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: When is my today event on calendar?
U2: When is my second event tomorrow?

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., $U_1$ 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).

### 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
<table>
<thead>
<tr>
<th></th>
<th>CalFlow</th>
<th>Taskmaster</th>
<th>Max-D</th>
<th>Ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orig</td>
<td>73.25</td>
<td>75.53</td>
<td>15.02</td>
<td>5.87</td>
</tr>
<tr>
<td>Gen</td>
<td>68.75</td>
<td>67.07</td>
<td>14.01</td>
<td>6.51</td>
</tr>
</tbody>
</table>

(a) Quality. (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.

<table>
<thead>
<tr>
<th></th>
<th>Hits@1</th>
<th>Hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random (250)</td>
<td>41.2</td>
<td>57.7</td>
</tr>
<tr>
<td>Random (1000)</td>
<td>53.9</td>
<td>71.8</td>
</tr>
<tr>
<td>Core-set (1000)</td>
<td>54.8</td>
<td>72.4</td>
</tr>
<tr>
<td>Uncertainty (1000)</td>
<td>56.2</td>
<td>73.3</td>
</tr>
</tbody>
</table>

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. 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 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 in top-k predicted programs and the accuracy.

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

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

⁴That 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.
⁵We consider dialogues with less than 70 average pairwise Levenshtein distance on predicted programs. We tune this parameter on the dev set.
References


Jason D Williams, Nobal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez,


## 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.
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.

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

Table 3: Examples of high and low quality original/generated utterances.
<table>
<thead>
<tr>
<th>Generated User Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 turn</strong></td>
</tr>
<tr>
<td>User: I need a meeting next Thursday at 3pm.</td>
</tr>
<tr>
<td><strong>2 turns</strong></td>
</tr>
<tr>
<td>User (1): Do I have any appointments today?</td>
</tr>
<tr>
<td>User (2): Do I have any meeting with Chris today?</td>
</tr>
<tr>
<td><strong>3 turns</strong></td>
</tr>
<tr>
<td>User (1): How the weather going to be in San Francisco next weekend?</td>
</tr>
<tr>
<td>User (2): Thanks!</td>
</tr>
<tr>
<td>User (3): So it will be sunny?</td>
</tr>
</tbody>
</table>

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