Active Dialogue Simulation in Conversational Systems

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-
genres 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 dialogue 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: |
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>CalFlow</th>
<th>Taskmaster</th>
<th>Max-D</th>
<th>Ent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orig</td>
<td>Gen</td>
<td>Orig</td>
<td>Gen</td>
</tr>
<tr>
<td></td>
<td>73.25</td>
<td>68.75</td>
<td>15.02</td>
<td>14.01</td>
</tr>
<tr>
<td></td>
<td>75.53</td>
<td>67.07</td>
<td>5.87</td>
<td>6.51</td>
</tr>
</tbody>
</table>
(a) Quality.
(b) Diversity, SMCalFlow.

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

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

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. 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 learning by informing the system of the most informative instances to prompt. To this end, 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.
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.

<table>
<thead>
<tr>
<th>Method</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>

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


<table>
<thead>
<tr>
<th></th>
<th>High-Quality</th>
<th>Low-Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Orig</strong></td>
<td>Add a team meeting to my calendar for today at 5 pm. When is Kwanzaa.</td>
<td>i need any job.</td>
</tr>
<tr>
<td><strong>SMCalFlow</strong></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.</td>
</tr>
<tr>
<td><strong>Taskmaster</strong></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>hey there do you know where to this new movie where everyone gaga over villan thanos snap?</td>
</tr>
<tr>
<td><strong>Gen</strong></td>
<td>I want to see some movies. Could you show me the movie times for the Eureka Theater 10?</td>
<td>Are you a human?</td>
</tr>
</tbody>
</table>

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.
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 utterance “meeting Monday 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.

<table>
<thead>
<tr>
<th>Generated User Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 turn User: I need a meeting next Thursday at 3pm.</td>
</tr>
<tr>
<td>2 turns User (1): Do I have any appointments today? User (2): Do I have any meeting with Chris today?</td>
</tr>
<tr>
<td>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?</td>
</tr>
</tbody>
</table>

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