On Task-Level Dialogue Composition of Generative Transformer Model

Prasanna Parthasarathi [*]
McGill University / Mila

Arvind Neelakantan [†]
OpenAI

Sharan Narang
Google Brain

Abstract

Task-oriented dialogue systems help users accomplish tasks such as booking a movie ticket and ordering food via conversation. Generative models parameterized by a deep neural network are widely used for next turn response generation in such systems. It is natural for users of the system to want to accomplish multiple tasks within the same conversation, but the ability of generative models to compose multiple tasks is not well studied. In this work, we begin by studying the effect of training human-human task-oriented dialogues towards improving the ability to compose multiple tasks on Transformer generative models. To that end, we propose and explore two solutions: (1) creating synthetic multiple task dialogue data for training from human-human single task dialogue and (2) forcing the encoder representation to be invariant to single and multiple task dialogues using an auxiliary loss. The results from our experiments highlight the difficulty of even the sophisticated variant of transformer model in learning to compose multiple tasks from single task dialogues.

1 Introduction

Recent years have seen a tremendous surge in the application of deep learning methods for dialogue in general (Vinyals and Le, 2015; Rojas-Barahona et al., 2017; Budzianowski et al., 2018; Lewis et al., 2017) and task-oriented dialogue (Wen et al., 2015; Einolghozati et al., 2019; Neelakantan et al., 2019) specifically. Task-oriented dialogue systems help users accomplish tasks such as booking a movie ticket and ordering food via conversation. Generative models are a popular choice for next turn response generation in such systems (Rojas-Barahona et al., 2017; Wen et al., 2017; Eric and Manning, 2017). These models are typically learned using large amounts of dialogue data for every task (Budzianowski et al., 2018; Byrne et al., 2019). It is natural for users of the task-oriented dialogue system to want to accomplish multiple tasks within the same conversation, e.g. booking a movie ticket and ordering a taxi to the movie theater within the same conversation. The brute-force solution would require collecting dialogue data for every task combination which might be practically infeasible given the combinatorially many possibilities.

While the ability of generative dialogue models to compose multiple tasks has not yet been studied in the literature, there has been some investigation on the compositionality skills of deep neural networks. Lake and Baroni (2017) propose a suite of tasks to evaluate a method's compositionality skills and find that deep neural networks generalize to unseen compositions only in a limited way. Kottur et al. (2017) analyze whether the language emerged when multiple generative models interact with each other is compositional and conclude that compositionality arises only with strong regularization.

Motivated by the practical infeasibility of collecting data for combinatorially many task compositions, we focus on task-level compositionality of text response generation models. We begin by studying the effect of training data size of human-human multiple task dialogues on the performance of Transformer (Vaswani et al., 2017) generative models. Next, we explore two solutions to improve task-level compositionality. First, we propose a data augmentation approach (Simard et al., 2003; Schmidhuber, 2012; Krizhevsky et al., 2012; Baird, 1992; Sennrich et al., 2016) where we create synthetic multiple task dialogues for training from human-human single task dialogue; we add a portion of one dialogue as a prefix to another to

[*] This work was done when the author was an intern at Google Brain. pparth2@cs.mcgill.ca

[†] This work was done when the author was a Research Scientist at Google Brain.

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 41–47
Online, November 19, 2020. ©2020 Association for Computational Linguistics
ISBN 978-1-952148-66-8

simulate multiple task dialogues during training. As a second solution, we draw inspiration from the domain adaptation literature (Ganin and Lempitsky, 2015; Tzeng et al., 2015; Xu and Yang, 2017; Chen et al., 2016; Xu et al., 2017; Sun et al., 2018) and encourage the model to learn domain invariant representations with an auxiliary loss to learn representations that are invariant to single and multiple task dialogues.

We conduct our experiments on the Multiwoz dataset (Budzianowski et al., 2018). The dataset contains both single and multiple task dialogues for training and evaluation. In Multiwoz, the tasks in multiple task dialogues are only the combinations of tasks in single task dialogues. This allows the dataset to be an appropriate benchmark for our experiments.

To summarize, our key findings are:

- We study task-level compositionality of text response generation models and find that they are heavily reliant on multiple task conversations at train time to do well on such conversations at test time.

- We explore two novel unsupervised solutions to improve task-level compositionality: (1) creating synthetic multiple task dialogue data from human-human single task dialogue and (2) forcing the encoder representation to be invariant to single and multiple task dialogues using an auxiliary loss.

- Highlighting the difficulty of composing tasks in generative dialogues with experiments on the Multiwoz dataset, where both the methods combined result only in a 8.5% BLEU (Papineni et al., 2002) score improvement when zero-shot evaluated on multiple task dialogues.

2 Background

Let $d_1, d_2, \ldots, d_M$ be the dialogues in the training set and every dialogue $d_m = ((u_m^1, a_m^1), (u_m^2, a_m^2), \ldots, (u_m^{n_m}, a_m^{n_m}))$ $(\forall m \in \{1, 2, \ldots, M\})$ consists of n_m turns each of user and assistant. Further each user and assistant turn consists of a sequence of word tokens. The individual dialogue could be either single task or multiple task depending on the number of tasks being accomplished in the dialogue.

The response generation model is trained to generate each turn of the assistant response given the conversation history. The generative model learns a probability distribution given by $P(a^i \mid (u^1, a^1), \ldots, (u^{i-1}, a^{i-1}), u^i)$. We drop the symbol m that denotes a particular training example for simplicity. The assistant turn a^i consists of a sequence of word tokens, $a^i = (w_1^i, w_2^i, \ldots, w_{l^i}^i)$. The response generation model factorizes the joint distribution left-to-right given by,

$$P(a^i \mid x^i) = \prod_{j=1}^{l^i} P(w_j \mid x^i, w_1^i, \ldots, w_{j-1}^i)$$

where $x^i = ((u^1, a^1), \ldots, (u^{i-1}, a^{i-1}), u^i)$ refers to the conversation history till the i^{th} turn.

We use a Transformer (Vaswani et al., 2017) sequence-to-sequence model to parameterize the above distribution. Given a training set of dialogues, the parameters of the Transformer model are learned to optimize the conditional language modelling objective given by,

$$L_{LM} = \sum_{m=1}^{M} \sum_{i=1}^{n_m} \log P(a^i \mid x^i, \Theta) \qquad (1)$$

where Θ refers to the parameters of the Transformer model.

3 Data Augmentation

The first solution we explore for task compositionality generates synthetic multiple task dialogues for training from human-human single task dialogues [1]. Here, we sample two dialogues from the training set, and add a portion of one dialogue as a prefix to another. While this procedure might not create dialogues of the quality equivalent to human-human multiple task dialogue, it is an unsupervised way to create approximate multiple task dialogues that the model could theoretically benefit from.

Concretely, we randomly sample two single task dialogues d_i and d_j from the training set and create a noisy multiple task dialogue by adding a fraction of the dialogue d_j as a prefix to dialogue d_i. The fraction of dialogue taken from dialogue d_j is given by the hyperparameter $augment\_fraction$. The number of times dialogue d_i is augmented by a randomly sampled dialogue is given by the hyperparameter $augment\_fold$.

We consider two strategies for sampling the dialogue d_j. In $Random\_Augment$, the dialogue is uniformly randomly sampled from the remainder of the training set. A potential issue with the random strategy is that it might create spurious

[1] Code repository

task combinations and the model might fit to this noise. Motivated by the spurious task combination phenomenon, we consider another sampling strategy $Targeted\_Augment$ where we create synthetic multiple task dialogues only for task combinations that exist in the development set. Here, d_j is sampled from a set of dialogues whose task is compatible with the task of dialogue d_i. The Transformer model is now trained on the augmented training set using the objective function given in Equation 1. The effect of the sampling strategy and the hyperparameters on the model performance is discussed in the experiments section (Section 5).

4 Domain Invariant Transformer

We propose Domain Invariant Transformer model (Figure 1) to maintain a domain invariant representation of the encoder by training the encoder representation for an auxiliary task. Here, the auxiliary task for the network is to predict the label ,$^{i}\hat{l}$, denoting the type of task (single or multi-task) in the encoded conversation history. The model takes as input the sequence of byte pair encoded tokens that are represented at the encoder hidden state as a set of attention weights from the multi-head multiple layer attention mechanism of transformer. The conditional language model (Equation 1) is learnt by a transformer decoder on top that attends over the encoder states.

The discriminator task network is trained with average pooling of the encoder summary over the attention heads (h_j)as shown in Equation 2.

$$^{i}e^{s} = \sum_{j=1}^{k} \frac{(h_j)}{k} \tag{2}$$

The average pooled encoder summary is passed as input to a two-layer feed forward discriminator. The discriminator network has a dropout (Srivastava et al., 2014) layer in-between the two fully connected layers (f_1 and f_2) (Equation 3).

$$\hat{y}_i = f_2\left(f_1\left(^{i}e^{s}\right)\right) \tag{3}$$

The binary cross-entropy loss, L_{disc}, for the predicted label, $\hat{y}_i$, an input context *i* is computed as in Equation 4.

$$L_{disc} = -\left(y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right) \tag{4}$$

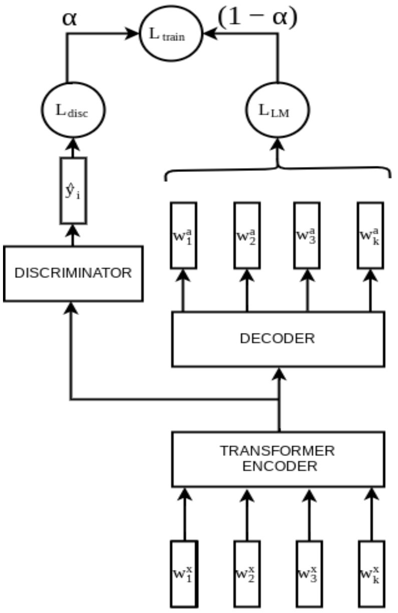

Figure 1: Domain Invariant Transformer Architecture.

The Domain Invariant Transformer model optimizes a convex combination of the two losses as shown in Equation 5.

$$L_{train} = \alpha * L_{disc} + (1 - \alpha) * L_{LM} \tag{5}$$

The language model loss makes sure that the model learns to generate the next utterance while the discriminator loss makes sure the model is aware of the nature of task. To understand the effect of the auxiliary loss we experiment with different values for α (ref Appendix).

5 Experiments

5.1 Importance of multiple task dialogues

We measure the importance of multiple task dialogue on the overall performance of transformer by training the model with varying amount of multiple task dialogues and keeping the task distribution between multiple and single domain dialogues almost similar in the experiments. We keep increasing the number of multiple task dialogues while reducing the single task dialogues to keep the total number of dialogues constant at $2,150$. The model should be able to learn to generalize to multiple tasks as the set of tasks are the same between the train and test sets with only the nature in which the task is posed by the user is different. We use the Tensor2Tensor (Vaswani et al., 2018) framework to run our experiments with (tiny) hyper-parameter setting in the framework.

Training Data		BLEU	
Single	Multiple	Multiple Only	Overall
2150	0	7.17	6.81
1836	314	7.25	6.87
1522	628	7.94	7.84
1208	942	8.68	8.68
894	1256	8.83	8.27
580	1570	9.33	8.84
266	1884	9.10	9.25

Table 1: Ablation study to understand the usefulness of Multiple task dialogues.

As shown in Table 1, the quality of the model improves significantly as number of multiple task dialogues increases. Interestingly, even though the total number of dialogues are kept fixed, the overall validation BLEU score also improves as the number of multiple task dialogues increase in the training set. The results show that the models may be better at decomposing than composing in the domain of goal oriented dialogues or the model at best can only mimic surface level token distribution (Appendix B). Though training with more multi-task dialogues can potentially improve the performance, it is not a scalable solution. We will test two of the out-of-the-shelf techniques to improve the task level compositionality in the following section.

5.2 Zero-shot Compositionality Experiments

We experiment on Transformer to evaluate the performance on handling zero-shot compositional tasks by training the baseline model only on single task dialogues, and with the proposed data augmentation techniques. The results, in Table 2, show that the *Targeted_Augment* technique increased the performance on multiple-task dialogues by 8.5% BLEU score while the scores of the model slightly dropped in the performance of all dialogues.

Data	BLEU	
	Multiple	Overall
SNG	7.17	6.81
SNG + RS	7.46	7.14
SNG +TS	7.78	7.09

Table 2: SNG: Single task dialogues, RS: Random_Augment Synthetic, and TS: Targeted_Augment Synthetic.

The reason for only a minor BLEU improvement could be due to the noise in generation process. Although the task distributions are matched, the token level distributions appear to be significantly different between the single and multiple-tasks. The results suggest that the method may inject more noise in the token level distribution thereby not improving the model performance significantly.

5.3 Domain Invariant Transformer

We compared the proposed architecture and the baseline Transformer model to understand the effects of domain invariant encoder representation towards language generation in multi-task dialogues. We observed from our experiments in Table 3 that Domain Invariant Transformer or Transformer model fails to generalize with few-shot multi-task dialogues. The data augmentation techniques too appear to not contribute towards improving the performance. But, Domain Invariant Transformer model improved the performance to a BLEU score when trained only on all of training data, which, though was not the intended objective. Although that seems good, the model is still heavily reliant on human-human multiple domain dialogues and zero-shot or few-shot generalization in compositional dialogues seem quite difficult to achieve.

Model	Training Data		BLEU	
	Multiple	Synthetic	Multiple	Overall
Transformer	1.00	No	14.06	14.00
Transformer	0.50	Yes	11.4	12.43
	1.00	Yes	11.89	12.32
Transformer Discriminator	0.50	No	12.24	12.13
	1.00	No	15.06	14.81
Transformer Discriminator	0.50	Yes	11.05	11.60
	1.00	Yes	11.29	12.13

Table 3: 0.5 and 1.0 correspond to half and all of multitask samples respectively during training. Synthetic refers to *Targeted_Augment* dialogues.

The poor performance of the data augmentation techniques can be due to the overwhelming noise in token distribution of input contexts, which skews the language model that the model learns.

6 Conclusion

We studied the problem of composing multiple dialogue tasks to predict next utterance in a single multiple-task dialogue. We found that even powerful transformer models do not naturally compose multiple tasks and the performance is severely relied on multiple task dialogues. In this paper, we explored two solutions that only further showed the difficulty of composing multiple dialogue tasks.

The challenge in generalizing to zero-shot composition, as observed in the experiments, hints at the possibility of transformer model potentially mimicking only the surface level tokens without understanding the underlying task. The token overlap distribution in Appendix B supports the possibility.

References

Henry Baird. 1992. Document image defect models. *Springer*.

Paweł 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. *EMNLP*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *EMNLP*.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*.

Arash Einolghozati, Panupong Pasupat, Sonal Gupta, Rushin Shah, Mrinal Mohit, Mike Lewis, and Luke Zettlemoyer. 2019. Improving semantic parsing for task oriented dialog. *arXiv*.

Mihail Eric and Christopher Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *EACL*.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. *ICML*.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. *EMNLP*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS*.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *Arxiv*.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *EMNLP*.

Arvind Neelakantan, Semih Yavuz, Sharan Narang, Vishaal Prasad, Ben Goodrich, Daniel Duckworth, Chinnadhurai Sankar, and Xifeng Yan. 2019. Neural assistant: Joint action prediction, response generation, and latent knowledge reasoning. *Arxiv*.

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 on association for computational linguistics*. ACL.

Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A network-based end-to-end trainable task-oriented dialogue system. *EACL*.

Jurgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. *CVPR*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *ACL*.

Patrice Y. Simard, Dave Steinkraus, and John C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. *ICDAR*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. 2018. Domain adversarial training for accented speech recognition. *Arxiv*.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. *ICCV*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *Arxiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *EMNLP*.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve J. Young. 2017. Latent intention dialogue models. *ICML*.

Gao Xu, Yongming Zhang, Qixing Zhang, Gaohua Lin, and Jinjun Wang. 2017. Domain adaptation from synthesis to reality in single-model detector for video smoke detection. *ArXiv*.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. *ACL*.

A Preprocessing

The MultiWoZ 2.0 dataset has a JSON metadata that maintains a dictionary of slot-value pairs provided by the user to the agent in every utterance. We use this metadata to construct a local and a global knowledge of slot-value shared by the user and split to relabel the dataset for single domain and multidomain dialogues. The preprocessing step removed the noise in the labeling of dialogues. We used this approach to keep a test set of multi-domain dialogues to evaluate the model performance on compositional tasks. On the clean split of single domain dialogues we generate synthetic multidomain dialogues using two different approaches:

A.1 Random Synthetic (RS)

In this approach, we pick a single task dialogue ${}^{i}D^{SNG}$ and randomly select a set of *K* single task dialogues,$\left({}^{i}D^{SNG}_{noise}\right)_{k=1}^{K}$, to inject noise in D^{SNG}. With an hyperparameter, *percentCopy*, we select the number of utterances to be copied from every dialogue in the set noiseDialogues and add it as a prefix to D^{SNG}. This results in K negative samples of synthetic multidomain dialogues, $\left({}^{i}D^{MUL}_{RS}\right)_{k=1}^{K}$, for every single domain dialogues in the dataset.

A.2 Targetted Synthetic (TS)

We bucket the single domain dialogues based on the conversation domain (*taxi, hotel, attraction* etc.,). Similarly, we bucket the multi-task dialogues in the training set to measure the topic distributions in multi-task dialogues. Using the computed distribution of composite tasks in *true* multidomain dialogues and the domain label of every ${}^{i}D^{SNG}$, we constrain the selection of random dialogues to conform to the training distribution of *true* composite tasks in the training set. The hyperparameters and the remainder of the procedure is similar to RS except when combining the single domain dialogues from two different domains $\left({}^{i}Dom,{}^{j}Dom\right)$, we inject the topic change exchanges randomly sampled from $TC^{\left({}^{j}Dom1,{}^{i}Dom2\right)}$.

For training the proposed Domain Invariant Transformer model, we create the labels for the auxiliary tasks using the preprocessing steps used to split the dataset into single and multi-domain dialogues

A.3 Experiments varying α

α	**BLEU (MUL)**	**BLEU**(BOTH)
0.0	14.07	13.94
0.00001	13.74	13.31
0.0001	14.13	14.11
0.001	15.06	**14.81**
0.01	14.61	14.40
0.1	14.70	14.41

Table 4: Varying the α to understand the effect of the discriminator on decoder performance

We experimented with different values of α to understand the influence of the discriminator loss. The results in Table 4 show that Domain Invariant Transformer performed better when α is 0.001. The experiment also shows consistent performance improvement in all the experiments with different α highlighting the usefulness of training an auxiliary network to train domain invariant encoder representations.

B Token distribution

We analyze the token distribution in the dataset to understand the negative result further. We observed that despite the task distributions are matched the underlying token distribution in different set up is not (Table 5). We looked at the overlap of the distribution of 4-grams in conversations on the different splits we used for training. We observed that Multi-task dialogues (MUL) training set has as much 4-gram overlap with MUL Valid and SNG (Single task dialogues) Valid sets as the combined (SNG + MUL) training data.

The analysis raises doubts in the performance of transformer model with increased MUL train dialogues that the performance improvement cannot be only because of the model's ability to decompose multiple tasks but may be because the MUL train has higher 4-gram overlap with SNG Valid and MUL Valid. This shows that despite the dialogues carrying rich information in task oriented dialogues, the model at best only mimics the surface level token distribution. Hence, it is not clear if the Transformer model can generalize to multi-task dialogues with an understanding of the underlying task structure.

USER:I'm looking for the Lensfield Hotel. Can you help me find it ?
ASSISTANT: Yes of course, it is located in 53-57 lensfield road. Would you like me to book you a reservation ?
USER: No thank you. Could you give me the postcode and tell me if they have free parking ?
ASSISTANT: Yes, they have free parking. The postcode is cb21en. Is there anything else you'd like to know about this hotel ?
USER: No, that's all I needed. Thanks! Bye!
ASSISTANT: You are welcome. Goddbye!

USER:I'd like to find a train that leaves after 17:00 on Saturday.
ASSISTANT: Absolutely, I need departure and destination information and a time you need to arrive by.
USER: I will be departing from Broxbourne and I'd like to go to Cambridge.
ASSISTANT: I have 7 available. The first leaves at 17:32, and arrives at 18:32. Would you like me to book that for you ?
USER:I am not looking to book at the moment, can I just get the train ID? Thanks.
ASSISTANT: Of course! TR0811 leaves at 17:32, arrives at 18:32, with a duration of 60 minutes. The price is 14.32 pounds. May I assist you with anything else ?
USER: That is all I need for today. Thank you. Good bye.
ASSISTANT: Okay glad I could help you. Enjoy your trip.

USER:I'm looking for the Lensfield Hotel. Can you help me find it ?
ASSISTANT: Yes of course, it is located in 53-57 lensfield road. Would you like me to book you a reservation ?
USER: No thank you. Coul you give me the postcode and tell me if they have free parking ?
ASSISTANT: Yes, they have free parking. The postcode is cb21en. Is there anything else you'd like to know about this hotel ?
USER:I'd like to find a train that leaves after 17:00 on Saturday.
ASSISTANT: Absolutely, I need departure and destination information and a time you need to arrive by.
USER: I will be departing from Broxbourne and I'd like to go to Cambridge.
ASSISTANT: I have 7 available. The first leaves at 17:32, and arrives at 18:32. Would you like me to book that for you ?
USER:I am not looking to book at the moment, can I just get the train ID? Thanks.
ASSISTANT: Of course! TR0811 leaves at 17:32, arrives at 18:32, with a duration of 60 minutes. The price is 14.32 pounds. May I assist you with anything else ?
USER: That is all I need for today. Thank you. Good bye.
ASSISTANT: Okay glad I could help you. Enjoy your trip.

Figure 2: An example of combining two single-task dialogues in color1 and color2 together to form a single multi-task dialogue.

(a) Table 1

MUL Train	492688
SNG Valid	16907
Intersection	9238
% Unseen	45.36

(b) Table 2

MUL Train	492688
MUL Valid	104261
Intersection	48076
% Unseen	53.89%

(c) Table 3

SNG Train	124038
MUL Valid	104261
Intersection	22254
% Unseen	78.66%

(d) Table 4

SNG Train	124038
SNG Valid	16907
Intersection	6562
% Unseen	61.19%

(e) Table 5

SNG+MUL Train	568674
SNG Valid	104261
Intersection	49999
% Unseen	52.04%

(f) Table 6

SNG+MUL Train	568674
SNG Valid	16907
Intersection	9746
% Unseen	42.36%

Table 5: Analysis of 4-gram overlap across different combinations of train and validation splits that were used in the experiments. The analysis show that the %Unseen in validation set is higher when training with SNG (Single domain dialogues) but considerably lower when trained with MUL. The composition task requires models to understand the underlying task structure but the data distribution and performance of transformer strongly correlate to show that the transformer model at best mimics the surface level token distribution than understanding the nature of task.