	51.28
USMDA	**43.27**	**48.50**	**56.01**	**57.91**
Performance with supervised learning				
Upper bound	45.32	48.94	59.66	61.09

Table 1: Primary results with Micro-F1 and Macro-F1 metrics on task-oriented dialogue datasets.

4.2 Other models

We use the BERT model as our baseline model and the backbone for our proposed method for a thorough comparison. The following related models with task-oriented dialogue pretraining or different unsupervised domain adaptation methods are implemented:

- ToD Bert (Wu et al., 2020) is pretrained with masked-language modeling strategy and response selection task on nine task-oriented dialogue datasets.
- WDGRL (Shen et al., 2018) learns domain invariant representations by reducing empirical Wasserstein distance with an adversarial strategy.
- DANN (Ganin et al., 2016) uses domain adversarial training to learn the features that can not discriminate in domain adaptation. The DANN method is most widely used for unsupervised domain adaptation task in natural language processing.

5 Results and Analysis

5.1 Overall performance

Table 1 shows primary experiment results, including the following models: (1) the baseline model and ToD Bert using only the source chitchat dataset, (2) several models with domain adaptation strategies and access to the user actions from the target data, (3) the BERT-based model with supervised learning on task-oriented datasets as upper bound.

We made the following notable observations:

(1) Our unsupervised domain adaptation strategy is effective in improving the performance for USE on two task-oriented dialogue datasets. USMDA leads to a significant improvement in Macro-F1 of 5.29% on MWOZ and 15.35% on SGD, and a performance improvement in Micro-F1 of 2.99% on MWOZ and 8.76% on SGD. Our proposed framework USMDA successfully solves the domain shift problem for USE from chitchat to task-oriented dialogue. USMDA, without any satisfaction labels in task-oriented data, achieves a competitive Micro-F1 48.50% on MWOZ, which is comparable to the upper bound model with supervised learning.

(2) Our framework USMDA achieves the best performance with domain adaptation for two datasets. On average, the models with domain adaptation have better performance than the baseline model. This suggests that the domain-invariant dialogue-related features boost the performance of the user satisfaction model. Compared to other domain adaptation approaches, USMDA leads to a comparatively significant improvement. We demonstrate that our proposed framework USMDA to learning domain-invariant dialogue-related features is more effective than WDGRL and DANN.

(3) Baseline model, using only source chitchat samples, does not perform competitively. Even though ToD-BERT is pretrained with nine task-oriented dialogue datasets, it has a subpar performance without domain adaptation in the USE task. The unsatisfactory results without domain adaptation suggest that specific domain features are valuable and necessary for USE in task-oriented dialogue.

5.2 Ablation study

To understand the impacts of different individual parts in our domain adaptation strategy, we conduct an ablation study on three simplified modules of our proposed framework (see Table 2). We can observe that by removing any module, this results in worse performance. Removing joint learning leads to the most significant loss in Micro-F1 by 6.96% on SGD. This indicates that user actions throughout the dialogue reflect user satisfaction and are important dialogue-related specific features in task-oriented dialogue.

Table 2 shows that the improvement transfers well across both datasets. Learning transferable features using MMD is beneficial because dropping MMD impairs the performance by 1.17% Macro-F1 and 0.85% Micro-F1 on SGD. Moreover, removing the pseudo-labeling degrades the performance

	MWOZ		SGD	
	Macro	Micro	Macro	Micro
w/o pseudo	-5.33	-0.58	-3.88	-1.20
w/o MMD	-0.37	-0.26	-1.17	-0.85
w/o joint	-0.22	-0.50	-6.27	-6.96

Table 2: Ablation study of USMDA on pseudo-labeling, joint learning and MMD. A negative value means a performance loss by removing module.

by 3.9-5.3% Macro-F1 and 0.6-1.2% Micro-F1, indicating the benefits of the data-centric approach to the USE task.

5.3 Discussion and future work

Compared to the kernelized method MMD, the WDGRL and DANN are adversarial training strategies. Table 1 shows that WDGRL improves the model performance only slightly and that DANN does not always lead to the increased target domain performance. While traditional adversarial training strategies are sometimes unable to gain improvements with pre-trained language models, simple MMD is efficient at learning domain-invariant features. Our proposed framework achieves impressive results on the two fixed datasets. In the future, we will evaluate this framework on real-life scenarios.

6 Conclusion

We adopt joint-learning, MMD, and pseudo-labeling with domain adaptation to improve the strong USE model in task-oriented dialogue. The results show that domain adaptation with user actions is effective in the USE task. MMD has positive effects on overall performance by learning domain-invariant dialogue-related feature representations. The pseudo-labeling is important for USE with unsupervised domain adaptation. Our proposed USMDA framework has comparable results like the supervised model, encouraging future work addressing domain adaptation in the USE task.

Acknowledgements

The BMW Group supported the content of this work. We thank Davide Cadamuro and the reviewers for the invaluable feedback.

References

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *WWW '22: The Web Conference 2022*.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

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, Marseille, France. European Language Resources Association.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Jiawei Liu, Zhe Gao, Yangyang Kang, Zhuoren Jiang, Guoxiu He, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2021. Time to transfer: Predicting and evaluating machine-human chatting handoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5841–5849.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PloS one*, 15(7):e0235367.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

A Appendix

A.1 Datasets

We perform experiments on dialogue corpora, using 713 dialogues from EmoryNLP, 1000 dialogues from MWOZ, and 1000 dialogues from SGD. A dialogue session is divided into dialogue windows. The number of considered exchange-level turns in a dialogue window is four.

EmoryNLP	Emotion	Satisfaction
Monica: Hey.	Neutral	Neutral
Rachel: Hey.	Neutral	Neutral
Monica: How's the big anniversary dinner?	Neutral	Neutral
Rachel: Well, we never actually got to dinner.	Sad	Unsatisfied
Monica: Ohhh, nice.	Sad	Unsatisfied
Rachel: No, we kinda broke up instead.	Sad	Unsatisfied
Monica: What?!	Scared	Unsatisfied
Rachel: God, Monica it's on the ceiling.	Scared	Unsatisfied

Table 3: Chitchat example from EmoryNLP.

EmoryNLP: EmoryNLP is an annotated chitchat dataset with fine-grained and coarse-grained emotions from the TV show, Friends. The EmoryNLP contains seven emotions: (1) Positive: powerful, joyful, peaceful, (2) Negative: mad, sad, scared, and (3) Neutral: neutral.

MWOZ: Multi-Domain Wizard-of-Oz 2.1 is a multi-domain task-oriented dialogue dataset. It contains dialogue utterances, user goals, and system actions over seven domains.

SGD: Schema-Guided Dialogue has multi-domain task-oriented dialogues over 16 domains and provides a challenging testbed for dialogue state tracking. SGD contains multiple schemas with a dynamic set of slots for task-oriented dialogue.

A.2 Training and hyperparameters

For the backbone of USMDA, we use the implementation of the BERT-base uncased model and of the ToD-Bert model from huggingface[1]. Both models are 768-dimensional Transformer self-attention encoders with 12 layers. The dialogue-level Transformer is a standard Transformer encoder with 2 layers.

We run each experiment four times with different seeds and calculate the average performance. The adamW optimizer is implemented with a learning rate of 2E-5, and the training epoch size is 2.

A.3 Dialogue Examples

The annotated chitchat and task-oriented dialogue examples are presented in Tables 3, 4.

[1]https://huggingface.co/

SGD	Action	Satisfaction
System: Okay, there is an American Airlines outbound flight that will leave at 11:40 am and the return flight will leave at 9:30 am. The cost of the ticket will be only $163.	Offer	-
USER: When is the arrival time of my return flight and to which airport it arrives?	Request	Neutral
System: The destination airport is Dulles International Airport and the flight will arrive at 1:42 pm.	Inform	-
USER: Sounds good to me.	Select	Satisfied
System: Do you want me to book tickets for this flight?	Offer	-
USER: Yup, please book 1 ticket in economy class which should be a refundable one, since my plan might change later on.	Affirm	Neutral
System: Okay! Please can you confirm me that you wish to fly from Atlanta to Washington in American Airlines on March 10th at 11:40 am and the return journey will be on March 14th at 9:30 am and you wish to book only 1 Economy ticket, right?	Confirm	-
USER: Yup, you're right. Is it a zero stops flight?	Affirm	Satisfied

Table 4: Task-oriented dialogue example from SGD.