Data Recombination for Neural Semantic Parsing

Presented by: Edward Xue
Robin Jia, Percy Liang
Intro

• Semantic Parsing: The translation of natural language into logical forms
• RNNs have had much success recently
  • Few domain specific assumptions allows them to be generally good without much feature engineering
• Good Semantic Parsers rely on prior knowledge
  • How do we add prior knowledge to an RNN model?
Sequence to Sequence RNN

• Encoder
  • Input utterance is a sequence of words: $x = x_1, \ldots x_m \in V^{(in)}$
  • Converts to sequence of context sensitive embeddings: $b_1, \ldots, b_m$
  • Through a bidirectional RNN
    • Forward direction: $h^F_i = \text{LSTM}(\phi^{(in)}(x_i), h^F_{i-1})$
  • Each embedding is a concatenation of the forward and backward hidden state
Sequence to Sequence RNN

• Decoder: Attention based model
  • Generates output sequence one token at a time: \( y = y_1, \ldots y_n \in V^{(out)} \)

\[
\begin{align*}
  s_1 &= \tanh(W^{(s)}[h^F_m, h^B_1]). \\
  e_{ji} &= s_j^T W^{(a)} b_i. \\
  \alpha_{ji} &= \frac{\exp(e_{ji})}{\sum_{i'=1}^m \exp(e_{ji'})}. \\
  c_j &= \sum_{i=1}^m \alpha_{ji} b_i. \\
  P(y_j = w \mid x, y_{1:j-1}) &\propto \exp(U_w [s_j, c_j]). \\
  s_{j+1} &= \text{LSTM}([\phi^{(out)}(y_j), c_j], s_j).
\end{align*}
\]
Attention Based Copying: Motivation

• Previously just chose next output word using a softmax over all words in the output vocabulary
• Does not generalize well for entity names
• Entity names often correspond directly to output tokens: eg “iowa” -> iowa
Attention Based Copying

• At each time step $j$ also allow the decoder to copy any input word directly to the output, instead of writing a word to the output

\[
P(a_j = \texttt{Write}[w] \mid x, y_{1:j-1}) \propto \exp(U_w[s_j, c_j]),\]

(8)

\[
P(a_j = \texttt{Copy}[i] \mid x, y_{1:j-1}) \propto \exp(e_{ji}).\]

(9)
Attention Based Copying Results

<table>
<thead>
<tr>
<th></th>
<th>GEO</th>
<th>ATIS</th>
<th>OVERNIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Copying</td>
<td>74.6</td>
<td>69.9</td>
<td>76.7</td>
</tr>
<tr>
<td>With Copying</td>
<td>85.0</td>
<td>76.3</td>
<td>75.8</td>
</tr>
</tbody>
</table>

Table 1: Test accuracy on GEO, ATIS, and OVERNIGHT, both with and without copying. On OVERNIGHT, we average across all eight domains.
Data Recombination

• This framework induces a generative model from the training data
• Then, it samples from the model to generate new training examples.
• The generative model here is a Synchronous CFG
Data Recombination

Start with training set $D$ of $(x, y)$ pairs.
$p(x, y)$ is the empirical distribution
Fit a generative model $\hat{p}(x, y)$ to $p$ which generalizes beyond with recombinat examples
To train model, maximize $E[\log p_{\theta}(y \mid x)]$
Data Recombination

• Synchronous CFG
  • Set of Production rules \( X \rightarrow \langle \alpha, \beta \rangle \)
• The generative model is the distribution over the pairs \((x,y)\) defined by sampling from \(G\)
• SCFG is only used to convey prior knowledge about conditional independence structure
• Initial grammar generated as \(\text{ROOT} \rightarrow \langle x, y \rangle\)
Data Recombination: Grammar Induction Strategies

• Abstracting Entities
  • Abstracts entities with their types

• Abstracting Whole Phrases
  • Abstracts both entities and whole phrases with their types

• Concatenation
  • For any $k \geq 2$, CONCAT-K creates two types of rules
  • ROOT going to a sequence of $k$ SENT’s
  • Then for each $\text{ROOT} \rightarrow <\alpha,\beta>$ in the input grammar, add rule $\text{SENT-} > <\alpha,\beta>$ to the output grammar
Examples
("what states border texas ?",
answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(texas))))))
("what is the highest mountain in ohio ?",
answer(NV, highest(V0, (mountain(V0), loc(V0, NV), const(V0, stateid(ohio))))))

Rules created by ABSENTITIES
ROOT → { "what states border STATEID ?",
        answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(STATEID))))}
STATEID → { "texas", texas }
ROOT → { "what is the highest mountain in STATEID ?",
        answer(NV, highest(V0, (mountain(V0), loc(V0, NV),
        const(V0, stateid(STATEID))))})
STATEID → { "ohio", ohio }

Rules created by ABSWHOLEPHRASES
ROOT → { "what states border STATE ?", answer(NV, (state(V0), next_to(V0, NV), STATE))}
STATE → { "states border texas", state(V0), next_to(V0, NV), const(V0, stateid(texas))}
ROOT → { "what is the highest mountain in STATE ?",
        answer(NV, highest(V0, (mountain(V0), loc(V0, NV), STATE))))

Rules created by CONCAT-2
ROOT → { SENT₁ </s> SENT₂, SENT₁ </s> SENT₂ }
SENT → { "what states border texas ?",
        answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(texas))))}
SENT → { "what is the highest mountain in ohio ?",
        answer(NV, highest(V0, (mountain(V0), loc(V0, NV), const(V0, stateid(ohio))))})
Datasets

• GeoQuery (GEO): questions about US geography paired with answers in database query form. 600/280 split.
• ATIS: queries for a flight database paired with corresponding database queries. 4473/448 split
• Overnight: Logical forms paired with natural language paraphrases over eight different subdomains. For each domain, random 20% as test, the rest split into 80/20 training/development set
Experiments: GEO and ATIS

<table>
<thead>
<tr>
<th>Previous Work</th>
<th>GEO</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zettlemoyer and Collins (2007)</td>
<td>88.9</td>
<td>84.6</td>
</tr>
<tr>
<td>Kwiatkowski et al. (2010)</td>
<td>88.9</td>
<td>84.6</td>
</tr>
<tr>
<td>Liang et al. (2011)$^2$</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>Kwiatkowski et al. (2011)</td>
<td>88.6</td>
<td>82.8</td>
</tr>
<tr>
<td>Poon (2013)</td>
<td>88.9</td>
<td>84.2</td>
</tr>
<tr>
<td>Zhao and Huang (2015)</td>
<td>88.9</td>
<td>84.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our Model</th>
<th>GEO</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Recombination</td>
<td>85.0</td>
<td>76.3</td>
</tr>
<tr>
<td>ABSENTITIES</td>
<td>85.4</td>
<td>79.9</td>
</tr>
<tr>
<td>ABSWHOLEPHRASES</td>
<td>87.5</td>
<td></td>
</tr>
<tr>
<td>CONCAT-2</td>
<td>84.6</td>
<td>79.0</td>
</tr>
<tr>
<td>CONCAT-3</td>
<td>87.5</td>
<td>77.5</td>
</tr>
<tr>
<td>AWP + AE</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td>AE + C2</td>
<td>88.9</td>
<td>78.8</td>
</tr>
<tr>
<td>AWP + AE + C2</td>
<td>89.3</td>
<td>83.3</td>
</tr>
<tr>
<td>AE + C3</td>
<td>89.3</td>
<td>83.3</td>
</tr>
</tbody>
</table>

Table 2: Test accuracy using different data recombination strategies on GEO and ATIS. AE is ABSENTITIES, AWP is ABSWHOLEPHRASES, C2 is CONCAT-2, and C3 is CONCAT-3.
Experiments: Overnight

<table>
<thead>
<tr>
<th></th>
<th>BASKETBALL</th>
<th>BLOCKS</th>
<th>CALENDAR</th>
<th>HOUSING</th>
<th>PUBLICATIONS</th>
<th>RECIPES</th>
<th>RESTAURANTS</th>
<th>SOCIAL</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Previous Work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang et al. (2015)</td>
<td>46.3</td>
<td>41.9</td>
<td>74.4</td>
<td>54.0</td>
<td>59.0</td>
<td>70.8</td>
<td>75.9</td>
<td>48.2</td>
<td>58.8</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Recombination</td>
<td>85.2</td>
<td>58.1</td>
<td>78.0</td>
<td>71.4</td>
<td>76.4</td>
<td>79.6</td>
<td>76.2</td>
<td>81.4</td>
<td>75.8</td>
</tr>
<tr>
<td>ABSENTITIES</td>
<td>86.7</td>
<td>60.2</td>
<td>78.0</td>
<td>65.6</td>
<td>73.9</td>
<td>77.3</td>
<td>79.5</td>
<td>81.3</td>
<td>75.3</td>
</tr>
<tr>
<td>ABSWHOLEPHRASES</td>
<td>86.7</td>
<td>55.9</td>
<td>79.2</td>
<td>69.8</td>
<td>76.4</td>
<td>77.8</td>
<td>80.7</td>
<td>80.9</td>
<td>75.9</td>
</tr>
<tr>
<td>CONCAT-2</td>
<td>84.7</td>
<td><strong>60.7</strong></td>
<td>75.6</td>
<td>69.8</td>
<td>74.5</td>
<td>80.1</td>
<td>79.5</td>
<td>80.8</td>
<td>75.7</td>
</tr>
<tr>
<td>AWP + AE</td>
<td>85.2</td>
<td>54.1</td>
<td>78.6</td>
<td>67.2</td>
<td>73.9</td>
<td>79.6</td>
<td><strong>81.9</strong></td>
<td><strong>82.1</strong></td>
<td>75.3</td>
</tr>
<tr>
<td>AWP + AE + C2</td>
<td><strong>87.5</strong></td>
<td>60.2</td>
<td><strong>81.0</strong></td>
<td><strong>72.5</strong></td>
<td><strong>78.3</strong></td>
<td><strong>81.0</strong></td>
<td>79.5</td>
<td>79.6</td>
<td>77.5</td>
</tr>
</tbody>
</table>

Table 3: Test accuracy using different data recombination strategies on the OVERNIGHT tasks.
Experiments: Effects of longer examples

![Graph showing test accuracy over number of additional examples for different conditions.]
Conclusions

• Data Recombination seems to provide better test accuracy in lieu of more training examples
  • Would this generalize well?
• Attention Based Copying is useful for certain datasets
Thank you