Q&A for Wikidata

CS294S/W Project Pitch
Silei Xu
Wikidata.org

A large open-domain knowledge base with 90 million items, 8K properties

Welcome to Wikidata

the free knowledge base with 89,489,798 data items that anyone can edit.

Introduction • Project Chat • Community Portal • Help

Want to help translate? Translate the missing messages.
## Q&A on Wikidata

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Publisher</th>
<th>STOA</th>
<th>Dataset Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSQA</td>
<td>1.6 Million</td>
<td>AAAI 2018</td>
<td>0.71 (F1)</td>
<td>Train &amp; evaluate on synthetic data</td>
</tr>
<tr>
<td>LC-Quad 2.0</td>
<td>30K</td>
<td>ISWC 2019</td>
<td>-</td>
<td>Train &amp; evaluate on paraphrase data</td>
</tr>
<tr>
<td>KQA Pro</td>
<td>117K</td>
<td>Arxiv 2020</td>
<td>35%</td>
<td>Train &amp; evaluate on paraphrase data</td>
</tr>
<tr>
<td>Schema2QA</td>
<td>470K per domain</td>
<td>CIKM 2020</td>
<td>70%</td>
<td>Train on synthetic+paraphrase, evaluate on real questions</td>
</tr>
</tbody>
</table>
Current Status

- Homework: build a Q&A agent for one domain in Wikidata
- Can we extend this to a multi-domain Q&A agent over the entire Wikidata?
  - Extract useful information to generate the manifest and parameter values needed for data synthesis
  - Generate synthetic dataset for all domains
  - Avoid conflicts
Challenges

● **Scalability**
  ○ More than 80GB of data
  ○ Extract useful information to generate the manifest and parameter values needed for data synthesis
  ○ Generate synthetic dataset for all domains
  ○ Avoid conflicts

● **Representation**
  ○ ThingTalk: qualifiers, joins

● **Compositionality**
  ○ Impossible to train on all possible combinations, we need to generalize to unseen programs
  ○ Can we leverage other information such as types?
Roadmap

1. Download the wikidata dump and extract manifest (1~2 weeks)
2. Build a baseline semantic parser with current infrastructure (1~2 weeks)
3. Find out where it fails
4. Improve the quality of representation (manifest, ThingTalk) & synthetic data (3~4 weeks)
5. Beat the benchmarks and profit!
Auto-IoT
Semantic Parser for IoTs

CS294S/W Project Pitch
Silei Xu
Recap: AutoQA

- Automatically generate Q&A agents from schema
  - Learn how to ask questions using pre-trained language models
  - Synthesize large training set with 800 templates
Auto-IoT

Automatically generate virtual assistants to control IoTs from IoT function signatures

**IoT function signatures**

```python
action set_power(in req power: Enum(on,off))
```

- Turn on/off the light
- Switch on/off the light
- Lights up!
- Lights out!
- ...

We have function signatures for 20+ IoT devices in Thingpedia
Difference between Q&A and VA commands

- **Generic verb phrases vs domain-specific verb phrases**
  - Most of Q&A tables can use generic verb phrase to query: “search”, “find”, “show”, “get”, etc.
  - IoTs have different verb phrases: “turn on/off”, “lower the temperature”, “open the garage door”, “change the color to blue”, etc.

- **Personalization**
  - In Q&A, everyone queries the same database
  - For IoT devices, people may have different set of devices, and may name them differently.
Roadmap

1. Learn available commands for IoTs and analyze their sentence structure (~1 week)
2. Implement a similar algorithm as the one in AutoQA for Auto-IoT (~2 weeks)
3. Find out where it fails
4. Improve the algorithm & investigate new methodologies (3~4 weeks)
5. Get integrated with Almond + Home Assistant
6. Profit!