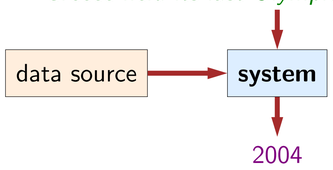
Compositional Semantic Parsing on Semi-Structured Tables

Panupong Pasupat Percy Liang

Motivation

Goal: answer factual questions

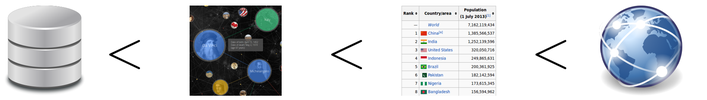
Greece held its last Olympics in which year?



The system is trained with many example question-answer pairs

Desiderata:

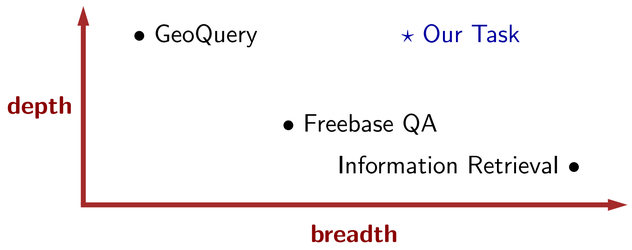
1. Breadth: cover a wide range of knowledge domains



2. Depth: handle complex language and different operations

Where was Barack Obama born? < **How many** *presidents* **after** *Abraham Lincoln were born in Ohio?*

Task Description & Related Work



GeoQuery: fixed domain (US geography) / focuses on compositionality:

What states border states that 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?

but the questions tend to be simpler factoid questions

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

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

Modified from en.wikipedia.org/wiki/Summer_Olympic_Games

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

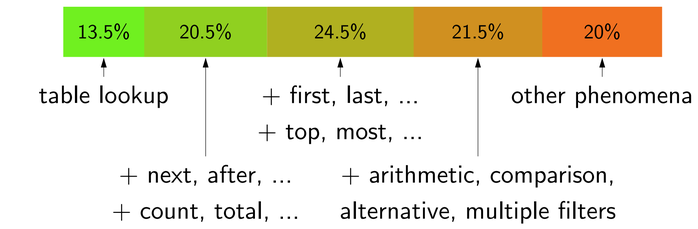
x: **How many more** *participants were in 1900 than the* **first** *year?*

y: 10

Dataset

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

- **3929** unique column headers = relations
  (GeoQuery: 30 relations, Free917 on Freebase: 635 relations)
- Breadth: Freebase can answer only ≈ 20% of the questions
- Tables in test data are not seen during training
  → Must learn to generalize to open-ended table schemata
- Depth: crowdsourced complex questions (≈ **10** words per question)

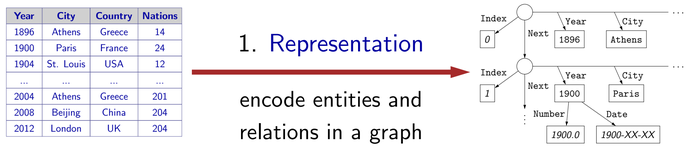


Semantic Parsing Approach

Semantic parsing: use a latent logical form z as:

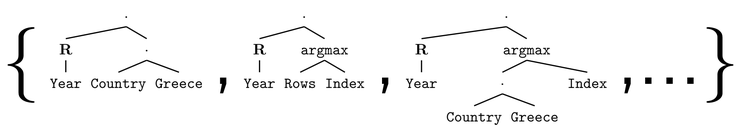
- an expressive semantic representation of the question x, and
- a query that can be executed on the table to get an answer y

Note that z is not given in training data

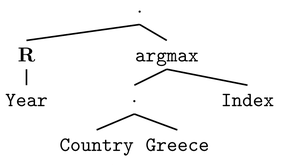


x: *Greece held its last Olympics in which year?*

2. Generation
parse x into candidate logical forms



3. Ranking
use a statistical model to score candidates and choose the highest-scoring formula z



z: $\mathbf{R}[\texttt{Year}].\texttt{argmax}(\texttt{Country.Greece}, \texttt{Index})$

4. Execution
execute z on the graph

y: 2004

Challenges

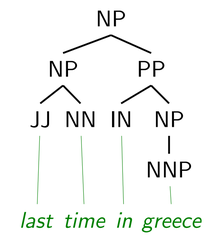
With increased breadth:

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

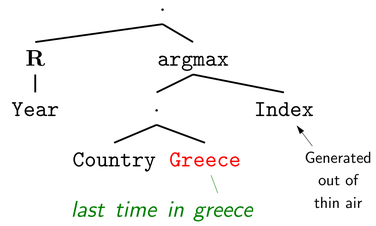
Solution: floating cells

- Allow generated predicates to not anchor to any phrase

anchored parsing (syntactic parse)



parsing with floating cells



- Let the statistical model relate phrases to formula predicates

With increased depth:

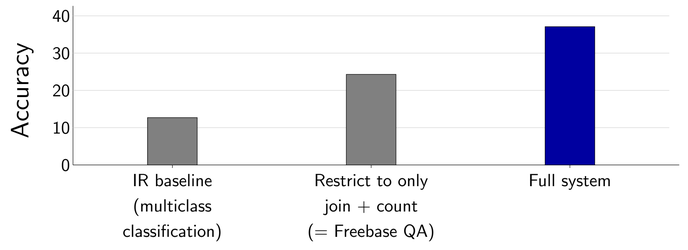
- Must handle more operations / larger parse trees
- Exponentially many possible formulas

Solution: generic recursive rules with type constraints

Records	→	Relation . Values
Values	→	$\mathbf{R}$[Relation] . Records
Records	→	argmax(Records, Index)
Records	→	argmax(Records, Relation)

- Prevent combinatorial explosion by pruning malformed or redundant formulas based on 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"**)