**Motivation**

**Goal:** answer factual questions

- Greece held its last Olympics in which year?

The system is trained with many example question-answer pairs

**Desiderata:**

1. **Breadth:** cover a wide range of knowledge domains
2. **Depth:** handle complex language and different operations

**Task Description & Related Work**

- **GeoQuery:** fixed domain (US geography) / focuses on compositionality: What states border states that border states that border Texas?

- **Freebase QA:** increases breadth to knowledge bases (e.g., Freebase): In which comic book issue did Kitty Pryde first appear?

**Information Retrieval:** Web-level coverage but less complexity: Stanford CS faculty

**Our Task:** complex questions on semi-structured tables from the Web

**Input:** a table \( t \) and a question \( x \)

**Output:** an answer \( y \)

- \( x \): Greece held its last Olympics in which year?
  - \( y \): 2004

- \( x \): In which city was the first time with at least 20 nations?
  - \( y \): Paris

- \( x \): How many more participants were in 1900 than the first year?
  - \( y \): 10

**WikiTableQuestions** (2108 tables, 22033 questions)

- **3929 unique column headers = relations** (GeoQuery: 30 relations, Free917 on Freebase: 635 relations)

- **Breadth:** Freebase can answer only \( \approx 20\% \) of the questions

- **Tables in test data are not seen during training:** Must learn to generalize to open-ended table schemata

- **Depth:** crowd-sourced complex questions (\( \approx 10 \) words per question)

**Semantic Parsing Approach**

Semantic parsing: use a latent logical form \( z \) as:

- an expressive semantic representation of the question \( x \), and
- a query that can be executed on the table to get an answer \( y \)

Note that \( z \) is not given in training data

**Dataset**

<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Country</th>
<th>Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Greece</td>
<td>14</td>
</tr>
<tr>
<td>1900</td>
<td>Paris</td>
<td>France</td>
<td>24</td>
</tr>
<tr>
<td>1904</td>
<td>St. Louis</td>
<td>USA</td>
<td>12</td>
</tr>
<tr>
<td>2004</td>
<td>Athens</td>
<td>Greece</td>
<td>201</td>
</tr>
<tr>
<td>2008</td>
<td>Beijing</td>
<td>China</td>
<td>204</td>
</tr>
<tr>
<td>2012</td>
<td>London</td>
<td>UK</td>
<td>204</td>
</tr>
</tbody>
</table>

**Challenges**

**With increased breadth:**

- Unlike knowledge bases, tables have no fixed schema
- Don’t know which phrase maps to unseen relations

**Solution:** floating cells

- Allow generated predicates to not anchor to any phrase

**With increased depth:**

- Must handle more operations / larger parse trees
- Exponentially many possible formulas

**Solution:** generic recursive rules with type constraints

**Results**

**Example correct answer:**

How many districts have a population density of at least 1000?

(Information retrieval alone will not be able to answer the question)

**Example errors:**

- Fail to anchor entities: How many Mexican swimmers ...? (table has "Mexico")
- Must interpret table cell content: How long is the program? (table has "2pm-3pm")
- Phrase and relation are obliquely related: Which airplane ...? (table has column "Model")