Datasets for Semantic Parsing

Alane Suhr
Cornell University

Research Focuses

- Generating executable representations
- Understanding in a situated environment
- Generalizing to broad domains
- Sequential language understanding

Executable Representations

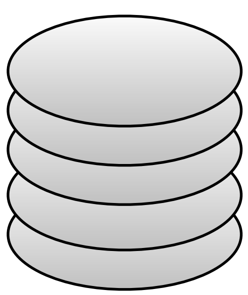
- **Goal:** generate compositional, executable, formal representation, e.g. code
- Formal representation requires following strict constraints:
  - Syntax (e.g., correct number of parentheses)
  - Semantics (e.g., calling function with the right type of arguments)

Question:

What is the largest city in Hawaii?

Logical form:

```
answer(A, largest(A, city(A), loc(A, B),
    const(B, stateid(hawaii))))
```



Database

Answer:

Honolulu

[Zelle and Mooney 1996; Tang and Mooney 2001]

Geoquery and Jobs

- Natural language interfaces to a database
- First corpus-based methods for learning an interface
- Early datasets for semantic parsing
- **Data collection:** in class exercise to write queries paired with sentences; later: web interface
- **Size:** <1,000 examples, relatively small database

John Zelle and Raymond Mooney. 1996. Learning to parse database queries using inductive logic programming. In AAAI.

Lappoon Tang and Raymond Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In ECML.

Training Data

(sentence, implementation) pairs

What is the largest city in Hawaii? → `answer(A, largest(A, city(A), loc(A, B), const(B, stateid(hawaii))))`

What is the capital of California? → `answer(A, capital(A), loc(A, B), const(B, stateid(california))))`

Evaluation

- Exact-match accuracy of code
- Compare results of executing code on database

Methods

Califf and Mooney 1999, Thompson et al. 1999, Zettlemoyer and Collins 2005, Zettlemoyer and Collins 2007, Kate and Mooney 2006, Wong and Mooney 2006, Wong and Mooney 2007, Kwiatkowski et al. 2010, Liang et al. 2013, Jia and Liang 2016, Dong and Lapata 2016…

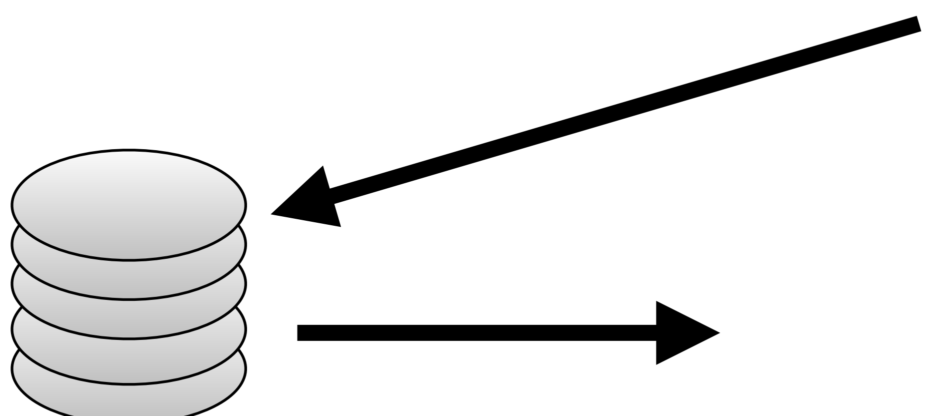
[Zelle and Mooney 1996; Tang and Mooney 2001]

Request:

Show me flights from Seattle to Boston next Monday

SQL query:

```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN
(SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND
(flight.to_airport IN (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN (SELECT city.city_code FROM city WHERE city.city_name =
'BOSTON'))) AND (flight.flight_days IN (SELECT days.days_code FROM days WHERE
days.day_name IN (SELECT date_day.day_name FROM date_day WHERE date_day.year = 1993 AND
date_day.month_number = 2 AND date_day.day_number = 8))));
```



Database

Result:

`31 flights available`

[Hemphill et al. 1990; Dahl et al. 1994]

ATIS

- Natural language interfaces to a database
- Spoken language understanding
- **Data collection:** Wizard-of-Oz experiments with experts
- **Size:** ~5,000 examples; database with 27 tables and ~160,000 entries

Charles Hemphill, John Godfrey, and George Doddington. 1990. The ATIS spoken language pilot corpus. In DARPA Speech & Natural Language Workshop.

Deborah Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In HLT.

Training Data

Natural language sentences paired with SQL queries

Evaluation

- Execute SQL on database
- Exact-match accuracy on resulting tables
- Exact representation accuracy when using lambda calculus or slot-filling frames

Methods

Pereira and Schabes 1992, Schabes 1992, Brill 1993, Ward and Issar 1994, Goodman 1996, Popescu et al. 2004, Raymond and Riccardi 2007, Zettlemoyer and Collins 2007, Kwiatkowski et al. 2011, Poon et al. 2013, Yao et al. 2013, Wang et al. 2014, Zhao and Huang 2015, Dong and Lapata 2016, Jia and Liang 2016, Iyer et al. 2017, …

[Hemphill et al. 1990; Dahl et al. 1994]

Request:

Copy the content of file 'file.txt' to file 'file2.txt'

↓

```
shutil.copy('file.txt', 'file2.txt')
```

Request:

Check if all elements in list 'mylist' are the same

↓

```
len(set(mylist)) == 1
```

[Yin et al. 2018]

CoNaLa

- Natural language pseudocode to Python implementation

- **Data collection:** question/answer pairs in Stack Overflow

- **Size:** ~2,500 curated examples; ~600,000 noisy examples

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. 2018. Learning to mine aligned code and natural language pairs from Stack Overflow. In MSR.

Training Data

Natural language utterance paired with an implementation

Evaluation

- Exact-match accuracy
- BLEU score

CoNaLa challenge and leaderboard:
conala-corpus.github.io/

[Yin et al. 2018]

Some Other Datasets

- **Django:** line-by-line pseudocode to method implementations [Oda et al. 2015]

- **HearthStone and Magic the Gathering:** trading card descriptions to class implementations [Ling et al. 2016]

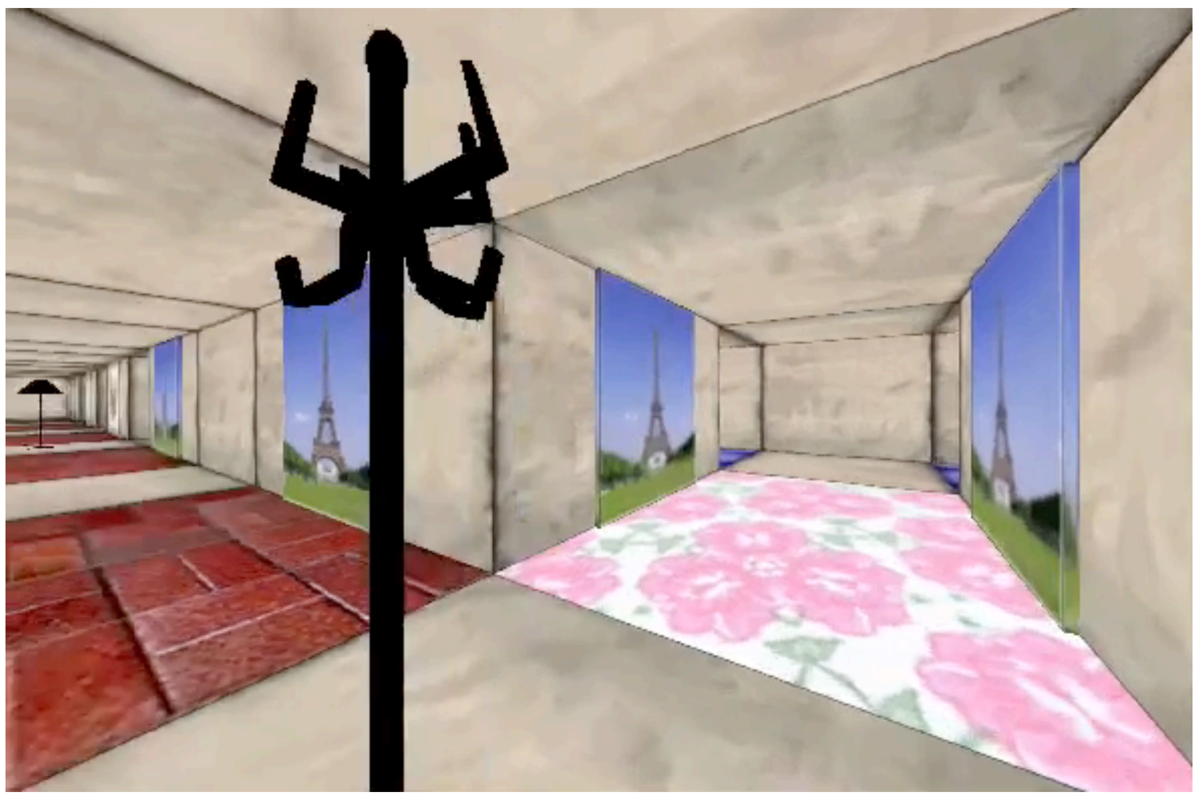
- **NL2Bash:** natural language requests to bash commands [Lin et al. 2018]

- **CONCODE:** method descriptions and existing code to method implementations [Upcoming from UW]

Research Focuses

✔ Generating executable representations

- Understanding in a situated environment
- Generalizing to broad domains
- Sequential language understanding



Instructions:

Place your back against the wall of the T intersection

Turn left

Go forward along the pink flowered carpet hall two segments to the intersection with the brick hall

Example from Artzi and Zettlemoyer 2013
[MacMahon et al. 2006; Chen and Mooney 2011]

SAIL

- Follow navigation instructions by moving in an environment
- Supervision on final world states only
- **Data collection:** virtual reality with instructors and followers
- **Size:** ~3,000 single sentences, three environments

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In AAAI.

David Chen and Raymond Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In AAAI.

Training Data

Weakly supervised learning: start state, instruction, goal state

Evaluation

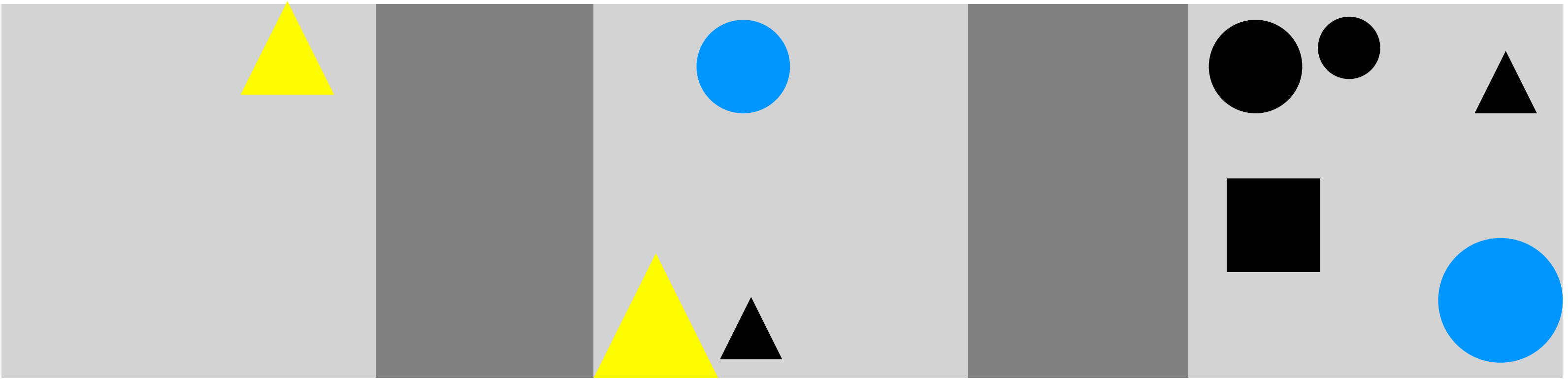
- Logical form exact match
- Execute in environment (move)
- Compare locations

Methods

Chen and Mooney 2011, Chen 2012, Kim and Mooney 2012, Kim and Mooney 2013, Artzi and Zettlemoyer 2013, Artzi et al. 2014, Andreas and Klein 2015, Mei et al. 2016, Fried et al. 2018

[MacMahon et al. 2006; Chen and Mooney 2011]

Image:



Structured representation:

```
Box 1:
1.  medium yellow triangle at (60, 0)

Box 2:
1. small black triangle at (30, 70)
                  ⋮
```

There is a box with 3 items of all 3 different colors.

TRUE

[Suhr et al. 2017]

NLVR

- Language understanding for visual reasoning
- Requires compositional reasoning: counting, comparisons, set theory
- **Data collection:** workers compare and contrast four images
- **Size:** ~4,000 unique sentences, ~92,000 total examples

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In ACL.

Training Data

- Weak supervision: sentences paired with images or structured representations, labeled as true or false
- Each sentence paired with multiple images

Evaluation

- Execute logical form on structured representation
- Accuracy: % of correct true/false predictions
- Consistency: % of unique sentences always correct

Methods

Tan and Bansal 2018, Goldman et al 2018; **Leaderboard:** http://lic.nlp.cornell.edu/nlvr/

[Suhr et al. 2017]

```
Box 1:
1.  medium yellow triangle at (60, 0)

Box 2:
1. small black triangle at (30, 70)
              ⋮
```

There is a box with 3 items of all 3 different colors.

```
λx. λy. λz. λw. box(x) ∧ count(x, 3) ∧
object(y) ∧ object(z) ∧ object(w) ∧ in(y, x) ∧
in(z, x) ∧ in(w, x) ∧ ¬(color(y) == color(z)) ∧
¬(color(y) == color(w)) ∧ ¬(color(z) == color(w))
```

TRUE

[Suhr et al. 2017, Goldman et al. 2018]

Some Other Datasets

- **Referring expressions:** resolve referring expressions to a lambda calculus expression [FitzGerald et al. 2012, FitzGerald et al. 2013]
- **Blocks world:** follow instructions to move blocks around [Bisk et al. 2016, Bisk et al. 2018]
- **Room2Room:** navigation in a realistic home environment [Anderson et al. 2017]

Research Focuses

✔ Generating executable representations

✔ Understanding in a situated environment

- Generalizing to broad domains
- Sequential language understanding

Sentence:

About 14,000 people fled their homes at the weekend after a local tsunami warning was issued, the UN said on its Web site

AMR:

```
(s / say-01
  :ARG0 (g / organization
          :name (n / name
                  :op1 "UN"))
  :ARG1 (f / flee-01
          :ARG0 (p / person
                  :quant (a / about
                           :op1 14000))
          :ARG1 (h / home
                  :poss p)
          :time (w / weekend)
          :time (a2 / after
                  :op1 (w2 / warn-01
                         :ARG1 (t / tsunami)
                         :location (l / local))))
  :medium (s2 / site
            :poss g
            :mod (w3 / web)))
```

[Banarescu et al. 2013]

AMR Bank

- Broad-coverage semantic parsing
- Map from <u>any</u> English sentence to a formal representations
- **Size:** ~40,000 sentences; multiple genres

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. <u>Abstract meaning representation for sembanking.</u> In Proceedings of the Linguistic Annotation Workshop.

AMR Bank

Training and Evaluation

Training Data

Sentences annotated with AMRs

Evaluation

SMATCH precision/recall/F1: measures semantic overlap between AMRs

Methods

Flanigan et al. 2014, Artzi et al. 2015, Peng et al. 2015, Werling et al. 2015, Wang et al. 2015, Misra et al. 2016, Damonte et al. 2016, Zhou et al. 2016, Konstas et al. 2017, …

[Banarescu et al. 2013]

WikiTableQuestions

Task

Rank	Nation	Gold	Silver	Bronze	Total
1	Venezuela	7	$\boxed{4}$	3	14
2	Bolivia	2	2	2	6
2	Chile	2	2	2	6
4	Peru	1	$\boxed{3}$	3	7
5	Ecuador	1	1	1	3
6	Colombia	0	1	2	3
Total	Total	13	13	13	39

Question:

How many countries in this competition had more than two silver medals?

Answer:

2

[Pasupat and Liang, 2015]

WikiTableQuestions

Task

Rank	Nation	Gold	Silver	Bronze	Total
1	Venezuela	7	4	3	14
2	Bolivia	2	2	2	6
2	Chile	2	2	2	6
4	Peru	1	3	3	7
5	Ecuador	1	1	1	3
6	Colombia	0	1	2	3
Total	Total	13	13	13	39

Question:

How many countries in this competition had more than two silver medals?

Logical form:

`count(R[Nation].Silver.Number.>.2)`

Answer:

2

[Pasupat and Liang, 2015]

WikiTableQuestions

- Answering questions about semi-structured tables from Wikipedia
- Must generalize to unseen tables during testing
- **Data collection:** workers ask question about tables given prompts
- **Size:** ~22,000 questions grounded in ~2,000 tables

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In ACL.

Training Data

Weak supervision: only (question, table, answer) triples

Evaluation

Generate a logical form and execute it on the table

Methods

Pasupat and Liang 2016, Haug et al. 2017, Neelakantan et al. 2017, Krishnamurthy et al. 2017

[Pasupat and Liang, 2015]

Some Other Datasets

- **Free917:** semantic parsing on Freebase [Cai and Yates, 2013]
- **WebQuestions:** semantic parsing on Freebase without logical form annotation [Berant et al. 2013]

Research Focuses

✔ Generating executable representations

✔ Understanding in a situated environment

✔ Generalizing to broad domains

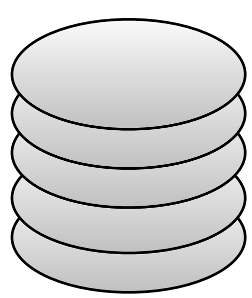
- Sequential language understanding

Request:

Show me flights from Seattle to Boston next Monday

SQL query:

```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN
(SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND
(flight.to_airport IN (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN (SELECT city.city_code FROM city WHERE city.city_name =
'BOSTON'))) AND (flight.flight_days IN (SELECT days.days_code FROM days WHERE
days.day_name IN (SELECT date_day.day_name FROM date_day WHERE date_day.year = 1993 AND
date_day.month_number = 2 AND date_day.day_number = 8))));
```



Database

Result:

`31 flights available`

[Hemphill et al. 1990; Dahl et al. 1994]

ATIS

Task

Previous request:

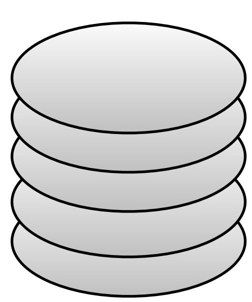
Show me flights from Seattle to Boston next Monday

Request:

On American Airlines

SQL query:

```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN
(SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND
(flight.to_airport IN (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN (SELECT city.city_code FROM city WHERE city.city_name =
'BOSTON'))) AND (flight.flight_days IN (SELECT days.days_code FROM days WHERE
days.day_name IN (SELECT date_day.day_name FROM date_day WHERE date_day.year = 1993 AND
date_day.month_number = 2 AND date_day.day_number = 8))) AND flight.airline_code = 'AA');
```



Database

Result:

`5 flights available`

[Hemphill et al. 1990; Dahl et al. 1994]

ATIS Interactions

- Natural language interfaces to databases
- Spoken language understanding
- **Data collection:** Wizard-of-Oz experiments with experts
- **Size:** ~1,000 interactions; database with 27 tables and ~160,000 entries

Charles Hemphill, John Godfrey, and George Doddington. 1990. The ATIS spoken language pilot corpus. In DARPA Speech & Natural Language Workshop.

Deborah Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In HLT.

Training Data

Sequence of natural language utterances paired with a logical form (e.g., lambda calculus or SQL query)

Evaluation

- Utterance-level accuracy
- Process user requests in sequence
- Execute SQL on database
- Lambda calculus: exact-match accuracy

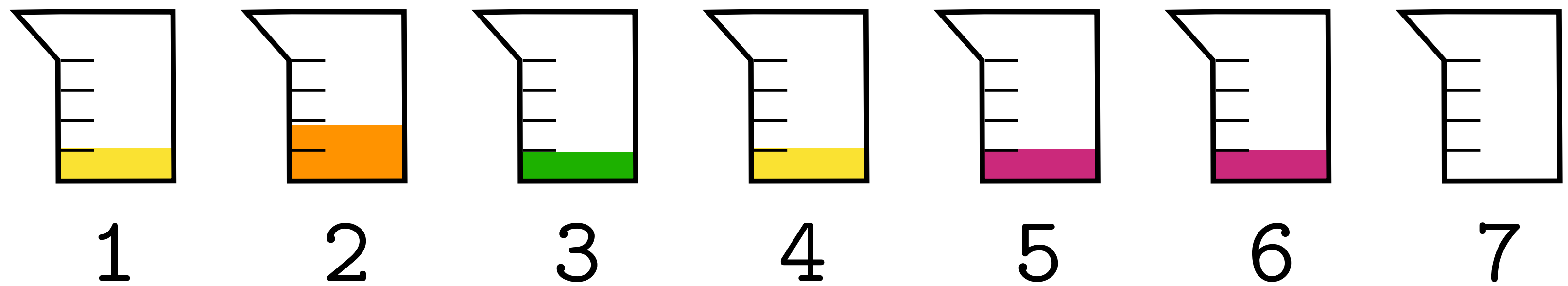
Methods

Miller et al. 1996, Zettlemoyer and Collins 2009, Suhr et al. 2018

[Hemphill et al. 1990; Dahl et al. 1994]

SCONE

Task



Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

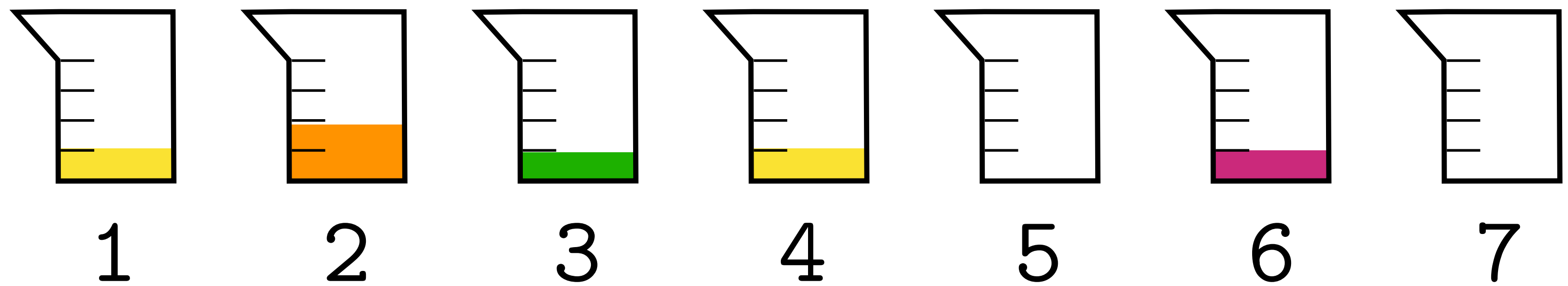
Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

SCONE

Task



Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

SCONE

Task

Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

→ *Mix it*

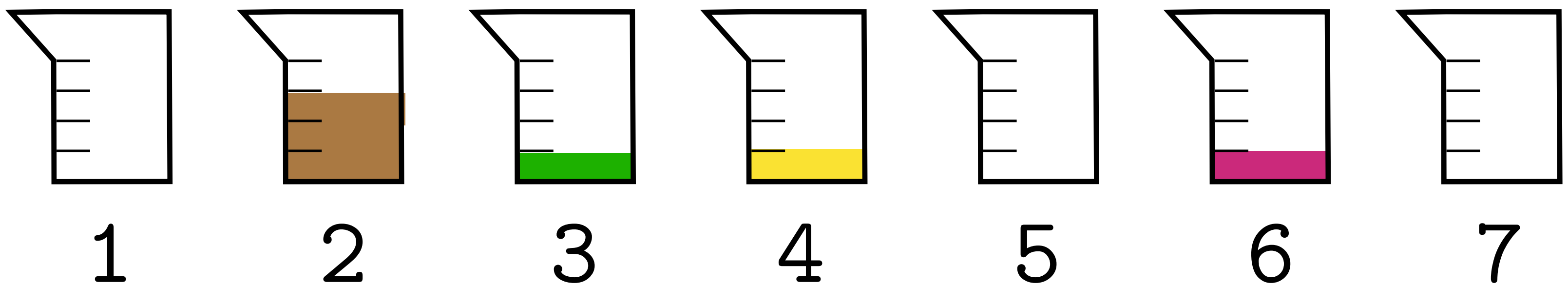
Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

SCONE

Task



Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

SCONE

Task

Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

Then, drain 1 unit from it

→ *Same for 1 more unit*

[Long et al. 2016]

SCONE

Task

Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

SCONE

- Sequential, context-dependent instructions in small, manipulable environments
- Only access to world state annotation during training
- **Data collection:** ask workers to describe changes in environment states
- **Size:** ~13,000 examples grounded in three domains

Reginald Long, Panupong Pasupat, and Percy Liang. 2016. Simpler context-dependent logical forms via model projections. In ACL.

Training Data

Start state and sequence of instructions and world states

Empty out the leftmost beaker of purple chemical

Then, add the contents of the first beaker to the second

Mix it

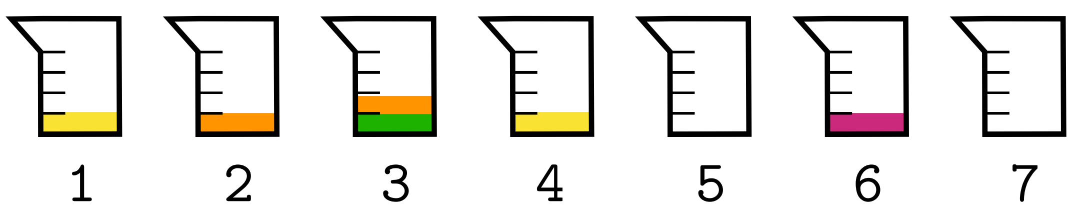
Then, drain 1 unit from it

Same for 1 more unit

[Long et al. 2016]

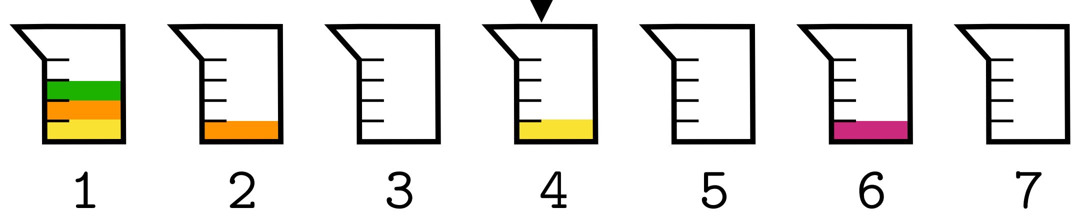
Evaluation

- Follow instructions one by one, changing state over time
- Compare final world state at end of execution



Add the third beaker to the first

```
pour(index(allObjects, 1), index(allObjects, 2))
```



Methods

Long et al. 2016, Guu et al. 2017, Fried et al. 2018, Suhr and Artzi 2018

[Long et al. 2016, Guu et al. 2017, Fried et al. 2018]

Some Other Datasets

- **SQA:** answer sequence of questions about Wikipedia tables [Iyyer et al. 2016]
- Task-oriented dialogues over knowledge bases [e.g., Eric et al. 2017]

Summary

✔ Generating executable representations

✔ Understanding in a situated environment

✔ Generalizing to broad domains

✔ Sequential language understanding