Active Dialogue Simulation in Conversational Systems

Anonymous ACL submission

Abstract

Semantic parsing helps conversational systems in satisfying users' requests through dialogues. To train these models, collecting annotated dialogues as a dataset is a very expensive and time-consuming process. In this paper, 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 the use of parser uncertainty on generated outputs as a selection criteria for annotation and contrast this with a strategy based on Core-sets. Our pipeline leads to more useful examples on average, motivating future work on *active generation* for bootstrapping semantic parsers.

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. We wish to guide synthetic dialogue generation to produce examples with most impact on semantic parser accuracy once annotated.

Building natural semantic parsing corpora requires (1) collecting examples of a user interacting with a software agent (i.e., user utterances in a form of a dialogue); and (2) annotating those utterances (i.e., tagging utterances with executable programs). In this work, we focus on step 1: 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) and Elgohary et al. (2021) exemplify strate-

gies 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 (using template) with 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; Zhao et al., 2019; Tseng et al., 2021). Such approaches have the benefit of complete data generation without a human annotation step, but have relied on template language generation, with dialogues created using logical forms (the target language of the parser), rather than true natural language.

In this work we are concerned with generation of natural language and adopt a different approach, directly incorporating large autoregressive language models (Radford et al., 2019; Brown et al., 2020) to simulate users based on dialague prompts. Moreover, concerning with efficiency of our pipeline, we utilize active learning schema (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 Active Simulated User

To generate examples of user interactions with a software agent, our framework consists of 3 steps: 1) *Generating user utterances* by prompting GPT-3, 2) *Actively filtering* generated utterances using active learning schema, and 3) *Generating dialogues* by iteratively prompting GPT-3 using filtered utterances and then sampling the most informative dialogues to a subset for manual annotation.

Step 1: Utterance Generation To generate dialogue utterances (turns), we start by generating the first user utterance. Incorporating GPT-3, we create prompts by randomly choosing first user utterances from the current training data (in a low-resource setting where you initially have a few hundred seed instances), and then asking GPT-3 to generate utterances similar to sampled instances in the prompt. Therefore, we construct a prompt like this:

> 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 4.1.

Step 2: Active Filtering We consider two approaches to select candidate generated utterances for annotation. We select based on: (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 obviously require a confidence-calibrated model: this is problematic if we wish to measure the relative level of uncertainty a parser may have in interpreting different synthetic user utterances. 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. Intuitively, the more distinct the examples produced by a model for a given utterance, the less confident the model is in its prediction. We investigate this empirically in Section 4.2. 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). We adopt the state-of-the-art semantic parser on SMCalFlow (Platanios et al., 2021) as our base parser throughout the paper.

Step 3: Dialogue Generation After filtering the generated utterances, to generate the whole dialogue, we first choose the number of turns that we plan to have for this dialogue uniformly from 1-3 turns. Then, we iteratively generate the next user utterance in the dialogue by creating a prompt containing the most similar dialogues (considering only user utterances) in the seed training data to our current generated dialogue (based on Levenshtein distance) with an equal or higher number of turns than our current turn in the generation. Then, concatenating our current generated dialogue to the prompt we ask GPT-3 to generate the next user turn. Assuming we want to generate the second user turn in a dialogue, we construct a prompt like this:

> Generate the next utterance in the dialogue.
> U1: When is the second event on my calendar for today?
> U2: When is my second event tomorrow?
> . . .
> U1: When is my sister's birthday? (this utterance is generated in step 1)
> U2:

	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.

To further improve the efficiency of our pipeline, after generating dialogues using GPT-3, we sample the most informative ones in an active setting by calculating a score for entire dialogues using the *max* of our utterance level score, whether uncertainty or diversity.[1]

4 Experiments

In this section, we first investigate the quality and diversity of our generated utterances prompted via GPT-3. Then, to incorporate uncertainty as a mechanism for active filtering, we first validate our approximation of model confidence, and then study the effect of different active learning samplings on the parser performance over SMCalFlow. Finally, we conduct a simulated study using generated dialogues with different active filtering methods, providing a lower bound on the parser performance incorporating our proposed pipeline.

4.1 Utility of Generated Utterances

The major challenge in utilizing GPT-3 generated dialogues/utterances to populate conversational system datasets is determining whether the generated instances are diverse and high quality enough (i.e., the probability that a user might bring up the generated utterances in a conversation about a specific domain). To study the quality of generated utterances using GPT-3, we adopt SMCalFlow (Andreas et al., 2020)—consisting of dialogues regarding calendars, people, locations, and weather—and Taskmaster-3 (Byrne et al., 2019)—consisting of dialogues about movie ticketing (more details in Appendix). 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 each utterance from 0-100, capturing the quality of each instance. 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, demonstrating their capability to replace humans in data collection.

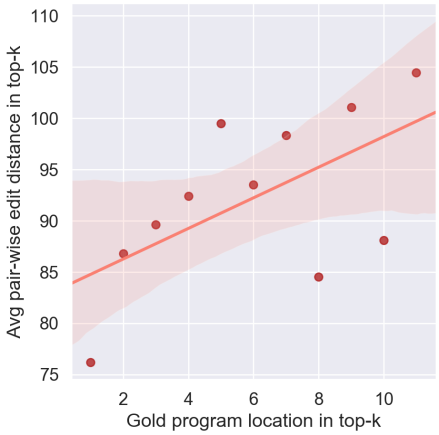


Figure 1: Approximating the parser confidence by investigating the correlation between average pairwise distance in top-k predicted programs and the accuracy.

Diversity We further investigate the diversity of generated utterances in comparison to original ones with over 20k random instances using two diversity measures. The results are presented in Table 1b. As shown, the generated utterances demonstrate a similar/better level of diversity in comparison to the original instances.

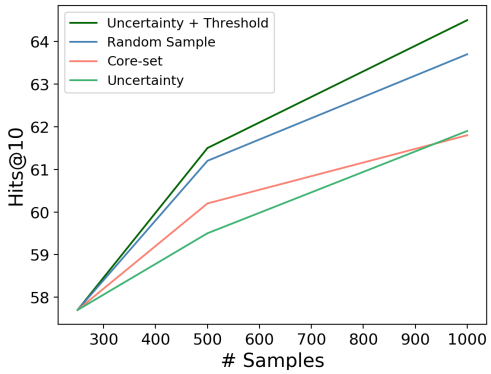
4.2 Active Generation

Approximating Uncertainty We investigate our approximation of uncertainty by capturing the correlation between the average pairwise distance between the top-10 predictions and the placement of the gold program in the top-10 predictions on SMCalFlow dev set. We adopt Levenshtein distance (Miller et al., 2009) to measure the similarity between the predicted programs[2]. The correlation between the similarity of predictions and the model 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.

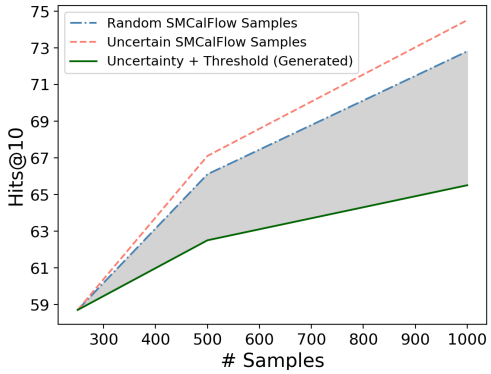
Active Learning in Conversational Systems In here, our goal is to evaluate the impact of active

[1] During development we confirmed that the *mean* utterance score of a dialogue was not effective.

[2] 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.

	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.

sampling (more specifically, our approximation of uncertainty) on the performance of the parser over SMCalFlow. We start with 250 random dialogues and increase the training size to 1000 instances (since we are more concerned with a limited labeled regime, we believe this is a reasonable interval) using different active learning approaches. The top-1 and top-10 exact match parser accuracy over SMCalFlow dev set is depicted in Table 2. As it shows, our 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 procedure, we conducted a simulated study. Starting with 250 random dialogues from the SMCalFlow training set, we start populating 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 by incorporating a parser trained on all SMCalFlow training data and consider the top predicted program from the parser as the gold annotation for generated utterances. The result of top-10 exact match for our proposed pipeline with different filtering strategies is provided in Figure 2a. As it shows, both of our active sampling approaches perform worse than the random strategy. We believe that this is because these methods choose the most uncertain instances, so there is a higher probability that the parser mispredicts them, resulting in augmenting more mislabeled samples into the training. To investigate this phenomenon, we consider another baseline in which we first filter the dialogues that the model is at a certain level of confidence in their prediction (we consider dialogues with less than 70 average pairwise Levenshtein distance on predicted programs. We tune this parameter on the dev set), to reduce the amount of mislabeled data. 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 dialogues in Figure 2b, demonstrating the room for improvement upon introducing human in the loop.

5 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 (Brown et al., 2020). 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 in the conversational system. Finally, we evaluate our active dialogue simulation in improving the parse performance, motivating future work on *active generation* for bootstrapping semantic parsers.

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*.

Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan. 2021. Nl-edit: Correcting semantic parse errors through natural language interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5599–5610.

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*.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. 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)*, pages 1208–1218.

		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.

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. 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.

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.

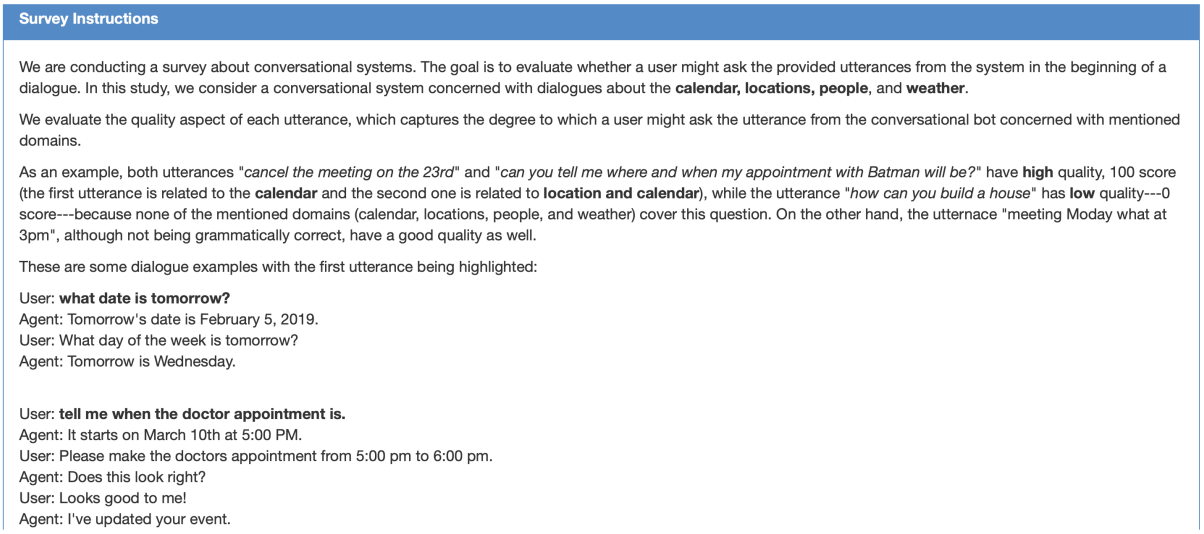


Figure 3: Screenshot of user study instruction.

	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.