An Iterative Approach to Synthesize Data Transformation Programs

Bo Wu and Craig Knoblock
University of Southern California
# Learning Transformation Programs by Example

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Target Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)</td>
<td>2000 Ford Expedition los angeles $4900</td>
</tr>
<tr>
<td>1998 Honda Civic 12k miles s. Auto. - $3800 (Arcadia)</td>
<td>2008 Mitsubishi Galant Sylmar CA $7500</td>
</tr>
<tr>
<td>2008 Mitsubishi Galant ES $7500 (Sylmar CA) pic</td>
<td>1998 Honda Civic Arcadia $3800</td>
</tr>
<tr>
<td>1996 Isuzu Trooper 14k clean title west covina $999 (west covina) pic</td>
<td>1996 Isuzu Trooper west covina $999</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Time complexity is **exponential** in the **number** and a **high polynomial** in the **length** of examples.
Reuse subprograms

Original: 2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)

Target: 2000 Ford Expedition los angeles $4900

Position program=(left context, right context, occurrence)

Learned Programs

After 1\textsuperscript{st} example

\((\text{START,NUM,1})\quad (\text{BNK,NUM,1})\)

After 2\textsuperscript{nd} example

\((\text{START,NUM,1})\quad (\text{BNK,NUM,1})\)

After 3\textsuperscript{rd} example

\((\text{START,NUM,1})\quad (\text{BNK,ANY,3})\)

7/30/15
Identify incorrect subprograms

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 Ford Expedition 11k runs great los angeles $4900 (los angeles)</td>
<td>2000 Ford Expedition los angeles $4900</td>
</tr>
<tr>
<td>1998 Honda Civic 12k miles s. Auto. - $3800 (Arcadia)</td>
<td>2008 Mitsubishi Galant Sylmar CA $7500</td>
</tr>
</tbody>
</table>

Program

```
(START, NUM, 1) (BNK, NUM, 1) ('(', WORD, 1) (LWRD, '),', 1) (ANY,BNK'$', 1) (NUM, BNK, 2)
```

Execution Result:

```
 0    -1
11   ≠
 0   23
```

Output:

```
2008 Mitsubishi Galant Sylmar CA $7500
```

Input:

```
2008 Mitsubishi Galant ES $7500 (Sylmar CA) pic
```
Update hypothesis spaces

Program

Hypothesis $H_3$

2000 Ford Expedition 11k runs great los angeles $4900$ (los angeles)

1998 Honda Civic 12k miles s. Auto. - $3800$ (Arcadia)

2008 Mitsubishi Galant ES $7500$ (Sylmar CA) pic
Evaluation

• Dataset
  – **D1**: 17 scenarios used in (Lin et al., 2014)
    • 5 records per scenario
  – **D2**: 30 scenarios collected from student data integration projects
    • about 350 records per scenario
  – **D3**: synthetic dataset
    • designed to evaluate scale-up

• Alternative approaches
  – **Our implementation of Gulwani’s approach**: (Gulwani, 2011)
  – **Metagol**: (Lin et al., 2014)

• Metric
  – Time (in **seconds**) to generate a transformation program
Program generation time comparisons

Table: time (in seconds) to generate programs on D1 and D2 datasets

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPBE</td>
<td>0</td>
<td>5</td>
<td>0.34</td>
<td>0</td>
</tr>
<tr>
<td>Gulwani’s approach</td>
<td>0</td>
<td>8</td>
<td>0.59</td>
<td>0</td>
</tr>
<tr>
<td>Metagol</td>
<td>0</td>
<td>213.93</td>
<td>55.1</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>D2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPBE</td>
<td>0</td>
<td>1.28</td>
<td>0.20</td>
<td>0</td>
</tr>
<tr>
<td>Gulwani’s approach</td>
<td>0</td>
<td>17.95</td>
<td>4.02</td>
<td>0.33</td>
</tr>
<tr>
<td>Metagol</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
</tbody>
</table>

Figure: scalability test on D3
Discussion

• Our iterative PBE approach significantly reduces time in synthesizing programs

Future work

• Extend to domains with only partial traces

• Help user to determine when to stop transforming on large datasets.
Thanks

Please come to my poster #23 for more details

Bo Wu
bowu@isi.edu
References


Different number of segments

Trace

Input: 1998 Honda Civic 130 k miles - $3800 (Arcadia)

Output: 1998 Honda Civic Arcadia $3800

Hypothesis Spaces:

H_3

h_{31} h_{32} h_{33}

Execution Result: 2000 Ford Expedition Los Angeles $4900

Old Program:

(START, NUM, 1) (BNK, NUM, 1) (BNK, LWRD, 2) (NUM, BNK, -1)

Start = 24  End = 39