Active Dialogue Simulation in Conversational Systems

Anonymous ACL submission

Abstract

Our goal is to utilize large language models and active learning to replace Wizard-of-Oz (WoZ) collection via crowdsourcing for bootstrapping training data for task-driven semantic parsers. We first demonstrate the utility of utterances generated by GPT-3 when seeded with prior training dialogues, as evaluated by human judges. We then explore two approaches for example selection: maximizing model (parser) uncertainty on generated outputs, and maximizing lexical diversity. We find that large language models can generate useful training data, and that there is a promising direction in *active generation* to maximize the impact of each such example.

1 Introduction

Semantic parsers power conversational systems in satisfying user requests, e.g., modifying calendar entries, making reservations, asking questions, and buying tickets through dialogues (Bordes et al., 2016; Yu et al., 2019a; Andreas et al., 2020). These parsers translate natural utterances into executable programs, typically constructed through access to a large amount of annotated training data (Guu et al., 2017; Yu et al., 2019b). The complex nature of natural dialogues and attendant semantic representations account for the fact that relatively few large-scale corpora exist, targeting a limited number of domains.

Building natural semantic parsing corpora requires (1) collecting examples of a user interacting with a software agent (i.e., user utterances in the form of a dialogue); and (2) annotating those utterances (i.e., writing an executable program for each utterance). In this work, we focus on the first step: how to efficiently produce examples of interactions with a software agent. Ideally, one might wish to simply deploy a conversational system to real users, then use those interactions as the data to drive future improvements to the agent. Yet in practice, real user interactions with software agents are often protected as a matter of privacy, and without initial annotated examples, there is no trained software agent to drive ongoing data collection.

We turn to the use of large language models (LLMs), focusing on GPT-3 (Brown et al., 2020), with the goal of replacing humans in generating example interactions (user utterances) with a software agent. We first consider the *utility* of GPT-3 prompted generation (to replace humans), measured for diversity and human assessed quality. Experimental results on conversational system benchmarks Taskmaster-3 (Byrne et al., 2019), and SMCalFlow (Andreas et al., 2020) illustrate the promise of this approach.

We then consider the cost of annotation: can we generate and select example dialogues that are most useful to annotate for improving a semantic parser? We first introduce an approximation of uncertainty for a black-box parser. Then, we investigate the effect of different active learning schemes in improving parser accuracy. Our findings suggest the combination of LLMs and active learning is an effective approach for bootstrapping initial data in rich semantic parsing domains.

2 Related Work

Semantic parsers play a major role in conversational systems by translating natural utterances into executable programs (Zettlemoyer and Collins, 2009; Dong and Lapata, 2018; Cheng et al., 2020).

Prior work has considered how to minimize the cost of semantic parsing training data collection. Work such as Williams et al. (2015) proposed active learning for example selection, while Yao et al. (2020) exemplify strategies for interactively providing feedback to a system on its interpretation of a given example. Shah et al. (2018), Lin et al. (2020) and Acharya et al. (2021) combine a user with a system simulator and crowdsourcing.

Closest to this work are efforts defining a *user* *simulator* interaction with a dialog system in a reinforcement learning (self-play) setting to gather the data (El Asri et al., 2014; Su et al., 2017; Tseng et al., 2021). Such approaches have the benefit of complete data generation without a human annotation step, but mostly have relied on templatic language generation, or logical forms.

In this work we are concerned with the generation of natural language and adopt a different approach, i.e., directly incorporating large autoregressive language models (Radford et al., 2019; Brown et al., 2020) to simulate users. Moreover, to maximize the efficiency of our annotation process, we consider active learning strategies (Sener and Savarese, 2018; Ren et al., 2020) to identify the most informative generated outputs from language models and augment them into the training set.

3 Actively Simulating a User

Here we describe our framework to generate examples of user interactions with a software agent for training the parser. We adopt the state-of-the-art semantic parser on SMCalFlow (Platanios et al., 2021) as our base parser throughout the paper. Since this base parser does not require agent responses, in this work we only focus on generating utterances for user's turns.

To generate dialogues, we start by generating the first utterances. We target the setting where we have relatively few examples in a domain, in this case $N = 250$. We prompt GPT-3 through a random selection of the k first utterances in dialogues from the N available dialogues, conditionally generating utterances similar to instances in the prompt. For example:

> Generate a similar utterance.
> U: What time is my dinner scheduled?
> ...
> U: Is it going to snow in Spokane?
> U: ___

A natural question that might arise is whether generating utterances based on our proposed approach will have good quality and diversity. We empirically investigate this in Section 4.

Using our constructed prompts, we can generate lots of first user utterances, but since many of them will be very similar, we need to filter the most informative ones to have more efficient dialogue generation. We utilize two approaches to select candidates from generated utterances: (1) parser uncertainty, or (2) example diversity. Typically, a semantic parser is employed in an environment such that the top-1 prediction is used in a downstream conversational system. Such use cases do not typically require a confidence-calibrated model. Here, to approximate the parser uncertainty, we illustrate a post-hoc confidence estimation strategy based on measuring the average pairwise differences between the elements of a k-best list of model predictions. We investigate this empirically in Section 5. As our diversity-based sampling baseline, we use the concept of Core-sets (Sener and Savarese, 2018) applied on sentence representations based on S-RoBERTa (Reimers and Gurevych, 2019).

After filtering the generated utterances, to generate the whole dialogue, we iteratively generate the next user utterance in the dialogue by prompting GPT-3 (we limit our generation to dialogues with 1-3 turns). To better capture dialogue history while generating the next utterance, instead of randomly choosing our prompts' examples, we choose the most similar dialogues from the seed training data. We measure similarity using Levenshtein distance of seed dialogues with our so far generated dialogue. Then, concatenating our current generated dialogue to the prompt (e.g., *U1* in the prompt below), we ask GPT-3 to generate the next user turn. Assuming we want to generate the second user turn in a dialogue, we construct prompts like this:

> Generate the next utterance in the dialogue.
> U1: When is my today event on calendar?
> U2: When is my second event tomorrow?
> ...
> U1: When is my sister's birthday? (this utterance was generated in the earlier stage)
> U2: ___

To further improve the efficiency of our pipeline, after generating dialogues using GPT-3, we select the most informative ones in an active setting. We calculate a score for dialogues by taking *max* over all utterances' score in a dialogue (whether using uncertainty or diversity for scoring).[1]

4 Intrinsic Generation Quality

The first challenge in utilizing GPT-3 to populate conversational system datasets is determining whether the generated instances are diverse and high quality. By quality we mean how likely a real user might state a given utterance in a conversation on the specific domain. Considering real users

[1] During development we confirmed that using the *mean* of utterances' score for scoring dialogues was not effective.

	CalFlow	Taskmaster
Orig	**73.25**	**75.53**
Gen	68.75	67.07

(a) Quality.

	Max-D	Ent
Orig	**15.02**	5.87
Gen	14.01	**6.51**

(b) Diversity, SMCalFlow.

Table 1: Quality and diversity of generated vs original utterances. We evaluate diversity in SMCalFlow by calculating pair-wise maximum distance (**Max-D**) and entropy (**Ent**) based on S-RoBERTa representations.

	Hits@1	Hits@10
Random (250)	41.2	57.7
Random (1000)	53.9	71.8
Core-set (1000)	54.8	72.4
Uncertainty (1000)	**56.2**	**73.3**

Table 2: Effect of active learning approaches in sampling SMCalFlow dialogues in a low-resource setting. We start from 250 random samples and add extra 750 samples based on different sampling methods.

might ask grammatically incorrect utterances, our goal here is not to assess the correctness (fluency) of utterances. To study the quality and diversity of generated utterances, we adopt SMCalFlow and Taskmaster-3. [2] To create the GPT-3 prompts, we observe that considering only 10 examples in each prompt yields desirable performance.

Quality To evaluate the quality of the generated utterances, we conduct a user study asking participants to score the quality of each utterance from 0-100. We consider 100 instances for each baseline and assign 3 users for every sample (screenshot of user study in addition to examples of low and high quality original/generated instances is provided in Appendix). The result of our user study on quality evaluation is provided in Table 1a. As shown, the outputs of our GPT-3 prompting scheme are comparable with the original utterances (human-created), demonstrating their possible capability to replace humans in data collection.

Diversity We also investigate the diversity of generated utterances in comparison to original training samples. We generate 20k utterances and use two diversity measures, entropy and pair-wise maximum distance (details in Appendix). As seen in Table 1b, the generated utterances demonstrate a similar or better level of diversity.

5 Active Generation

Our goal is to generate examples that will be annotated by humans. To limit costs we would like to

[2]Background on these datasets are in the Appendix

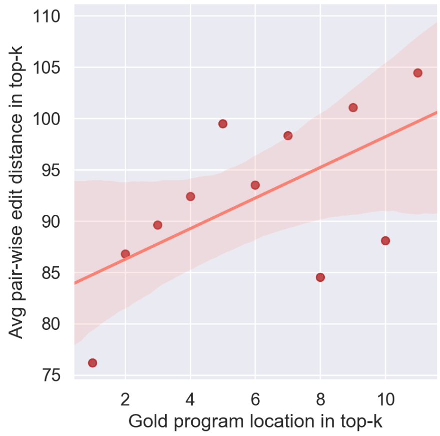


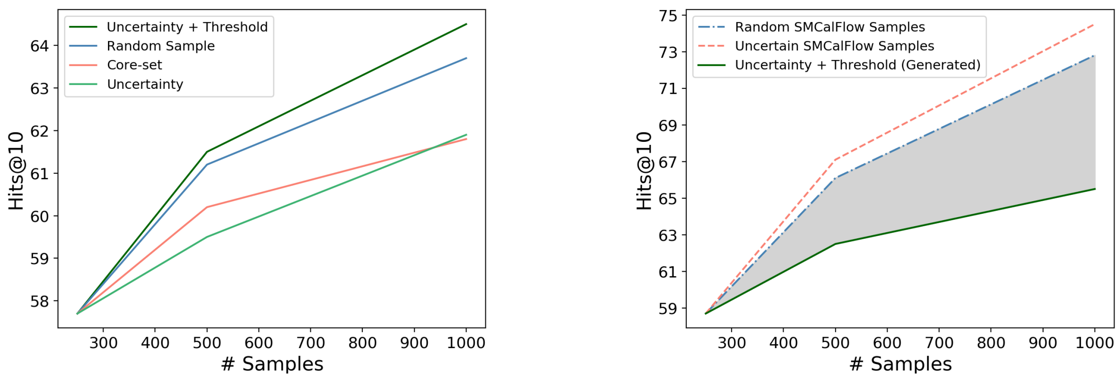
Figure 1: Approximating the parser confidence by investigating the correlation between average pairwise distance in top-k predicted programs and the accuracy.

minimize the number of such examples while maximizing their impact. Here we consider uncertainty-based methods as a mechanism for active filtering. We first propose and validate a black-box approximation of model's (parser) confidence on individual utterances. We then study the effect of different active learning strategies on parser performance.

In traditional active learning research, a dataset is pre-annotated and the goal is to identify the most informative subset. Annotations are hidden from the selection mechanism, but the impact on model performance can be studied automatically, by "revealing" the annotations performed before the study began. In active generation, this is not viable: examples are created without their annotations. We simulate a human annotator with a high resource system and discuss the trade-offs of this approach. We conduct a simulated study incorporating this approach on top of our pipeline, providing a lower bound on the parser performance.

Approximating Uncertainty We investigate our approximation of uncertainty by capturing the correlation between the average pairwise distance of the top-10 predictions and the placement of the gold program (gold annotation) in the top-10 predictions on the SMCalFlow dev set. We adopt Levenshtein distance (Miller et al., 2009) to measure the similarity between the predicted programs.[3] The correlation between the similarity of predictions and the accuracy is depicted in Figure 1. As it shows, there is a high correlation between the average pairwise similarity of predicted programs and model accuracy, thereby validating our conjecture.

[3]We investigate a variety of similarity metrics and Levenshtein distance shows the highest correlation with accuracy.



(a) Generated dialogues

(b) Generated vs SMCalFlow dialogues

Figure 2: Semantic parser performance by actively simulating dialogues in a low-resource setting.

Active Learning in Conversational Systems We study the potential impact of active sampling of generated utterances, by first subsampling from existing annotated data in the training set. We start with 250 random dialogues and increase this to 1000 using different active learning approaches, then simulate a human labeling by revealing the gold annotations. The top-1 and top-10 exact match parser accuracy on SMCalFlow dev is depicted in Table 2. Uncertainty approximation performs better than other baselines, outperforming the random sampling with 2-3% gain over accuracy. Moreover, the Core-set sampling also demonstrates a minor improvement over random sampling.

Active Dialogue Simulation To investigate the degree by which we can replace users in collecting data, we conducted a simulated study. The goal here is to see if our pipeline can help improve parser performance by generating informative dialogues in a limited label regime. Starting with 250 random dialogues from the SMCalFlow training set, we populate the training data using our proposed pipeline (examples of generated dialogues with different number of user turns is provided in Appendix). We simulate the user annotation process (writing executable programs for generated utterances) by incorporating a parser trained on all SMCalFlow training data and consider the top predicted program as the gold annotation. The result of top-10 exact match for our proposed pipeline with different filtering strategies is provided in Figure 2a. As it shows, our generated dialogues can help improve the performance by bootstrapping the parser. Moreover, both of our active sampling approaches perform worse than the random strategy. We suspect that since these sampling strategies choose the most uncertain instances, there is a higher probability that the high-resource parser mispredicts them, resulting in augmenting the training set with more mislabeled instances.[4]

To reduce the amount of mislabeled dialogues, we consider another baseline in which we first filter the dialogues that the parser is at a certain level of confidence in their prediction (using our approximation of uncertainty).[5] This baseline successfully outperforms the random sampling, setting a lower bound on the parser performance. We also compare the performance of parser trained with our generated dialogues versus SMCalFlow human-created dialogues in Figure 2b, demonstrating the room for improvement upon using human annotations instead of annotating based on high resource parser.

6 Conclusion

Collecting annotated dialogues constitutes a promising approach to train semantic parsers in conversational systems. However, gathering natural dialogues and annotating them is prohibitively expensive. In this work, we investigate whether we can automate this process by generating dialogues prompted via GPT-3. We first demonstrate that GPT-3 can generate high-quality and diverse utterances. Then providing an approximation for the parser uncertainty, we investigate the impact of active learning approaches. Finally, we evaluate our active dialogue simulation in improving the parser performance, motivating future work on *active generation* for bootstrapping semantic parsers.

[4] That is, we are actively selecting those cases where our proxy-annotator—the high resource parser—is most likely to get wrong, resulting in more mislabeled generated dialogues.

[5] We consider dialogues with less than 70 average pairwise Levenshtein distance on predicted programs. We tune this parameter on the dev set.

References

Anish Acharya, Suranjit Adhikari, Sanchit Agarwal, Vincent Auvray, Nehal Belgamwar, Arijit Biswas, Shubhra Chandra, Tagyoung Chung, Maryam Fazel-Zarandi, Raefer Gabriel, et al. 2021. Alexa conversations: An extensible data-driven approach for building task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 125–132.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, et al. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *56th Annual Meeting of the Association for Computational Linguistics*, pages 731–742. Association for Computational Linguistics.

Layla El Asri, Romain Laroche, and Olivier Pietquin. 2014. Task completion transfer learning for reward inference. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Kelvin Guu, Panupong Pasupat, Evan Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1051–1062.

Chien-Wei Lin, Vincent Auvray, Daniel Elkind, Arijit Biswas, Maryam Fazel-Zarandi, Nehal Belgamwar, Shubhra Chandra, Matt Zhao, Angeliki Metallinou, Tagyoung Chung, et al. 2020. Dialog simulation with realistic variations for training goal-oriented conversational systems. *arXiv preprint arXiv:2011.08243*.

Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.

Emmanouil Antonios Platanios, Adam Pauls, Subhro Roy, Yuchen Zhang, Alex Kyte, Alan Guo, Sam Thomson, Jayant Krishnamurthy, Jason Wolfe, Jacob Andreas, et al. 2021. Value-agnostic conversational semantic parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Pei-Hao Su, Paweł Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 147–157.

Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable dialogue systems and user simulators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 152–166.

Jason D Williams, Nobal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. In *Natural language dialog systems and intelligent assistants*, pages 1–13. Springer.

Ziyu Yao, Yiqi Tang, Wen-tau Yih, Huan Sun, and Yu Su. 2020. An imitation game for learning semantic parsers from user interaction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6883–6902, Online. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. 2019a. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1962–1979.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, et al. 2019b. Sparc: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523.

Luke S Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. *Association for Computing Machinery*.

A Conversational System Benchmarks

In this work, we adopt SMCalFlow (Andreas et al., 2020), a conversational system dataset consisting of around 40K natural dialogues regarding calendars, people, locations, and weather. An example of such conversation is as follows:

> **User** : What do I have scheduled for Friday?
> **Agent** : I didn't find any matching events on your calendar.
> **User** : Schedule time on Friday from 1pm to 3 pm to meditate.
> **Agent** : I've put that on your calendar.
> **User** : Mark my calendar as shopping from 4 to 6 pm on Friday.
> **Agent** : Is this good?
> **User** : Yes.

We also consider Taskmaster-3 (Byrne et al., 2019), a dataset consisting of 23,789 dialogues about movie ticketing, i.e., conversations in which users try to purchase tickets after deciding on the theater, time, movie name, number of tickets, and date. An example of movie ticketing conversation is as follows:

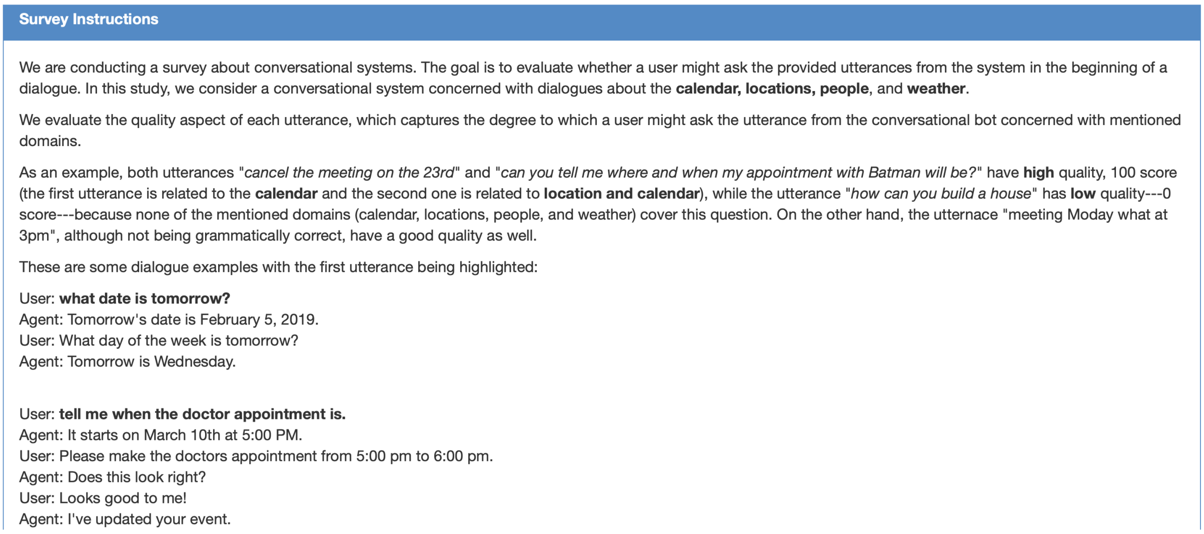
> **User** : I am looking for tickets tonight at the AMC Mountain 16.
> **Agent** : No problem. Is there a particular movie you are looking for?
> **User** : No Time To Die.
> **Agent** : OK. I see one action movie playing at AMC Mountain 16: No Time To Die. Remaining showtimes are 6:40pm and 9:10pm. Does any of those work?
> **User** : Yes, 9:10 is perfect.
> **Agent** : Great. And how many tickets.
> **User** : Just one.
> **Agent** : OK. Your tickets are purchased. Can I help with anything else?
> **User** : No thanks.
> **Agent** : OK. Enjoy your movie!

B Generated Samples

We provide the examples of low and high quality original/generated user utterances in Table 3. Moreover, examples of generated dialogues with different number of user turns is provided in Table 4.

C User Study

We provide the screenshot of our user study's instruction assessing the quality of generated and original utterances in Figure 3.



Survey Instructions

We are conducting a survey about conversational systems. The goal is to evaluate whether a user might ask the provided utterances from the system in the beginning of a dialogue. In this study, we consider a conversational system concerned with dialogues about the **calendar, locations, people**, and **weather**.

We evaluate the quality aspect of each utterance, which captures the degree to which a user might ask the utterance from the conversational bot concerned with mentioned domains.

As an example, both utterances "*cancel the meeting on the 23rd*" and "*can you tell me where and when my appointment with Batman will be?*" have **high** quality, 100 score (the first utterance is related to the **calendar** and the second one is related to **location and calendar**), while the utterance "*how can you build a house*" has **low** quality---0 score---because none of the mentioned domains (calendar, locations, people, and weather) cover this question. On the other hand, the utternace "meeting Moday what at 3pm", although not being grammatically correct, have a good quality as well.

These are some dialogue examples with the first utterance being highlighted:

User: **what date is tomorrow?**
Agent: Tomorrow's date is February 5, 2019.
User: What day of the week is tomorrow?
Agent: Tomorrow is Wednesday.

User: **tell me when the doctor appointment is.**
Agent: It starts on March 10th at 5:00 PM.
User: Please make the doctors appointment from 5:00 pm to 6:00 pm.
Agent: Does this look right?
User: Looks good to me!
Agent: I've updated your event.

Figure 3: Screenshot of user study instruction.

D Diversity Measures

To investigate the diversity of generated utterances in comparison to those taken from the original training set, we generate 20k utterances using our pipeline and compare them to 20k random first utterances from original training set. We consider two diversity metrics: 1) entropy, in which we first map the utterances (generated and original ones) into vector space using S-RoBERTa. Then, by partitioning the vector space into grids, we assign a probability to each grid by dividing the number of utterances that fall into that grid by the total number of utterances (20k). We then calculate the entropy using grids' probability. The higher entropy means that utterances are divided more uniformly into grids (space) thus providing more level of diversity. And 2) pair-wise maximum distance, in which we first map the utterances into vector space using S-RoBERTa, and then find the two data points that have the maximum distance from each other. The higher the maximum distance demonstrates the higher level of diversity.

		High-Quality	**Low-Quality**
SMCalFlow	Orig	Add a team meeting to my calendar for today at 5 pm.	i need any job.
		When is Kwanzaa.	Hello.
	Gen	Add Pick up Cake to my schedule at 2:30 today.	i am sick.
		find descriptions and url's of unread emails in my inbox.	Maybe.
Taskmaster	Orig	I'd like to see a move.	hello sir.
		Can you book two tickets for me to see Parasite tonight at AMC Norwalk 20 around 6PM?	hey there do you know where to this new movie where everyone gaga over villan thanos snap?
	Gen	I want to see some movies.	Hello.
		Could you show me the movie times for the Eureka Theater 10?	Are you a human?

Table 3: Examples of high and low quality original/generated utterances.

	Generated User Turns
1 turn	User: I need a meeting next Thursday at 3pm.
2 turns	User (1): Do I have any appointments today? User (2): Do I have any meeting with Chris today?
3 turns	User (1): How the weather going to be in San Francisco next weekend? User (2): Thanks! User (3): So it will be sunny?

Table 4: Random examples of generated dialogues with different number of user turns.