A Generative Dialogue Model for Travel Booking

Valeriia Lelik

July 2025

Abstract

This project proposes a generative dialogue model aimed at travel booking tasks, including flight and hotel reservations. The model is trained on the Schema-Guided Dialogue dataset, which contains richly annotated dialogues with intents, slots, and service schemas. We employ a T5-small transformer architecture with a "Show, Don't Tell" input representation approach. Evaluation is conducted on a held-out test subset of 500 dialogues using standard automatic metrics and additional manual analysis. The system demonstrates basic fluency and contextual relevance in generating structured dialogue responses in the travel domain.

1 Introduction

Task-oriented dialogue systems play a crucial role in domains such as travel and hospitality, where clear and informative communication is essential. Creating such systems involves solving complex problems like intent and named-entity recognition and natural language generation.

In this project, we focus on building a generative model for travel-related dialogues using the Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2019). We use the T5-small model and apply a linearized input format inspired by the "Show, Don't Tell" strategy (Gupta et al., 2022), where dialogue acts and user utterances are explicitly encoded into the model input.

Although we experimented with a Named Entity Recognition (NER) component based on BERT (Devlin et al., 2019), the current pipeline does not yet integrate NER outputs into training. Future iterations could benefit from tighter integration of structured semantic information.

2 Dataset and Preprocessing

2.1 Schema-Guided Dialogue Dataset

The SGD dataset contains over 20,000 task-oriented dialogues across 16 domains. Each dialogue includes turns between a user and system, labeled with intents, slots, and system actions.

For this work, we focus on booking scenarios (flights, hotels, etc.). The dataset splits used are:

- **Training set**: 5,000 dialogues randomly sampled from the SGD `train` split.
- **Validation set**: 1,000 dialogues from the `validation` split were loaded but not finally used.
- **Test set**: 500 dialogues from the official SGD `test` split, selected, scored automatically and then estimated manually.

2.2 Linearized Input Format

To transform dialogues into model-friendly format, we use a linearized input structure. For each system turn, we construct:

- The user's utterance from the previous turn.
- System actions in the form `ACT(slot=value)`, extracted from frame annotations.
- A target string: the system response (utterance).

This is in line with the "Show, Don't Tell" principle, which encourages the model to condition generation on explicit semantic cues rather than raw dialogue history alone.

2.3 Training Details

The model is based on `t5-small`. We tokenize input-target pairs with a maximum length of 128/64 tokens, respectively. Training is conducted over one epoch with batch size 1 and gradient accumulation of 4 steps, using Huggingface's `Trainer`. No intermediate evaluation or early stopping was used.

Although a BERT-based NER module (`dslim/bert-base-NER`) was loaded, it was used only for exploratory purposes and not included in the final training data.

3 Evaluation

3.1 Automatic Metrics

We evaluated the fine-tuned model on a manually curated test subset of 500 dialogues from the SGD `test` split. For each system turn, we reconstructed prompts and compared generated responses to the ground-truth utterances.

Metrics used:

- **BLEU** (Papineni et al., 2002): 0.0374
- **ROUGE-L** (Lin, 2004): 0.1360

These relatively low scores reflect the difficulty of the task, especially for small generative models and under limited supervision. Nevertheless, some outputs captured slot structure and intent alignment well.

3.2 Manual Analysis of Generated Outputs

A manual qualitative review of generated responses reveals that the model is capable of producing relevant and contextually appropriate outputs, especially when system actions and slot-value pairs are explicitly and clearly encoded in the input. For example, the model often correctly generates confirmations reflecting reservation details such as date, time, number of seats, and venue names.

However, several recurring issues were identified that contribute to the modest BLEU and ROUGE scores:

- **Entity Substitution Errors:** The model sometimes confuses or replaces entities inaccurately, e.g., misnaming restaurants or omitting location details, which leads to incomplete or misleading responses.
- **Hallucinations:** Unprompted content occasionally appears in outputs, such as mentioning vegetarian options or user ratings without relevant input, indicating generation of unsupported information.
- **Omission of Key Slots:** Critical information like reservation time or party size is sometimes missing from the generated utterances, reducing their informativeness and fidelity.
- **Vague Confirmations:** Responses occasionally provide generic acknowledgments without explicitly confirming important slot values, which, while conversationally natural, lowers overlap with reference texts.
- **Dependence on Input Encoding:** The quality and specificity of system action encodings in the prompt heavily influence output accuracy. Clear, detailed semantic annotations yield better grounding and slot realization.

These findings suggest that while the model demonstrates promising capabilities in task-oriented dialogue generation, further improvements are necessary to mitigate hallucinations, enhance slot fidelity, and increase output specificity, thereby improving automatic evaluation metrics.

4 Conclusion

We present a baseline generative model for travel booking dialogue generation, trained on the Schema-Guided Dialogue dataset with a T5-small architecture. Despite modest evaluation scores, the model demonstrates basic fluency and contextual relevance.

Our pipeline uses structured prompts that include user utterances and system action annotations, following the “Show, Don’t Tell” approach. While NER was not fully integrated into training, this infrastructure is in place for future experimentation.

5 Future Work

Next steps may include:

- Integrating NER outputs directly into model prompts to highlight slot entities.
- Using a validation set for tuning and early stopping.
- Scaling up to larger models (e.g., T5-base) and longer input contexts.
- Applying reinforcement learning or contrastive learning to improve response grounding.
- Incorporating human evaluation to assess dialogue coherence and task completion.

These improvements are likely to lead to better slot realization, context tracking, and overall task-oriented performance.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. https://aclanthology.org/N19-1423

Gupta, P., Rastogi, A., & Williams, J. D. (2022). Show, don't tell: A simple and effective approach to task-oriented dialogue. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI (NLP4ConvAI)*, 41–51. https://aclanthology.org/2022.nlp4convai-1.6

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. https://aclanthology.org/W04-1013

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, 311–318. https://doi.org/10.3115/1073083.1073135

Rastogi, A., Zang, Q., Sunkara, S., Gupta, R., & Khaitan, P. (2019). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*. https://doi.org/10.48550/arXiv.1909.05855