TSQA: Tabular Scenario Based Question Answering

Xiao Li, Yawei Sun, Gong Cheng
State Key Laboratory for Novel Software Technology, Nanjing University, China
{xiaoli.nju, ywsun}@mail.nju.edu.cn, gcheng@nju.edu.cn
Outline

- Motivation
- Task
- Dataset
- Approach
- Experiments
- Error analysis & Future work
- Conclusion
Motivation

Existing Table-based QA can be directly answered solely on the basis of tabular data, based on semantic parsing or retrieval, with a certain gap from real life problem.

### Scenario

The educational level on products (ELP) refers to the average educational level of workers that produce products. To compute it, educational stages are quantified as follows: 1 for lower secondary or below, 3 for higher secondary, 5 for junior college, 7 for undergraduate, and 9 for graduate or above. The following table shows the development of ELP in China. Please read the table and answer the question.

<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>

*Question:* The main reason for the change of ELP after year 2000 is

(A) improvement in the quality of education  
(B) industrial upgrading  
(C) rural labor migration  
(D) counterurbanization

A sample of existing Table-based QA

Pasupat et al. 2015, Compositional Semantic Parsing on Semi-Structured Tables
Task

TSQA (Tabular Scenario Based Question Answering):
Answering questions which contextualized by a scenario consisting of a textual passage and a set of tables.

Given:

- passage scenario $P$
- a set of table $T$
- a question $Q$
- a set of candidate options $O$

Goal:

select an option from $O$ as the answer to $Q$ contextualized by $< P, T >$
Dataset

Dataset Construction of GeoTSQA

- Collecting Questions
  - Domain: geography problems from China’s high-school exams
  - Source: GeoSQA (Huang et al. 2019) and Web
- Identifying Tables
  - Tables, or charts that can be straightforwardly converted to tables (e.g., histograms, line charts)
- Extracting Tables
  - We recruited 15 annotators to annotate table from the raw images
- Filtering Questions
  - Filtered out questions that can be answered without using any table to ensure every question is contextualized by a tabular scenario
Dataset

• Statistics about GeoTSQA

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenarios</td>
<td>556</td>
<td></td>
</tr>
<tr>
<td>Chinese characters per passage</td>
<td>52.42</td>
<td>±32.99</td>
</tr>
<tr>
<td>Tables per scenario</td>
<td>1.58</td>
<td>±0.93</td>
</tr>
<tr>
<td>Cells per table</td>
<td>26.98</td>
<td>±17.51</td>
</tr>
<tr>
<td>Questions</td>
<td>1,012</td>
<td></td>
</tr>
<tr>
<td>Chinese characters per question</td>
<td>44.02</td>
<td>±15.89</td>
</tr>
</tbody>
</table>

• Features
  • Most of tables (96%) in GeoTSQA is numerical
  • Question can be only answered from the fusing of table, question and scenarios
Approach

• Why do existing table-based QA models fail in GeoTSQA?
  • TSQA requires synthesizing data in multiple cells and combining tables with texts and domain knowledge to infer answers
  • Semantic parsing or retrieval based models do not work as answers do not appear in tables in most cases

• Can we extract information in table that needed for QA and then machine reading comprehension?
Approach

- Overview of our approach: a two stage model
Approach

- Template based sentence generation

<table>
<thead>
<tr>
<th>Year</th>
<th>ELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1998</td>
<td>2.465</td>
</tr>
<tr>
<td>1999</td>
<td>2.476</td>
</tr>
<tr>
<td>2000</td>
<td>2.504</td>
</tr>
<tr>
<td>2001</td>
<td>2.49</td>
</tr>
<tr>
<td>2002</td>
<td>2.482</td>
</tr>
<tr>
<td>2003</td>
<td>2.473</td>
</tr>
</tbody>
</table>

**Table**

**Templates**

- Extremum
- Special values
- Comparison with average
- Monotonicity
- Trend
- Range comparison

**Sentence**

- ELP reaches a maximum of 2.504 at Year 2000
- ELP reaches a minimum of ...
- ELP at Year 2000 is 2.504
- ...
- ELP is relatively large between Year 2000 and 2002
- ...
- ELP decreases between Year 2000 and 2003
- ...
- ELP generally increases and then decreases
- ...
- ...
Approach

- Not all table information is useful for QA.
- Two level sentence ranking: sentence-level ranking and template-level ranking
- Then feed the MRC model with top-k table sentences to select answer
Experiments

- Experiments on GeoTSQA
- We extend sota table-to-text method with MRC model
- Table-to-text Methods
  - Table-Infusing (Chen et al., 2020): encode table cells and then generate
  - GPT-Linearization (Chen et al., 2020): describe all cells of table and feed the GPT-2 with table passage to generate
  - Coarse-to-Fine (Chen et al., 2020): first generate template and then fill out template
  - Linearization: describe all cells of table into a passage
  - Templation: concatenate all sentences of $S$ into a table passage
Experiments

Metric: Accuracy, i.e., the proportion of correctly answered questions

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field-Infusing</td>
<td>0.353</td>
</tr>
<tr>
<td>GPT-Linearization</td>
<td>0.370</td>
</tr>
<tr>
<td>Coarse-to-Fine</td>
<td>0.367</td>
</tr>
<tr>
<td>GPT-Linearization$^+$</td>
<td>0.348</td>
</tr>
<tr>
<td>Coarse-to-Fine$^+$</td>
<td>0.359</td>
</tr>
<tr>
<td>Linearization</td>
<td>0.235</td>
</tr>
<tr>
<td>Temptation</td>
<td>0.243</td>
</tr>
<tr>
<td>TTGen</td>
<td>0.397</td>
</tr>
<tr>
<td>Gold-Standard Sentence</td>
<td>0.418</td>
</tr>
</tbody>
</table>

Table 4: Accuracy of TSQA. We mark the results of baselines that are significantly lower than TTGen under $p < 0.01$ (*) or $p < 0.05$ (°).
Experiments

Varying k (top-k ranked sentences are used in MRC)

<table>
<thead>
<tr>
<th></th>
<th>k = 1</th>
<th>k = 2</th>
<th>k = 3</th>
<th>k = 4</th>
<th>k = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.390</td>
<td>0.397</td>
<td>0.352</td>
<td>0.343</td>
<td>0.330</td>
</tr>
</tbody>
</table>

Table 5: Accuracy of TSQA by varying k in TTGen.

Ablation study

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTGen</td>
<td>0.397</td>
</tr>
<tr>
<td>TTGen w/o tabular data</td>
<td>0.372</td>
</tr>
<tr>
<td>TTGen w/o domain knowledge</td>
<td>0.380</td>
</tr>
</tbody>
</table>

Table 6: Accuracy of TSQA (ablation study).
Error analysis & Future work

Error analysis:
- Lack of external knowledge base (76%)
- Weak Reasoning Capabilities (62%)
- Sentence Ranking Error (54%)

Future work:
- More powerful MRC model which is able to fuse scenario passage, table sentences and domain knowledge
- More powerful reasoning module
- More powerful question-aware table-to-text generator
Conclusion

• We constructed the first TSQA dataset: GeoTSQA.
• We use six templates encapsulating predefined operations for synthesizing tabular and then select the most useful sentences.
• We use MRC model, fusing scenario passage, question, table information and domain knowledge to select answer.

Resources: https://github.com/nju-websoft/TSQA