How Would You Say It? Eliciting Lexically Diverse Data For Supervised Semantic Parsing

Abhilasha Ravichander, Thomas Manzini, Matthias Grabmair, Graham Neubig
Jonathan Francis, Eric Nyberg
Language Technologies Institute, Carnegie Mellon University

Motivation
Building dialogue interfaces for real-world scenarios often entails training semantic parsers starting from zero examples. How do we build datasets that capture the variety of ways in which users phrase their queries?

Introduction
Orienting a dialogue-capable intelligent system is accomplished by training its semantic parser with utterances that capture the nuances of the domain.

Previous work[1] proposes a methodology that generates example natural language utterances for logical forms, which are then paraphrased by crowdworkers.

Three main limitations:
• Seed utterances may induce bias towards the language of the canonical utterance, specifically with regards to lexical choice.
• Generic grammar suggested cannot be used to generate all the queries we may want to support in a new domain.
• No check on the correctness or naturalness of the canonical utterances themselves, which may not be logically plausible.

Method

Lexicon
- FOOD[bread] →
- FRIDGE[refrigerator] →
- FOOD_STATE[expired state] →
- FOOD_STATE[count] →

Grammar
- FRIDGE[x] → FRIDGE_NP[x]
- FOOD[x] → FOOD_NP[x]
- FD_NP[x] in the FRIDGE_NP[]
- what is the FOOD_STATE[] of the FD_SING[x]
- Q{x} → Q{None, None, x, checkState(+="s"), "root"[x]}

 Canonical & Logical Forms
what is the expired state of the

Crowdsourced Paraphrases
ROOT["(None, ‘refrigerator’, ‘bread’, ‘getFood’ checkStateExp=’expired state’)"

- "is the bread in the fridge go bad?"
- "is the bread in the refrigerator expired yet?"
- "is the bread in the fridge bad?"

• 195 Turker Sessions over three days.
• Each logical form shown to 5 Turkers for paraphrasing.
• Each Turker asked to enter a total of 60 paraphrases.
• 8294 unique paraphrases collected over 948 logical forms.

Results and Discussion

Table 1. Comparison of data creation methodology of [1] and this work.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Vocal Size</th>
<th>TTR</th>
<th>Lexical Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test (Wang et al., 2015)</td>
<td>291</td>
<td>0.44</td>
<td>5.50</td>
</tr>
<tr>
<td>Test Image (ours)</td>
<td>438</td>
<td>0.66</td>
<td>4.79</td>
</tr>
</tbody>
</table>

Table 3. Test accuracy results of different systems on the SMARTHOME dataset as compared to OVERNIGHT and GEO

<table>
<thead>
<tr>
<th>System</th>
<th>SMARTHOME</th>
<th>OVERNIGHT</th>
<th>GEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>18.0%</td>
<td>24.82%</td>
<td>40.7%</td>
</tr>
<tr>
<td>Neural Ranker</td>
<td>30.3%</td>
<td>41.91%</td>
<td>60.2%</td>
</tr>
<tr>
<td>Seq2Seq(2)</td>
<td>42.1%</td>
<td>75.8%</td>
<td>85.0%</td>
</tr>
</tbody>
</table>

Table 2. Number of word types in the language compared to the logical form. Larger ratio indicates more lexical diversity for the same complexity of the logical form.

Conclusion
• A mixture of text and images elicits more lexically diverse paraphrases from crowdworkers with limited loss of correctness.
• SMARTHOME dataset for semantic parsing.
• Domain, cardinality and complementary formulations also contribute to difficulty.

Most errors stem due to the following types of queries in SMARTHOME, which are not present in OVERNIGHT or GEO:
• Singular and plural forms (eg. radio/radios)
• Unseen semantically equivalent phrases (eg. Does Bob not have energy should be mapped to the logical form for Is Bob tired).
• Indirect phrases (eg. Do I need to change the lights in the living room not mapping to the logical form for living room lights not working correctly).
• Complementary Terms (on/off).

References

Acknowledgements
This research is supported by Robert Bosch LLC as part of a joint project with CMU on the development of a context-aware, dialog-capable personal assistant.

Dataset: https://github.com/oaqa/resources
E-mail: {aravicha, tmanzini}@cs.cmu.edu