From Schema to Q&A Agents

Silei Xu
CS294S September 17, 2020

Joint work with Giovanni Campagna, Sina Semnani, Jian Li, and Monica S. Lam
Alexa: Handcode 1 question at a time

- get me an upscale restaurants
- What are the restaurants around here?
- What is the best restaurant?
- search for Chinese restaurants

100K Alexa skills
Sep 2019
Commercial Assistants

Alexa: Handcode 1 question at a time

get me an upscale restaurants
What are the restaurants around here?
What is the best restaurant?
search for Chinese restaurants

100K Alexa skills
Sep 2019

1.8 billion websites
Genie: Synthesize Question/Code from a Schema

**User**

**Schema**

<table>
<thead>
<tr>
<th>Name</th>
<th>Price</th>
<th>Cuisine</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Genie**

**800 Domain-Independent Templates**

- What is the <prop> of <subject>?
- What is the <subject>'s <prop>?
- get me an upscale restaurants
- What are the restaurants around here?
- What is the best restaurant?
- search for Chinese restaurants
- What is the best restaurant within 10 miles?
- Find restaurants that serve Chinese or Japanese food
- What is the best non-Chinese restaurant near here?
- Show me a cheap restaurant with 5-star review.
- Are there any restaurant with at least 4.5 stars?
- What is the phone number of Wendy’s?
- I’m looking for an Italian fine dining restaurant.
- Give me the best Italian restaurant.
- Find me the best restaurant with 500 or more reviews
- Show me some restaurant with less than 10 reviews
Outline

• Representing Questions in ThingTalk

• High-quality Low-cost Training Data Generation by Genie

• Apply Genie on the Web

• AutoQA: Automate Everything!
ThingTalk for Questions
ThingTalk for QA

`table [filter]`?

`@QA.restaurant(), geo == new Location("Stanford")` now => notify

Show me restaurants in Stanford
@QA.restaurant(), geo == new Location("Stanford")
&& servesCuisine =~ "Chinese"

Show me Chinese restaurants in Stanford
Show me Chinese restaurants in Stanford

```
sort fn asc|desc of table [, filter]?

@QA.restaurant(), geo == new Location("Stanford")
  && servesCuisine ~= "Chinese"
now => notify
=> notify
```

Stanford University
ThingTalk for QA

Show me top-rated Chinese restaurants in Stanford

```
sort fn asc|desc of table [, filter]?

sort aggregateRating.ratingValue desc of ( 
@QA.restaurant(), geo == new Location("Stanford") 
&& servesCuisine =~ "Chinese" ) 
now => 
=> notify
```

Stanford University
Show me top-rated Chinese restaurants in Stanford

```plaintext
show fn asc|desc of table [., filter]?[join table [., filter]?]*

sort aggregateRating.ratingValue desc of ( 
@QA.restaurant(), geo == new Location("Stanford")
&& servesCuisine =~ "Chinese" )

now => notify
```

Stanford University
ThingTalk for QA

<table>
<thead>
<tr>
<th>ThingTalk Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>sort fn asc</td>
</tr>
<tr>
<td>sort aggregateRating.ratingValue desc of (</td>
</tr>
<tr>
<td>@QA.restaurant(), geo == new Location(“Stanford”)</td>
</tr>
<tr>
<td>&amp;&amp; servesCuisine =~ “Chinese” )</td>
</tr>
<tr>
<td>join ( @QA.review(), in_array(id, review)</td>
</tr>
<tr>
<td>&amp;&amp; author = “bob” )</td>
</tr>
<tr>
<td>now =&gt; =&gt; notify</td>
</tr>
<tr>
<td>Show me top-rated Chinese restaurants in Stanford</td>
</tr>
<tr>
<td>reviewed by Bob</td>
</tr>
</tbody>
</table>

Stanford University
ThingTalk for QA

\[
\text{sort } fn \text{ asc|desc of } \text{ table [, filter]?} \left[ \text{join table [, filter]?} \right]^* \\
\text{[fn}^+ \text{ of}]? \text{ table [, filter]?} \\
\text{aggregate min|max|sum|avg|count fn of } \text{ table [, filter]?}
\]
What is the top-rated Chinese restaurant in Palo Alto?

```
now =>
  sort aggregateRating.ratingValue desc of (@QA.restaurant(),
  geo == new MakeLocation("Stanford")
  && servesCuisine =~ "Chinese")
=> notify;
```
**Synthesizing Training Data with Templates**

- **Templates**: Map natural language to database operators

<table>
<thead>
<tr>
<th>DB Operator</th>
<th>Natural Language</th>
<th>Template</th>
<th>ThingTalk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>restaurants with rating equal to 4 restaurants with rating greater than 4 restaurants with rating less than 4 ...</td>
<td>&lt;table&gt; with &lt;property&gt; equal to &lt;value&gt; &lt;table&gt; with &lt;property&gt; greater than &lt;value&gt; &lt;table&gt; with &lt;property&gt; less than &lt;value&gt;</td>
<td>table, property == value table, property &gt;= value table, property &lt;= value ...</td>
</tr>
<tr>
<td>Projection</td>
<td>rating of restaurant</td>
<td>&lt;property&gt; of &lt;table&gt;</td>
<td>[property] of table</td>
</tr>
<tr>
<td>Aggregation</td>
<td>the number of restaurants</td>
<td>the number of &lt;table&gt;</td>
<td>aggregate count of table</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **Generate natural language and ThingTalk pairs**
Discussion

Why this won’t work?
Variety in Natural Language

• Fact: “Dr. Smith is Ann’s doctor”

<table>
<thead>
<tr>
<th>Relation</th>
<th>Unknown: Ann</th>
<th>Unknown: Dr. Smith</th>
<th>Part-of-Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor</td>
<td>Who has Dr. Smith as a doctor?</td>
<td>Who does Ann have as a doctor?</td>
<td>Noun (has ...)</td>
</tr>
<tr>
<td></td>
<td>Who is Dr. Smith a doctor of?</td>
<td>Who is a doctor of Ann?</td>
<td>Noun (is ...)</td>
</tr>
<tr>
<td></td>
<td>Whom does Dr. Smith treat?</td>
<td>Who treats Ann?</td>
<td>Active verb</td>
</tr>
<tr>
<td></td>
<td>Who is treated by Dr. Smith?</td>
<td>By whom is Ann treated?</td>
<td>Passive verb</td>
</tr>
<tr>
<td>Patient</td>
<td>Who does Dr. Smith have as a patient?</td>
<td>Who has Ann as a patient?</td>
<td>Noun (has ...)</td>
</tr>
<tr>
<td></td>
<td>Who is a patient of Dr. Smith?</td>
<td>Who is Ann a patient of?</td>
<td>Noun (is ...)</td>
</tr>
<tr>
<td></td>
<td>Who consults with Dr. Smith?</td>
<td>With whom does Ann consult?</td>
<td>Active verb</td>
</tr>
<tr>
<td></td>
<td>By whom is Dr. Smith consulted?</td>
<td>Who is consulted by Ann?</td>
<td>Passive verb</td>
</tr>
</tbody>
</table>

Previous work: train with paraphrase data based on synthesized sentences
Natural Language Annotations

- POS-based annotation for each property

<table>
<thead>
<tr>
<th>POS</th>
<th>People: worksFor</th>
<th>Restaurants: servesCuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active verb</td>
<td>works for &lt;value&gt;</td>
<td>serves &lt;value&gt; cuisine, offer &lt;value&gt; food</td>
</tr>
<tr>
<td>Passive verb</td>
<td>employed by &lt;value&gt;</td>
<td>-</td>
</tr>
<tr>
<td>Is-a Noun</td>
<td>an employer of &lt;value&gt;</td>
<td>-</td>
</tr>
<tr>
<td>has-a Noun</td>
<td>employee &lt;value&gt;</td>
<td>&lt;value&gt; food, &lt;value&gt; cuisine</td>
</tr>
<tr>
<td>Adjective</td>
<td>-</td>
<td>&lt;value&gt;</td>
</tr>
<tr>
<td>Prepositional</td>
<td>from &lt;value&gt;</td>
<td>-</td>
</tr>
</tbody>
</table>
Domain-Independent Templates

- A comprehensive set of 800 templates that captures:
  - Different parts of speech
    ```
    now => QA.restaurant(),
    servesCuisine =~ "Chinese" => notify;
    ```
    Show me <table> that <verb>.  
    Show me <table> with <noun>.  
    Show me <adjective> <table>.  
    Show me restaurants that serve Chinese cuisine.  
    Show me restaurants with Chinese food.  
    Show me Chinese restaurants.
  - Connectives
    Show me restaurant that serve Chinese cuisine and with more than 100 reviews.  
    Show me restaurant with Chinese food and at least 100 reviews.  
    Show me Chinese restaurant that have more than 100 reviews.
  - Different types
    when does the restaurant open?  
    who owns the restaurant?  
    how far is the restaurant?
Genie Pipeline

Natural Language Annotations

- cuisine of the restaurant
- restaurant’s cuisine
- cuisine served by the restaurant

Schema

<table>
<thead>
<tr>
<th>Name</th>
<th>Price</th>
<th>Cuisine</th>
<th>…</th>
</tr>
</thead>
</table>

ThingsTalk Grammar

Domain-Independent Templates

- What is the <prop> of <table>?
- What is the <table>’s <prop>?

Synthesize sentence/code pairs

Paraphrase

Parameter & data augmentation

iterate

Training Data

iterate

Natural language

Q&A Agent

ThingTalk
BERT-LSTM Neural Model
Applying Genie to the Web
How do we scale to the web?

• The web has a schema: **Schema.org**
  • Structure data to mark up web pages
  • Mainly used by search engines
  • It covers many domains, including restaurants, hotels, people, recipes, products, news ...

40% of the websites use it!

```html
<script type="application/ld+json">
{
  @type: "restaurant",
  name: "The French Laundry",
  servesCuisine: "French",
  aggregateRating: {
    @type: "AggregateRating",
    reviewCount: 2527,
    ratingValue: 4.5
  }
}
...
</script>
```

Schema.org markup on Yelp
### Experiment domains

- 5 domains: restaurant, people, movie, book, and music

<table>
<thead>
<tr>
<th></th>
<th>Restaurant</th>
<th>People</th>
<th>Movie</th>
<th>Book</th>
<th>Music</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website</td>
<td>Yelp</td>
<td>LinkedIn</td>
<td>IMDb</td>
<td>Goodreads</td>
<td>Last.fm</td>
<td>-</td>
</tr>
<tr>
<td># of properties</td>
<td>25</td>
<td>13</td>
<td>16</td>
<td>15</td>
<td>19</td>
<td>17.6</td>
</tr>
<tr>
<td># of annotations</td>
<td>122</td>
<td>95</td>
<td>111</td>
<td>96</td>
<td>103</td>
<td>105.4</td>
</tr>
<tr>
<td>Synthesized</td>
<td>270,081</td>
<td>270,081</td>
<td>270,081</td>
<td>270,081</td>
<td>270,081</td>
<td>270,081</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>6,419</td>
<td>7,108</td>
<td>3,774</td>
<td>3,941</td>
<td>3,626</td>
<td>4,973.6</td>
</tr>
<tr>
<td><strong>Total (augmented)</strong></td>
<td><strong>508,101</strong></td>
<td><strong>614,841</strong></td>
<td><strong>405,241</strong></td>
<td><strong>410,141</strong></td>
<td><strong>425,041</strong></td>
<td><strong>472,673</strong></td>
</tr>
</tbody>
</table>
Evaluation Data Collection

- Evaluating on paraphrase data is misleading!
- Evaluate on a challenging realistic dataset
Evaluation Data Collection

- Evaluating on paraphrase data is misleading!
- Evaluate on a challenging realistic dataset

<table>
<thead>
<tr>
<th></th>
<th>Restaurant</th>
<th>People</th>
<th>Movie</th>
<th>Book</th>
<th>Music</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>1 property</td>
<td>221</td>
<td>127</td>
<td>140</td>
<td>107</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>2 properties</td>
<td>219</td>
<td>346</td>
<td>226</td>
<td>222</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>3+ properties</td>
<td>88</td>
<td>26</td>
<td>23</td>
<td>33</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>528</td>
<td>499</td>
<td>389</td>
<td>362</td>
<td>326</td>
</tr>
<tr>
<td>Test</td>
<td>1 property</td>
<td>200</td>
<td>232</td>
<td>130</td>
<td>114</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>2 properties</td>
<td>245</td>
<td>257</td>
<td>264</td>
<td>241</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>3+ properties</td>
<td>79</td>
<td>11</td>
<td>19</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>524</td>
<td>500</td>
<td>413</td>
<td>410</td>
<td>288</td>
</tr>
</tbody>
</table>

- Over 2/3 of questions have 2+ properties
- Contains unseen values
Experimental Results

Query Accuracy on Test Set

Restaurants | People | Movies | Books | Music | Average

- 1 property
- 2 properties
- 3+ properties
- Overall

Stanford University
Experimental Results (Synthetic Only)

Query Accuracy with Models Trained with Only Synthetic Data

- Restaurants
- People
- Movies
- Books
- Music
- Average

Stanford University
Comparison with Commercial Assistants

Genie vs Commercial Assistants on Restaurant Domain

Stanford University
## Example Questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Siri</th>
<th>Google</th>
<th>Alexa</th>
<th>Genie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show restaurants near Stanford rated higher than 4.5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Show me restaurants rated at least 4 stars with at least 100 reviews</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>What is the highest rated Chinese restaurants in Hawaii?</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>How far is the closest 4 star and above restaurant?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Find a W3C employee that went to Oxford</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Who worked for both Google and Amazon?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Who graduated from Stanford and won a Nobel prize?</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Who worked for at least 3 companies?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Show me hotels with checkout time later than 12PM</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Which hotel has a swimming pool in this area?</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
</tbody>
</table>
Evaluate on Common Questions

restaurant

name
cuisine
address
rating
reviews
...

questions

annotate

Stanford University
Comparison with Commercial Assistants on Common Questions

Genie vs Commercial Assistants on Restaurant Domain

- Siri
- Google Assistant
- Alexa
- Genie

1 property | 2 properties | 3+ properties | Overall

Stanford University
Discussions

Why do commercial assistants do a poor job on the first task but do a much better job in the second?
Discussions

• Why do commercial assistants do a better job in the second experiment?
  • they are tuned for common questions
  • they do a great job on recognizing common named entities
  • they can answer question correctly even with limited understanding of the question

• Why do commercial assistant do a poor job in the first experiment?
  • they are not tuned for complex long-tail questions
  • they don’t even include some of the less-common properties (e.g., review count)
  • they do a poor job on numeric comparison
Error Analysis

• 50% of the errors are due to named entity recognition
  • work in progress (potential class project)

• 14% of the error can potentially be solvable with new templates
  • E.g., two fields with the same value: “movies produced and directed by Steven Spielberg”

• If we fix these two, we can get close to 90%!

• Others: typos, joins operators
Can We Do Better?
Manual Steps in Genie Pipeline

• Natural language annotations
  • We ask developers to provide natural language annotations, and it takes a few iterations to get a good quality set of annotations

• Paraphrase
  • We ask crowd workers to manually paraphrase synthetic sentences
  • We can only do this for a small sample of synthetic because of cost

• Can we replace them with something automatic?
Automatic NL Annotation Generation

- Generate context-aware synonyms by a language model

  A Sample Sentence
  Automatically Constructed based on property name

  | Show me restaurants with Italian cuisine. |
  | BERT (pretrained) |
  | Show me restaurants with Italian dishes. |
  | Show me restaurants with Italian food. |
  | Show me restaurants with Italian menu. |
  | ... |

  Generate Context-aware Synonyms

  Templatize
  noun: “# cuisine | dishes | menu ... ”
Automatic NL Annotation Generation (cont.)

- Predict adjective qualifiers by a language model

- Construct a sample sentence with mask

- BERT (pretrained)

- Show me a good restaurant.
- Show me a Chinese restaurant.
- ...

- Look up predicted words in property value sets

- servesCuisine – adjective: “#”
- ...

- Add adjective annotation to found properties
Automatic Paraphrasing

Paraphrase dataset

GPT-2 (Pretrained)

Synthetic Training Examples

Show me restaurants with Chinese cuisine.

Fine-tune

GPT-2 Paraphraser

What is a restaurant that is Chinese?
Give me Chinese dining places.
Show me top rated Chinese restaurants.
...

Inference

Model Trained w/ Synthetic data

Filter paraphrases that do not preserve meaning

Paraphrased Examples

What is a restaurant that is Chinese?
Give me Chinese dining places.
...

Stanford University
Experimental Result

Query Accuracy on Test Set*

<table>
<thead>
<tr>
<th>Domain</th>
<th>Manual Annotation + Manual Paraphrase</th>
<th>Auto Annotation + Auto Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>90%</td>
<td>85%</td>
</tr>
<tr>
<td>People</td>
<td>95%</td>
<td>92%</td>
</tr>
<tr>
<td>Movies</td>
<td>80%</td>
<td>75%</td>
</tr>
<tr>
<td>Books</td>
<td>85%</td>
<td>82%</td>
</tr>
<tr>
<td>Hotels</td>
<td>80%</td>
<td>75%</td>
</tr>
<tr>
<td>Average</td>
<td>85%</td>
<td>82%</td>
</tr>
</tbody>
</table>

* evaluated on an older version of the dataset with fewer properties per domain
Thank you!