Leveraging Graph Structures and Large Language Models for End-to-End Synthetic Task-Oriented Dialogues

Demonstration Track

Maya Medjad
UCBL, CNRS, Centrale Lyon, INSA Lyon, Univ. Lumière Lyon 2, LIRIS, UMR5205
69622 Villeurbanne, France
maya.medjad@univ-lyon1.fr

Hugo Imbert
Reecall
69000 Lyon, France
hugo@reecall.co

Bruno Yun
UCBL, CNRS, Centrale Lyon, INSA Lyon, Univ. Lumière Lyon 2, LIRIS, UMR5205
69622 Villeurbanne, France
bruno.yun@univ-lyon1.fr

Raphaël Szymocha
Reecall
69000 Lyon, France
raphael@reecall.co

Frédéric Armetta
UCBL, CNRS, Centrale Lyon, INSA Lyon, Univ. Lumière Lyon 2, LIRIS, UMR5205
69622 Villeurbanne, France
frederic.armetta@univ-lyon1.fr

ABSTRACT

Training task-oriented dialogue systems is both costly and time-consuming, due to the need for high-quality datasets encompassing diverse intents. Traditional methods depend on extensive human annotation, while recent advancements leverage large language models (LLMs) to generate synthetic data. However, these approaches often require custom prompts or code, limiting accessibility for non-technical users. We introduce GraphTOD, an end-to-end framework that simplifies the generation of task-oriented dialogues. Users can create dialogues by specifying transition graphs in JSON format. Our evaluation demonstrates that GraphTOD generates high-quality dialogues across various domains, significantly lowering the cost and complexity of dataset creation.

KEYWORDS

Task-oriented dialogue; Large-language models; Synthetic data

ACM Reference Format:
Maya Medjad, Hugo Imbert, Bruno Yun, Raphaël Szymocha, and Frédéric Armetta. 2025. Leveraging Graph Structures and Large Language Models for End-to-End Synthetic Task-Oriented Dialogues: Demonstration Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Task-Oriented Dialogue Systems (TODS) are increasingly used in domains like customer support, personal assistants, and enterprise solutions to help users achieve specific objectives through natural language conversations [2, 4, 13, 18].

Traditional TODS rely on machine learning models trained on predefined schemas [4, 14], but they struggle with complex/nuanced dialogues, especially when domain-specific data is scarce. In contrast, LLM-based TODS enable more human-like and engaging responses [1, 2, 7, 16]. However, these systems are prone to hallucinations [5, 6], needing fine-tuning for specific use cases [11, 15].

Fine-tuning LLM-based TODS requires large amounts of training data, with diverse and high-quality structures, which are costly and time-consuming to collect [19]. A diverse dataset of realistic dialogues is essential to allow these systems to grasp the nuances and unique patterns of human conversation. This becomes even more problematic when several TODS are required, each for a different task, spanning different domains (e.g., hotel booking).

Data has previously been collected with human workers [3], this type of dataset is particularly expensive to create even when users are assisted by computers [8, 14]. Synthetic data generation is now a common approach to acquire data for TODS [16, 17].

For example, *SynTOD* [15] proposes a new approach to model TODS behavior using a state transition graph. However, it was inaccessible to non-experts, as the graph and the corresponding prompts had to be implemented manually in Python. To alleviate this problem, we propose GraphTOD, a generalisable framework powered by LLMs with a new state machine-based prompt that allows non-technical users to generate task-oriented dialogues by specifying transition graphs in JSON format. The source code and the demo video are available on Github[1] and Youtube[2] respectively.

2 THE GRAPHTOD GENERATION PIPELINE

GraphTOD is based on two agents (system and user) which simulate dialogue utterances by navigating an *action transition graph* (see Figure 1). We formalize each of those elements.

An action transition graph is a tuple $G = (V, Ac, E, t, s, f)$, where V is a set of nodes, Ac is a set of actions, $s, f \in V$ are the initial and final states, $E \subseteq (V \setminus \{f\}) \times Ac$ is a set of available actions

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1] https://github.com/reecall/GraphTOD
[2] https://youtu.be/5pXa4yGcc58

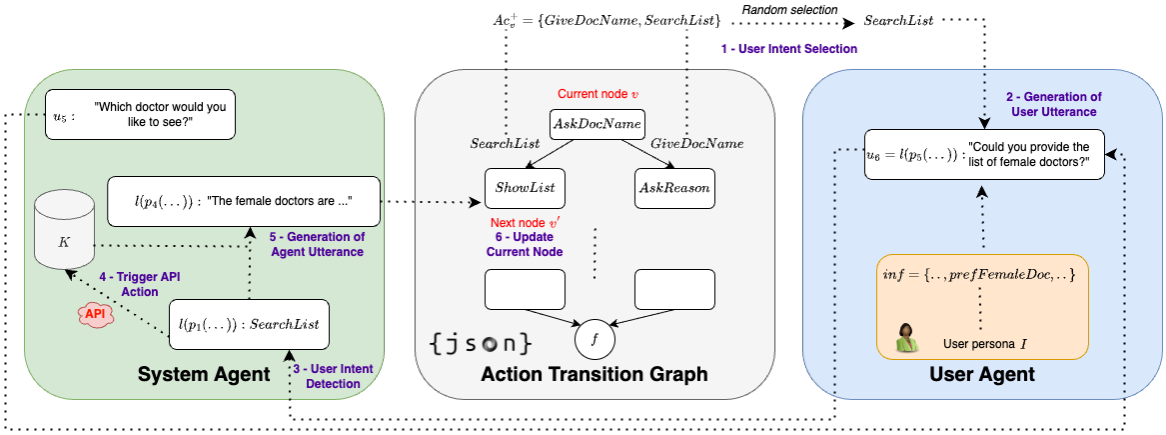

Figure 1: Representation of one turn of generation of the GraphTOD generation pipeline

at each non-final node, and $t : (V \setminus \{f\}) \times Ac \to V$ is the transition function. The action transition graph serves as the link between the two agents and can be specified quickly in JSON format.

Given an action transition graph G, a subset of actions $F \subseteq Ac$, called *function calls*, are associated with APIs to obtain external knowledge. For a node $v \in V$, the possible actions at v are denoted $Ac_v^+ = \{a \in Ac \mid (v, a) \in E\}$. Our pipeline makes use of five carefully crafted prompt templates[3] denoted by p_i, $1 \leq i \leq 5$. These prompt templates take as input a set of parameters and return a formatted string for an LLM. A dialogue history is $H = (u_1, u_2, u_3, \dots, u_n)$, where u_{2j} and u_{2j+1} represent the user's and system's utterances, respectively, at time $j \geq 0$. We denote an LLM as a (possibly non-deterministic) function l that outputs $l(x)$ for input x.

The system agent is defined as $A_s = (G, F, P_s, K, u_1)$, where G is an action transition graph, F is a set of function calls, $P_s = \{p_1, \dots, p_4\}$ is a set of system prompt templates, K is an agent knowledge database initialized at $\emptyset$, and u_1 is a starting utterance. Considering a dialogue history $H = (u_1, u_2, \dots, u_{2j+1})$, a user utterance u_{2j+2}, and the current node $v \in V$, the system agent performs a two-steps reasoning. First, $p_1(H, u_{2j+2}, K, Ac_v^+)$ is used to make the LLM detect the user's intention. Second, depending on the detected intention, potential APIs are triggered to collect knowledge, and the system utterance is generated to either state that the intent was not recognized, end the conversation, or continue the conversation (using the corresponding prompt templates p_2, p_3, or p_4).

The user agent is defined as $A_u = (I, P_u)$, with $P_u = \{p_5\}$ a set of user prompt templates, and $I = (age, name, gender, prefs)$ the agent's persona, where $age \in \{18, 19, \dots, 80\}$, $name$ is generated using the Faker library[4], $gender \in \{male, female\}$, and $prefs$ is a list of preferences generated based on G and can be topics or places linked to it. For $v \in V$ and $a \in Ac_v^+$, the user agent uses $p_5(a, G, H, I)$ to generate the next user utterance reflecting action a at v.

Example 2.1. Consider the generation of a dialogue between a medical chatbot and a user, guided by the action transition graph of Figure 1. At node $v = AskDocName$, the system agent initiates with u_5 = "Which doctor would you like to see?". A random user intention $SearchList \in Ac_v^+$ is selected as the next action. Based on this intent and the user preference for female doctors ($prefFemaleDoc$), the user agent formulates the response u_6 ="Could you provide the list of female doctors?". The system agent processes the user intent ($SearchList$), queries the relevant API to retrieve the list of doctors, and generates the next system utterance u_7 ="The female doctors are...". Finally, the current node is updated to $ShowList$.

[3]We refer the reader to the Github repository for more details.
[4]https://faker.readthedocs.io/. *name* is kept consistent with the agent's *gender*.

3 EVALUATION

We generated 150 conversations on 4 domain scenarios (Recipe, Hotel, RentCar, Doctor) using GraphTOD and OpenAI's GPT-4[5]. We evaluated the dialogues using 3 UniEval [20] metrics (naturalness, coherence, understandability) as classic NLP metrics (e.g., BLEU [12] or ROUGE [10]) are insufficient to portray the difference between advanced generation models. In Table 1, GraphTOD performs consistently well overall and reports similar performances to human-in-the-loop approaches based on LLMs such as LAPS [9].

Model - Scenario	Nat.	Coher.	Under.	Mean
(Baseline) LAPS - Recipe	0.867	0.891	0.860	0.872
(Baseline) LAPS - Movie	0.874	0.897	0.868	0.880
(Ours) GraphTOD - Recipe	0.899	0.857	0.890	0.882
(Ours) GraphTOD - Hotel	0.888	0.853	0.879	0.873
(Ours) GraphTOD - RentCar	0.887	0.768	0.878	0.844
(Ours) GraphTOD - Doctor	0.835	0.800	0.827	0.820

Table 1: Performance comparison with UniEval metrics.

4 CONCLUSION

GraphTOD is an end-to-end LLM pipeline designed to generate high-quality task-oriented conversations by using an action transition graph and a generalized prompting approach. GraphTOD also includes the automatic generation of user-agent preferences from the input graph and LLM-powered intent detection, resulting in a fully automated and fault-tolerant pipeline.

[5]Refer to the Github repository for the corresponding action transition graphs.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[2] Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, Alexandra Birch, Andrew M. Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh (Eds.). Association for Computational Linguistics, 15–22. https://doi.org/10.18653/V1/D19-5602

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2020. MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. arXiv:1810.00278 [cs.CL] https://arxiv.org/abs/1810.00278

[4] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Found. Trends Inf. Retr.* 13, 2-3 (2019), 127–298. https://doi.org/10.1561/1500000074

[5] Vojtěch Hudeček and Ondřej Dušek. 2023. Are LLMs All You Need for Task-Oriented Dialogue? arXiv:2304.06556 [cs.CL] https://arxiv.org/abs/2304.06556

[6] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[7] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023). https://doi.org/10.48550/ARXIV.2310.06825 arXiv:2310.06825

[8] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 796–806. https://doi.org/10.1145/3626772.3657815

[9] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024 (SIGIR 2024)*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 796–806. https://doi.org/10.1145/3626772.3657815

[10] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[11] Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024. TOAD: Task-Oriented Automatic Dialogs with Diverse Response Styles. arXiv:2402.10137 [cs.CL] https://arxiv.org/abs/2402.10137

[12] 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* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[13] Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end Task-oriented Dialogue: A Survey of Tasks, Methods, and Future Directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5925–5941. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.363

[14] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. arXiv:1909.05855 [cs.CL] https://arxiv.org/abs/1909.05855

[15] Chris Samarinas, Pracha Promthaw, Atharva Nijasure, Hansi Zeng, Julian Killingback, and Hamed Zamani. 2024. Simulating Task-Oriented Dialogues with State Transition Graphs and Large Language Models. *CoRR* abs/2404.14772 (2024). https://doi.org/10.48550/ARXIV.2404.14772 arXiv:2404.14772

[16] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. A Survey on Recent Advances in Conversational Data Generation. arXiv:2405.13003 [cs.CL] https://arxiv.org/abs/2405.13003

[17] Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024. LUCID: LLM-Generated Utterances for Complex and Interesting Dialogues. arXiv:2403.00462 [cs.CL] https://arxiv.org/abs/2403.00462

[18] Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-Oriented Dialogue System as Natural Language Generation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2698–2703. https://doi.org/10.1145/3477495.3531920

[19] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* 63, 10 (2020), 2011–2027.

[20] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. arXiv:2210.07197 [cs.CL] https://arxiv.org/abs/2210.07197