

# Domain Adaptive Task-Oriented Dialog System for Deep Understanding of Multimodal Conversation

Dongcheol Son<sup>1</sup>, Youngjae Chang<sup>2</sup>, Youngjoong Ko<sup>3\*</sup>, Jaehwan Lee<sup>4</sup>

Sungkyunkwan University,<sup>1,2,3</sup> LG Electronics<sup>4</sup>

sondongcheol94@gmail.com,<sup>1</sup> youngjaechang0@gmail.com,<sup>2</sup> yjko@skku.edu,<sup>3</sup> jaehwan314.lee@lge.com<sup>4</sup>

## Abstract


Multimodal dialog systems have attracted attention as intuitive approaches to provide users with richer information and appropriate services. However, the existing studies on the task-oriented dialog systems have focused on extracting meaningful information from the dialog context, and other studies on multimodality have not focused on the study of the task-oriented dialog system. Recently, new task and dataset called SIMMC have been proposed to study multimodal task-oriented dialog systems. SIMMC includes subtasks for successfully building a multimodal task-oriented dialog system. Therefore, in this study, we propose a model that effectively solves the SIMMC task as a study of a multimodal task-oriented dialog system. The proposed model outperformed other models in two of four subtasks, except for the response retrieval subtask in which we did not participate, and the final result shows that the proposed model is a promising approach for a multimodal task-oriented dialog system.

## Introduction

The task-oriented dialog system communicates with users through natural language to provide appropriate services or answers expected by users, such as product and route recommendations, and ticket reservations (Seneff and Polifroni 2000). The system should accurately analyze the user intention and track the user state, and generate a response that the user expected based on the analyzed and tracked information.

Existing studies on traditional task-oriented dialog systems have focused on the dialog context and have developed in a pipeline structure to deeply understand the user utterance and effectively generate a response in the form of the text. Recently, with advances in neural network approaches, end-to-end trainable dialog systems using pre-trained language models such as GPT-2 (Radford et al. 2019) have been studied such as (Gao, Galley, and Li 2019). The neural network approaches are used to improve several limitations of the modular dialog systems, such as error backpropagation and the high construction cost of labeled train data, and they showed significant performance improvements in many dialog system tasks.

In recent studies, dialog systems that combine not only dialog contexts but also multimodal contexts such as images



| UTTERANCES |   | ANNOTATIONS  |
|------------|---|--|
| USER:      | Hey, could you show me some of your more expensive jackets?   | REQUEST:GET, slots: {type: jacket, price: expensive} |
| ASSISTANT: | Sure, the grey and white jacket in the bottom left corner of the wall, the light blue jacket in the middle on the front table, and the grey jacket directly behind it are all in the expensive price range. | INFORM:GET, objects: [[11, 8, 0]]                    |

Figure 1: SIMMC 2.0 dataset consists of dialog context between the system and the user, the virtual store image called the scene, and metadata mapped for each object in the scene. The system should classify the disambiguation from the user utterance, track the user intention and state, and generate an appropriate response using the given data.

and videos, have attracted wide research attention. (Antol et al. 2015) proposed a new task and dataset called Visual Question Answering (VQA), where the system should find an appropriate answer to a given image and query. (Das et al. 2017) also proposed a new task and dataset called Visual Dialog (VisDial), where the system should generate an appropriate response to a given image and dialog context, including queries. However, most existing studies related to multimodal dialog systems have focused only on generating responses to queries well, and have not paid much attention to task-oriented dialogues such as analyzing the user intention, tracking user state, and searching for the product and services user wants (Liao et al. 2018).

SIMMC (Moon et al. 2020; Kottur et al. 2021) is a task with a dataset for a multimodal task-oriented dialog system and it was proposed as one track of the 9th Dialog System Technology Challenge (DSTC9) and the 10th Dialog System Technology Challenge (DSTC10). The dataset are based on virtual shopping scenarios that mimic the shopping experience in the real world fashion and furniture domains consisting of dialog contexts, images called scenes, and metadata mapped for each product. To recommend the appropriate product to the user, the system should analyze the user

\*Corresponding author

intention from the given dialog context, track the user state on the preferred product, and effectively combine the visual information extracted from the scene with the dialog context to search for the desired product that user wants. Since the system should communicate with users consistently in natural language with the analyzed and tracked information, SIMMC is an important task for a multimodal task-oriented dialog system.

Therefore, the capacity to effectively solve the SIMMC task is necessary to build a successful multimodal task-oriented dialog system, and we propose methods to solve the four subtasks of SIMMC. The four subtasks of SIMMC 2.0, in DSTC10, are as follows:

- **#1. Multimodal Disambiguation** The subtask involves classifying whether the assistant should disambiguate in the next turn, given a dialog context including the last user utterance. For example, when the user says, “*How much is that red shirt?*” the system can clearly understand the utterance and provide an appropriate service or information. However, when the user says, “*How much is the one over there?*”, the system cannot clearly understand the utterance. Therefore, the system should classify the utterance as disambiguated in the next turn to ask for additional information from the user.
- **#2. Multimodal Coreference Resolution** The subtask involves resolving referent objects to their canonical ID(s) as defined by the catalog, given a dialog and multimodal contexts. It is important to build a best multimodal context with metadata mapped with each object that exists in the scene to search for object ID(s) that the user wants, and the image processing model can be used to extract and use visual metadata for a multimodal context.
- **#3. Multimodal Dialog State Tracking (MM-DST)** The subtask involves tracking user belief states across multiple turns. The system should analyze the user intention and track the product information that the user wants from the dialog context and multimodal context.
- **#4. Multimodal Dialog Response Generation & Retrieval** The subtask involves generating assistant responses or retrieve from a candidate pool. The system should not only generate or search for a natural response, but also the response should include useful information on the recommended product to the user or answer the question.

In this study, we propose an encoder-based dialog model with RoBERTa (Liu et al. 2019) that has shown strong natural language understanding so far and leverages post-training (Xu et al. 2019, 2020) to learn domain-specific information for subtask 1. SIMMC provides the dataset that has domain-specific information related to fashion, furniture, and shopping scenarios, but existing studies have not focused on learning this domain information.

In addition, we proposed an encoder decoder-based model using BART (Lewis et al. 2020) for subtasks 2–4. To utilize a rich multimodal dialog context, the proposed model extracts visual metadata from the images using ResNet (He et al. 2016) and takes the visual metadata input in the form

of text tokens. Finally, the model generates a dialog state, object ID(s), and response.

As a result, we demonstrated the superiority of our proposed model by winning two of four subtasks in DSTC10 Track 3 where SIMMC 2.0 dataset were used.

## Method

This section describes the baseline and our approach for solving each subtask, including the image processing model used to extract visual metadata from the scenes.

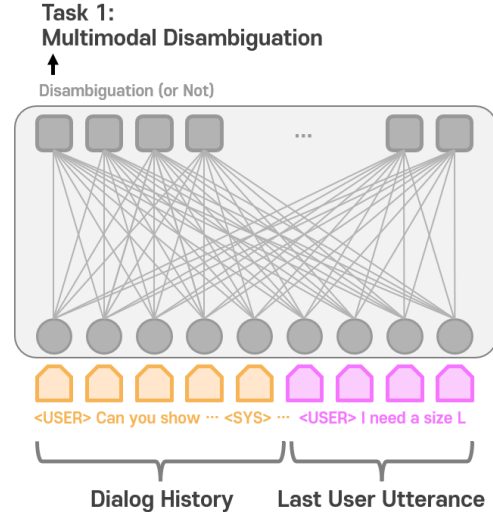


Figure 2: Overview of the proposed model for subtask 1. The dialogue context including the last user utterance is inputted in the encoder-based model, and the output [CLS] token embedding is used to classify the disambiguation from the dialog context.

## Baseline

For the study of the multimodal task-oriented dialog system, SIMMC 2.0 proposes four subtasks, and the released baseline is GPT-2. GPT-2 is a pre-trained language model based on Transformers (Vaswani et al. 2017) and has shown high performance in various natural language generation tasks. The dialog context and the object candidate ID(s) in the previous turn are provided as input data for the baseline, and the baseline should output natural language including the user intention and states, the object ID(s) that the user wants, and the response. The baseline achieved accuracy 73.5%, object F1 44.1%, slot F1 83.8%/intent F1 94.1%, and BLEU-4 0.445, for each subtask, respectively.

## Image Processing

The SIMMC 2.0 consists of dialog contexts, scenes, and metadata mapped with each object in the scene. The metadata includes visual (e.g., color and pattern) and non-visual (e.g., price and size) information of each object, but the system cannot refer to the visual information in the inference time. Therefore, the system should extract visual information from the scene using an image processing model.

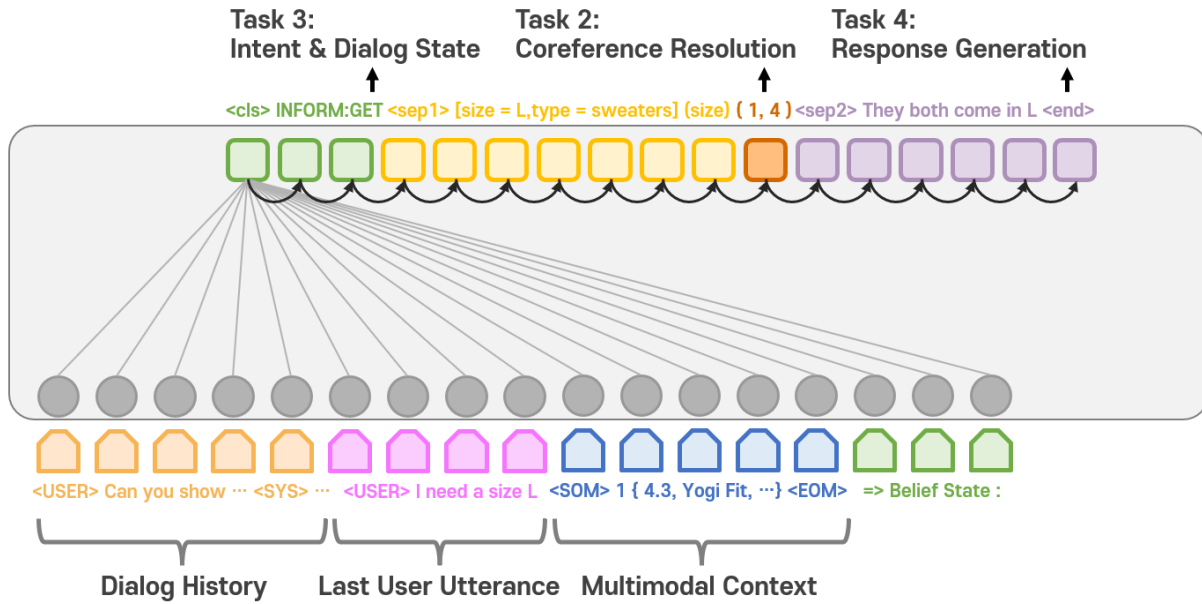


Figure 3: Overview of the encoder-decoder-based model for subtasks 2-4. The model takes the dialog and multimodal context input and it should generate the output containing the user intention and states, object ID(s) that the user wants, and natural response. The special tokens for input are used to distinguish a user utterance and system utterance in the dialog context. And the special tokens for output are used to distinguish the different types of the generated output.

In this study, we used ResNet as an image processing model to build a multimodal context with rich information for model input. ResNet got input the scene and bounding box information of each object and was fine-tuned to correctly classify the visual information of each object by using the visual information in the metadata file as the ground-truth labels.

### Subtask #1

**Encoder-based Model** Figure 2 shows an overview of the proposed model for subtask 1. We propose an encoder-based model structure for our approach in a multimodal disambiguation subtask. It is important to understand the dialog context for classifying the disambiguation of the user utterance, so the capacity of strong natural language understanding is necessary to fully solve the subtask. However, the released baseline, GPT-2, is a decoder-based model that has an advantage in natural language generation tasks. Therefore, we changed the model to encoder-based models such as BERT (Devlin et al. 2019), ALBERT (Lan et al. 2019), and RoBERTa with strong natural language understanding for this subtask. We eventually used RoBERTa, which has shown high performance as far in our proposed model in the experiment.

The model takes only the dialog context input, including the last user utterance, and outputs the [CLS] token embedding. The embedding is inputted to the MLP layer and trained as a binary classification task to classify the disambiguation.

**Post-training** In this study, we use a post-training method to add domain information to the pre-trained model. The

method aims to deepen the understanding of domain-specific information by learning the language representation of the target data. For example, while the word "killer" generally connotes a negative sentiment, "killer contents" is used as a phrase of positive sentiment in some domains, such as broadcast and movie.

Since SIMMC 2.0 consists of virtual shopping scenarios for fashion and furniture domains, it is a domain-specific dataset. However, in existing studies on SIMMC, no studies have focused on learning domain information. Therefore, we here leverage post-learning to learn domain-specific information of SIMMC and improve performance.

MLM is a learning method that randomly masks the tokens in the input data and predicts the masked token based on the context, and it is effective for learning the language representation. In this study, we used MLM based post-learning to reduce the learning bias of the existing pre-trained model that was trained only with articles and dictionary data; at the same time, we can add domain-specific information of the target data to our encoder-based model. We set the random masking probability of the MLM to 15%, and the loss of post-training was computed using the same method as the standard BERT.

### Subtask #2-4

**Encoder Decoder-based Model** Figure 3 shows an overview of the proposed model for subtasks 2- 4. Herein, we use the encoder-decoder-based model, BART. To successfully conduct three subtasks including multi-modal coreference resolution, multimodal dialog state tracking, and response generation, it is important to correctly extract

Input : Okay, then how much is the brown dress again?

Model Output : ... <sep1> [ ] (price) < 44 > <sep2> It costs \$224.99. <end>

Ground Truth Output : ... <sep1> [ ] (price) < 43 > <sep2> It costs \$239.99. <end>

Task3: Task2: Task4:  
Dialog State Object ID(s) Response

(\*Object ID 44 is a \$224.99 dress and Object ID 43 is a \$239.99 dress)

Figure 4: The typical example of the error propagation problem in the image processing model. Although the proposed model correctly tracked the user states, it failed to search for the appropriate product ID(s) due to incorrect visual metadata, and generated responses with incorrect metadata. The system had to search for object ID 43 with a price of \$239.99 as the correct object the user wants, but incorrectly search for object ID 44 with a price of \$224.99 due to incorrect visual information of the image processing model.

visual metadata from scenes and combine the dialog context and the multimodal context to track user intention and states. However, because these subtasks are closely related, the performance of multimodal coreference resolution and response generation subtasks might be affected if the user intention or states are tracked incorrectly. Therefore, it is important to correctly understand and track the information deeply from the given contexts, but the baseline GPT-2 is a decoder-based model that has an advantage in natural language generation tasks. That is a reason why we employ an encoder decoder-based model, BART, which has shown high performance in natural language generation based on strong natural language understanding.

Our model gets the dialog and multimodal context input. The multimodal context consists of non-visual and visual metadata; visual metadata is predicted as values represented in text token form for color, type, pattern, and material from the given image using the fine-tuned ResNet. The model is trained end-to-end to generate natural languages with user intention, dialog states, and object ID(s) that the user wants and the response.

## Experiments

### Dataset

Table 1 shows the statistics of SIMMC 2.0 dataset used for this experiment. The total number of data is 11,244 consisting of virtual shopping scenarios for fashion and furniture domains, and each dialog consists of an average of 10.4 utterances. Each dialog is mapped to a virtual image called a scene, and a dialog might be mapped to more than two scenes. The system might refer to metadata mapped to each object in the scene, but visual metadata should not be referred to in the inference time. Therefore, the system should build a multimodal context by extracting visual metadata from the given image.

### Experiment environment

Table 2 shows the distribution of the entire data split into train, dev, and test. 64% of the data was used as the training set, 5% as development set, and the remaining 30% as test-dev (15%) and test-std (15%). Test-std were used to evaluate

|                                    |         |
|------------------------------------|---------|
| Total # dialogs                    | 11,244  |
| Total # utterances                 | 117,236 |
| Total # scenes                     | 3,133   |
| Avg # words per user turns         | 12      |
| Avg # words per assistant turns    | 13.7    |
| Avg # utterances per dialog        | 10.4    |
| Avg # objects mentioned per dialog | 4.7     |
| Avg # objects in scene per dialog  | 19.7    |

Table 1: Statistics of SIMMC 2.0 dataset

all participating team models in the DSTC10 Track 3; these data were not released.

In this experiment, we used the RoBERTa-large model to solve subtask 1, and the hyperparameters were set as follows: the learning rate, max sequence length, and epoch were 2e-5, 512, and 5, respectively. For subtasks 2–4, we used the BART-large model, and the hyperparameters were set as follows: the learning rate, max sequence length, epoch, and dialog context length were 5e-5, 1,024, 10, and 6, respectively. The mini-batch size of both models was set to 6, and Adam (Kingma and Ba 2014) was used as a function for optimization.

As for evaluation metrics, we used accuracy to evaluate the number of examples that the model correctly predicted for subtask 1, and used slot F1 score and intent F1 score for subtask 2. The F1 score is a metric that is computed as the harmonic average of the precision +and recall and is effective when evaluating the performance in an environment where the amount of data is unbalanced for each category.

| Split          | Number of Dialogs |
|----------------|-------------------|
| Train (64%)    | 7,307             |
| Dev (5%)       | 563               |
| Test-Dev (15%) | 1,687             |
| Test-Std (15%) | 1,687             |

Table 2: Distribution of SIMMC 2.0 dataset

| Team ID  | #1. MM-Disamb. | #2. MM-Coref | #3. MM-DST   |           | #4. Response Generation |
|----------|----------------|--------------|--------------|-----------|-------------------------|
|          | Accuracy       | Object F1    | Slot F1      | Intent F1 | BLEU-4                  |
| Baseline | 73.5%          | 44.1%        | 83.8%        | 94.1%     | 0.202                   |
| 1        | -              | 52.1%        | 88.4%        | 96.3%     | 0.285                   |
| 2        | -              | 78.3%        | -            | -         | -                       |
| 3        | 89.5%          | 42.2%        | 87.8%        | 96.2%     | 0.256                   |
| 4        | 93.9%          | <b>75.8%</b> | 90.3%        | 95.9%     | 0.295                   |
| 5        | 93.8%          | 56.4%        | 89.3%        | 96.4%     | <b>0.322</b>            |
| 6 (Ours) | <b>94.7%</b>   | 59.5%        | <b>91.5%</b> | 96.0%     | 0.309                   |
| 7        | 93.1%          | 57.3%        | -            | -         | -                       |
| 8        | 93.1%          | 68.2%        | 4.0%         | 41.4%     | 0.297                   |
| 9        | -              | 73.3%        | -            | -         | -                       |
| 10       | 93.6%          | 68.2%        | 87.7%        | 95.8%     | 0.327                   |

Table 3: The final results of DSTC10 Track 3 using SIMMC 2.0 task and data. Our proposed model is Team 6, which ranks 1st in the Multimodal Disambiguation and Multimodal Dialog State Tracking subtasks.

## Experimental Result

Table 3 shows the final result of the baseline, our proposed model, and other team models for the four subtasks of SIMMC 2.0. Our proposed model outperformed other models as a winner in two of the four subtasks, except the subtask for the response retrieval, where we did not participate. Multimodal disambiguation is a subtask to classify whether the assistant should disambiguate in the next turn, given the dialog context, and our proposed model achieved 94.7% accuracy. It is not only about 21.2%p higher performance than the baseline, but also about 0.8%p higher than the second-ranked performance. The results demonstrate that our approach to solve the multimodal disambiguation problem is very effective, and learning the domain information of the target data is also helpful for understanding the data for multimodal task-oriented dialogue systems. The multimodal coreference resolution is a subtask to correctly search for the object ID(s) that the user wants on the dialog and multimodal context, and our proposed model achieved 59.5% of object F1 score. Performance was ranked fifth among all 10 teams. Multimodal dialog state tracking is a subtask to understand user intentions and track states such as color, type, and size of the product that the user wants. Our proposed model scored 91.5% for slot F1, and 96.0% for intent F1 ranking 1st in the cumulative scoring. The response generation is a subtask of generating an adequate response for the last user utterance, and our model achieved 0.309 on the BLEU-4 metric, ranking 3rd, which was a small gap compared with the 1st and 2nd teams.

## Analysis

In this section, we analyze the cause of the low performance of the proposed model in two subtasks. The proposed model showed relatively low performance in multimodal coreference resolution and response generation subtasks. Solving these two subtasks requires the metadata of correctly ex-

tracted objects. However, as the performance of the image processing model, ResNet, was not good, it generates lots of incorrect visual metadata. We think that it is a main reason why we obtained low performance in multimodal coreference resolution and response generation subtasks in spite of using a strong natural language understanding and generation model, BART.

Figure 4 shows a typical example of the error propagation problem. The proposed encoder decoder-based model outperformed other models on the MM-DST subtask owing to the better natural language understanding, which means that the system correctly tracked the user intention and states. As shown in Figure 4, the model correctly tracked information about the price that the user want to know. However, because the visual information of the object was predicted incorrectly, the system failed to search for the correct object ID(s). In Figure 4, The system had to search for object ID 43 with a price of \$239.99 as the correct object the user wants, but incorrectly search for object ID 44 with a price of \$224.99 due to incorrect visual information of the image processing model. As a result, the system generated a response with incorrect information. Therefore, if we would improve the performance of the image model on SIMMC 2.0 dataset in future work, we could achieve excellent performances in all the subtasks of the SIMMC 2.0.

## Conclusions

In this study, we proposed an encoder-based model based on post-training to deepen the understanding of SIMMC dataset using RoBERTa for subtask 1. To the best of our knowledge, this is the first study to learn the domain information of SIMMC data. Furthermore, we proposed an encoder decoder-based model using BART with ResNet to conduct subtask 2-4. The final result shows that our proposed model is a promising method for solving multimodal task-oriented dialog task.



## Acknowledgments

This work was supported in part by LG Electronics (CTO/AI lab) and in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

## References

- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Gao, J.; Galley, M.; and Li, L. 2019. *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations and Trends.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. *arXiv preprint arXiv:2104.08667*.
- Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880. Online: Association for Computational Linguistics.
- Liao, L.; Ma, Y.; He, X.; Hong, R.; and Chua, T.-S. 2018. Knowledge-Aware Multimodal Dialogue Systems. In *Proceedings of the 26th ACM International Conference on Multimedia, MM '18*, 801–809. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356657.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Moon, S.; Kottur, S.; Crook, P. A.; De, A.; Poddar, S.; Levin, T.; Whitney, D.; Diffranco, D.; Beirami, A.; Cho, E.; et al. 2020. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Seneff, S.; and Polifroni, J. 2000. Dialogue Management in the Mercury Flight Reservation System. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.